

2022-09-19

Technical Debt is an Ethical Issue

Paul John Gibson

Télécom SudParis, paul.gibson@telecom-sudparis.eu

Yannis Stavrakakis

Technological University Dublin, jonstavrakakis@gmail.com

Massamaesso Narouwa

Télécom SudParis

See next page for additional authors

Follow this and additional works at: <https://arrow.tudublin.ie/ascnetcon>



Part of the [Applied Ethics Commons](#), [Computer Sciences Commons](#), and the [Ethics and Political Philosophy Commons](#)

Recommended Citation

Gibson, J.P., Gordon, D., Collins, M., Narouwa, M., O'Sullivan, D., and Turner, J. Minding the Gap: Computing Ethics and the Political Economy of Big Tech. Proceedings of Ethicomp, 20th International Conference on the Ethical and Social issues in Information and Communication Technologies, Turku, Finland, 26-28th July, 2022. DOI: 10.21427/Y73Z-6P82

This Conference Paper is brought to you for free and open access by the Applied Social Computing Network at ARROW@TU Dublin. It has been accepted for inclusion in Conference Papers by an authorized administrator of ARROW@TU Dublin. For more information, please contact arrow.admin@tudublin.ie, aisling.coyne@tudublin.ie, gerard.connolly@tudublin.ie.



This work is licensed under a [Creative Commons Attribution-NonCommercial-Share Alike 4.0 License](#)
Funder: European Union

Authors

Paul John Gibson, Yannis Stavrakakis, Massamaesso Narouwa, Damian Gordon, Dympna O'Sullivan, Jonathan Turner, and Michael Collins



UNIVERSITY
OF TURKU



Proceedings of the ETHICOMP 2022

Effectiveness of ICT ethics – How do we help solve ethical problems in the field of ICT?

Eds. Jani Koskinen, Kai K. Kimppa, Olli Heimo, Juhani Naskali, Salla Ponkala
and Minna M. Rantanen

UNIVERSITY OF TURKU, Turku, Finland

ISBN 978-951-29-8989-8 (PDF)

Proceedings of the ETHICOMP 2022

Effectiveness of ICT ethics – How do we help solve ethical problems in the field of ICT?

Editors:

Jani Koskinen, Kai K. Kimppa, Olli Heimo, Juhani Naskali, Salla Ponkala and Minna M. Rantanen

Publisher:

University of Turku, Turku, Finland

Copyrights:

Copyright © 2022 for the individual papers by the papers' authors. Copyright © 2022 for the volume as a collection by its editors. This volume and its papers are published under the Creative Commons License Attribution 4.0 International ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

Cover art

“Northern Star” by Minna M. Rantanen

ISBN 978-951-29-8989-8 (PDF)

Table of Contents

Preface.....	6
The need for multiple approaches for ethics in technology education (Norberto Patrignani and Iordanis Kavathatzopoulos)	8
Prospects of equity in post-pandemic technology education (Isabel Alvarez and Nuno Silva)	15
Securing Meetings in D2D IoT Systems (Sabina Szymoniak and Olga Siedlecka)	30
Intelligent Identity Authentication, Using Face and Behavior Analysis (Mariusz Kubanek, Janusz Bobulski and Łukasz Karbowski)	42
Zero Trust Model and the Shift in the Ethos of Cybersecurity – Towards the Deleuzian Society of Control (Jukka Vuorinen and Ville Uusitupa)	52
Building the Learning Environment for Sustainable Development: a Co-creation approach (Ewa Duda).....	62
Capitalism, Technology and the Elderly: The Ethic of Inclusive Digital Conscience for African Values (Thando Nkohl-Ramunenyiwa).....	78
Technology and Ethics for Alternative Medicine (T V Gopal)	90
Endless Forms Most Deceitful (William M. Fleischman, Nick Langan and Leah N. Rosenbloom)...	104
IT ‘Experts’ Considered Harmful: A Concerning Case of False Expertise in the Legal System (Reuben Kirkham).....	123
From the Page to Practice: Support for Computing Professionals Using a Code of Ethics(Don Gotterbarn, Michael S. Kirkpatrick, and Marty J. Wolf)	136
“It Belongs in a Museum!”: A Practical Take on the Question of Artificial Intelligence and Moral Patency (Lauri Tuovinen)	153
Working with Affective Computing: Exploring UK Public Perceptions of AI enabled Workplace Surveillance (Lachlan Urquhart, Alexander Laffer and Diana Miranda)	164
Evaluation of the impact of health technologies on life expectancy and Avoidable mortality in Asian and Europe countries: case study of Japan, South Korea, Lithuania and Luxembourg (Younes Karrouk, Driss Ezouine, Mohammed Hassani Zerrouk and Mario Arias Oliva).....	178
Release the robot dogs: Teaching with professional codes of ethics (Stacy A. Doore and Azalea Yunus).....	187
Transparency and product safety regarding medical diagnostic systems (Daria Onitiu).....	204
An Interdisciplinary Approach to European Trustworthy Digital Environments (Martin Griesbacher and Hristina Veljanova)	219
Computing Apples and Oranges? Implications of Incommensurability for (Fair) Machine Learning (Arto Laitinen and Otto Sahlgren)	232
Trust and Explainable AI: Promises and Limitations (Sara Blanco)	245
Technical Debt is an Ethical Issue (J Paul Gibson, Massamaesso Narouwa, Damian Gordon, Dymrna O’Sullivan, Jonathan Turner and Michael Collins)	257
EKIP: designing a practical framework for embedding ethics in AI software development (Rebecca Raper and Mark Coeckelbergh)	266
Responsibility by Design: Actionable strategies and a tool for leveraging technology ethically and enabling innovation responsibly (Veikko Ikonen, Emad Yaghmaei, Janika Miettinen, Giovanna Sanchez Nieminen)	278
Factors contributing to the intention to use exoskeletons in the workplace – exploring the role of ethics (Stéphanie Gauttier).....	292

What does privacy mean to users of voice assistants in their homes? (Edith Maier, Michael Doerk, Michelle Muri, Ulrich Reimer and Uwe Riss)	299
From Liberal Democracies to Blockchain Systems: The Instrumentalization of Decentralization (Sofia Cossar, Wessel Reijers and Primavera De Filippi and Rodrigo Alonso-Alcalde)	314
Industrial Limitations on Academic Freedom in Computer Science (Reuben Kirkham)	328
Guidelines to Develop Trustworthy Conversational Agents for Children (Escobar-Planas Marina, Gómez Emilia and Martínez-Hinarejos, Carlos-D)	342
Emerging values in ICT during uncertain times: the case of COVID-19 (Ana Saraiva, Luís Tavares and Nuno Silva)	361
Perceived Risks of the Data Economy: Autonomy and the Case of Voice Assistants (Uwe V Riss, Edith Maier and Michael Doerk)	374
Mapping and understanding human factors in effective cybersecurity: a finance-sector organisation case study (Robin Renwick, Eliza Jordan, Eliseo Venegas Mayoral, Amanda Segura Gonzalez and Leire Cubo Arce)	387
Probabilistic Analysis of Security Protocols Using Probabilistic Timed Automata (Olga Siedlecka-Lamch and Sabina Szymoniak)	401
Choosing the Right Cybersecurity Solution: A Review of Selection and Evaluation Criteria (Rafał Leszczyna)	418
Bullshit Blockchain Mining? (Kazuyuki Shimizu)	436
The Hidden Side of Digital Technologies for Family Use: Privacy and Other Related Ethical Issues in Digital Distance Education and Telecommuting at Home (Ryoko Asai and Sachiko Yanagihara)	446
Smart patches in mass-casualty incidents (Katerina Zdravkova and Ana Madevska Bogdanova)	457
Does Artificial Intelligence Evolve or Degenerate Sports? (Kiyoshi Murata, Yohko Orito and Yasunori Fukuta)	470
The social implications of brain machine interfaces for people with disabilities: Experimental and semistructured interview surveys (Yohko Orito, Tomonori Yamamoto, Hidenobu Sai, Kiyoshi Murata, Yasunori Fukuta, Taichi Isobe and Masahi Hori)	486
IT Ethics Protocol and Usability for Enhancing Culture in the Spanish Museums (Raquel García-Martín, Ana María Lara-Palma, Bruno Baroque-Zanón and Rodrigo Alonso-Alcalde)	501
Rationalizing Emotion through Technology: Or Addressing the Political and Liberal Narrative in Emotion Technology (Eugenia Stamboliev and Mark Coeckelbergh)	
Water Use as Remote Monitoring Technology: An Ethical Analysis (Tania Moerenhout and Katleen Gabriels)	526
Sustainability and Technology: Digital Marketplaces in Voluntary Market Offsetting (Rafael Canorea-García and Mario Arias-Oliva)	532
Using VLE Engagement Tracking to Offer Leniency to Under Performing Students (Brian Keegan, Paul Doyle, Brian Gillespie and Dympna O'Sullivan)	538
Towards A Theory of Artificial Justice: Rawlsian Ethics Guidelines for Fair AI (Salla Ponkala)	542
Minding the Gap: Computing Ethics And the Political Economy of Big Tech (Ioannis Stavrakakis, Damian Gordon, J. Paul Gibson, Dympna O'Sullivan and Anna Becevel)	545
Regulating AI in Europe: from Ethics to Legal Rules building on the GDPR example (Costas Popotas and Maria Bottas)	550
Addressing ethical issues in the design of smart home technology for older adults and people with disabilities (Jonathan Turner, Dympna O'Sullivan, Damian Gordon, Yannis Stavrakakis, Brian Keegan and Emma Murphy)	554

The rights of caregivers to their personal data in the context of disability services (Anne-Marie Tuikka and Ville Kainu)	560
Snooping, Stalking, Breakup and Letting Go - Examining Emerging Norms, Values, and Technological Drivers around Relationships & Break-Ups online (Wilhelm Klein and Aara Cho)...	564
The ethical dilemma while using technology to mitigate covid-19 pandemic restrictions (Ala Ali Almahameed, Jorge Pelegrín-Borondo, Jorge de Andrés Sánchez, Orlando Lima Rua, Maria Alesanco Llorente, Alba García Milon and Mario Arias-Oliva)	569
Care and Data: How can we use healthcare data ethically? (Ryoko Asai)	573
“Be a Pattern for the World”: The Development of a Dark Patterns Detection Tool to Prevent User Loss (Jordan Donnelly, Alan Dowley, Yunpeng Liu, Yufei Su, Quanwei Sun, Lan Zeng, Andrea Curley, Damian Gordon, Paul Kelly, Dympna O’Sullivan and Anna Becevel)	577
Chinese Soft Power: Data Politics (Nehme Khawly and Mario Arias-Oliva)	582
Abolitionist Technology Ethics for Secondary Education(Leah Rosenbloom and William Fleischman)	586
Free Speech and Computing Professionals: Moral Considerations and Tensions (Michael Kirkpatrick, Mandy Burton and Marty J. Wolf)	590
Access Control Meets Genetics: The Challenges of Information Leakage and Non-users (Michael Kirkpatrick)	595

Preface

Jani Koskinen¹[0000-0001-8325-9277], Kai Kimppa¹[0000-0002-7622-7629], Olli Heimo¹[0000-0001-9412-0393], Juhani Naskali¹[0000-0002-7559-2595], Salla Ponkala¹[0000-0002-1441-8673] and Minna Rantanen¹[0000-0001-8832-5616]

¹ Information Systems Science, Turku School of Economics, University of Turku, Turku, Finland
jasiko@utu.fi

This Ethicomp is again organized in exceptional times. Two previous ones were forced to turn to online conferences because of Covid-pandemic but it was decided that this one would be the physical one or cancelled as the need for real encounters and discussion between people are essential part of doing philosophy. We need possibility to meet people face to face and even part of the presentation were held distance—because of insurmountable problems of arriving by some authors— we manage to have real, physical conference, even the number of participants was smaller than previous conferences.

The need of Ethicomp is underlined by the way world nowadays is portrayed for us. The truthfulness and argumentation seem to be replaced by lies, strategic games, hate and disrespect of humanity in personal, societal and even global communication. EThicomp is many times referred as community and therefore it is important that we as community do protect what Ethicomp stands for. We need to seek for goodness and be able to give argumentation what that goodness is. This lead us towards Habermass communicative action and Discourse ethics which encourages open and respectful discourse between people (see eg.Habermass 1984;1987;1996). However, this does not mean that we need to accept everything and everybody. We need to defend truthfulness, equality and demand those from others too. There are situations when some people should be removed from discussions if they neglect the demand for discourse. Because by giving voice for claims that have no respect for argumentation, lacks the respect of human dignity or are not ready for mutual understanding (or at least aiming to see possibility for it) we cannot have meaningful communication. This is visible in communication of all levels today and it should not be accepted, but resisted. It is duty of us all.

References

- Habermas, J. (1984) *The theory of communicative action*, Vol. 1, Polity Press.
- Habermas, J. (1987) *The Theory of Communicative Action: Lifeworld and Systems, a Critique of Functionalist Reason*, Volume 2, Polity Press.
- Habermas, J. (1996) *Between facts and norms: Contributions to a discourse theory of law and democracy*, Polity press.

Full Papers

The need for multiple approaches for ethics in technology education

Norberto Patrignani¹ and Iordanis Kavathatzopoulos²

¹ Politecnico of Torino, Torino, Italy
norberto.patrignani@polito.it

² Uppsala University, Uppsala, Sweden
iordanis.kavathatzopoulos@it.uu.se

Abstract. The need for ethics education in technology curricula is universally recognized. The difficult question is how to deploy this principle into the real universities' courses. This paper introduces a methodology based on multiple approaches: pneumatophores (computer experts with a high ethical profile), stakeholders' network (the analysis of the complex relationships among the many stakeholders of a digital complex system) and a psychological awareness (the personal skills needed by responsible designers of digital complex systems). All these approaches are currently used in computer ethics courses with encouraging positive results from students.

Keywords: Computer Ethics, Pneumatophores, Stakeholders' Network, Psychology of technologists

1 Introduction

Information and communication technologies (ICT) they have come to shape society and the planet in disturbing ways, by now they are an integral part of the challenges of the Anthropocene, that is, the first geological era in which human activities are able to influence the atmosphere and alter its balance (Crutzer, Stoermer, 2000). To face these challenges, the data-information-knowledge production-chain must move on to a systemic view of the *Infosphere* (Floridi, 2001) by developing future complex sociotechnical systems that are *socially desirable*, *environmentally sustainable* and *ethically acceptable* (Patrignani, 2020). This means that designers of future digital systems should take into account multiple requirements: *a human-centered approach* (accessibility, learnability, respect for privacy, etc.), *the minimization of environmental impact* (in terms of material used, power consumption, repairability and recyclability by-design, etc.), and a special *care on working conditions* alongside the entire ICT supply-chain (health hazards for workers, their required skills and social contracts, etc.). This challenging scenario for future designers of complex digital systems requires a strong ethical competence, but the question becomes: how to introduce ethics in technology education? What are the best strategies? A strategy based on a multiple approach is described in the following.

2 Multiple approaches

The challenging task of teaching ethics in engineering courses at university level is well known (Patrignani and Kavathatzopoulos, 2017) and it is recognized that a unique “silver bullet” does not exist for it (Johnson, 2020). A possible way for overcoming these limits is to go for a *multiple approach* and it is up to the teacher to select the most appropriate one in accordance with the context (social and cultural, age of students, type of university, time availability, etc.).

2.1 Pneumatophores

In this paper, with the term “pneumatophores” are indicated people who act as “spirit bearers”, that is, people who can be seen as exemplary by computer professionals (Huff and Barnard, 2009). Their acts and the story of their life can be an inspiration for young students that are preparing to become designers of future complex digital systems. Some of the most important *pneumatophores* in the history of computing are:

René Carmille (1886-1945), expert from the French “*Service de la démographie*” during the Nazi occupation of France in the WWII, he modified the machines for punched cards so that they did not print column 11 (which indicated religion), thus saving many Jews from concentration camps, arrested in 1944, he wouldn't talk even under torture, he was sent to Dachau camps where he died on 25 January 1945, it can be considered the first “ethical hacker” (Davis, 2015);

Norbert Wiener (1894-1964), professor at MIT, considered one of the founders of the computer age (with Alan Turing, John Von Neumann, and Claude Shannon), he introduced the new discipline of “cybernetics” (Wiener, 1948) and is recognized as one of the first computer scientists that refused to collaborate with the military apparatus: “*I do not expect to publish any future work of mine which may do damage in the hands of irresponsible militarists...*” (Wiener, 1947);

Joseph Weizenbaum (1923-2008), computer scientist and professor at MIT, considered one of the founder of the so called *artificial intelligence* by writing the first chatbot “Eliza” in 1966, he warned about the social impact of computing by introducing a distinction between *deciding* (a computational activity, something that can ultimately be programmed since it is something very logical, a type of System 2 thinking according to Kahneman, 2011) and *choosing* (the product of judgment, not calculation, it is the capacity to choose that ultimately makes us human) and by recommending that human functions that require “*judgement, respect, understanding, caring and love*” ought not to be substituted by computers (Weizenbaum, 1976), this warning is becoming particularly important in a time when humans are slowly living an *epistemological shift* from using technologies for *forecasting* (*prae-videre*, to see in advance with the mind's eyes, to estimate how something will be in the future) to use technologies for *predicting* (*prae-dicere*, to tell in advance, announcing

something will happen in the future, a prophecy, a statement of what will happen in the future), and finally to using technologies for *prescribing* to humans (*prae-scribere*, to write in advance, a prescription is a written order, a direction, like written directions from a doctor), a terrible example of this *epistemological shift* is the delegation to kill other humans (a *choice* that cannot be reduced to a *decision* programmed into a machine) happening now with *Lethal Autonomous Weapons Systems* (UN, 2021), a threshold that, also according to Red Cross International, should not be passed (ICRC, 2022);

Severo Ornstein, one of the leading scientists in the history of computing, he worked at the Lincoln Laboratory of MIT in 1955, pioneered the Internet at *Bolt Beranek & Newman* in 1969, founder of the *Computer Professionals for Social Responsibility* (CPSR) association in 1982, a computer professional with a deep awareness of the context and limitations of many working environments in the computer field: “... *I was concerned that many colleagues had their heads down and ... were not paying attention to the social consequences of what they were doing*” (Ornstein and Gould, 1994);

Batya Friedman and Terry Winograd, who contributed to the definition of the mission of CPSR: “... *support public discussion of, and public responsibility for decisions involving the use of technology in systems critical to society... dispel popular myths about the infallibility of technologies... challenge the assumption that technology alone can solve political and social problems... critically examine social and technical issues within the ICT profession,... encourage the use of information technology to improve quality of life*” (Friedman and Winograd, 1990);

Donald Gotterbarn, Professor Emeritus at East Tennessee State University and founder of the *Software Engineering Ethics Research Institute*, developed the *Software Development Impact Statement* process (SoDIS) that encourages those involved in software project management to consider the wider ramifications of their work, chair of the *ACM Committee on Professional Ethics* and main contributor to the *Software Engineering Code of Ethics* in 1999 and the new *ACM Code of Ethics* in 2018, received the *Joseph Weizenbaum Award* in 2010 “... *for his role in developing the moral consciousness of the profession*” (ACM, 2010);

Aaron Swartz (1986-2013), software developer, writer, hacker, Internet activist, in July 2011 was arrested for downloading 4.8 million scientific articles from the *JSTOR* academic database, the *symbolic act* (for underlining the restricted access to scientific papers imposed by publishing companies) was considered a crime and the cyber fraud trial included a potential imprisonment of up to 50 years and a fine of up to four million dollars, he took his own life on January 11, 2013, Lawrence Lessig said: “*the question this government needs to answer is why it*

was so necessary that Aaron Swartz be labeled a felon” (Gillmor, 2013), Tim Berners Lee said: “*we have lost a mentor, a wise elder*” (Knappenberger, 2014);

Edward Snowden, from inside the *National Security Agency* (NSA), the top US organization dedicated to national security, revealed publicly the existence of secret mass surveillance programs based on digital technologies, after unsuccessfully raising ethical issues internally, he decided to quit his job and became an international case when, on 7 June 2013, newspapers such as the *New York Times* and *Der Spiegel*

published his revelations, in 2013, he was named “person of the year” by the British newspaper *The Guardian* (Snowden, 2019).

And of course, this in an incomplete list.

The proposed methodology starts by introducing the history of computing and showing that many of the ethical dilemmas, currently at the co-shaping intersection between technology and society, have been always present for many years. Then alongside this history, introducing the main “pneumatophores” and to propose to study their lives, their context in terms of technology, social, economic and political terms. Very few students know, for example, that even Leonardo Da Vinci reflected on the possible dual-use of his submarine invention: “*I do not describe my method of remaining under water,... I do not publish nor divulge these, by reason of the evil nature of men, who would use them for assassinations at the bottom of the sea*” (Da Vinci, 1506).

This first approach based on historical background and real lives of computer experts, is useful for students as an “escaping” way from the dominant ICT “short-termism”. It is important to show them that a critical thinking approach to “techno-determinism” and to the relationship between technology and society has been proposed by philosophers and social researchers since the 1930s (Mumford, 1934; Ellul, 1954; 2014); and also to demonstrate that human choices and critical thinking can be more innovative and can have a positive impact even in complex digital systems design. Also important is to underline the need for “ethical hackers” in the next future, society will need more and more people that are experts in computing and, at the same time, also aware of the social and ethical impact if ICT, *ethical hackers* that put the public interest as a priority, that “... *contribute to society and to human well-being, acknowledging that all people are stakeholders in computing*” as suggests the *ACM Code of Ethics and Professional Conduct* (ACM, 2018). For example, taking into account the growing importance of cybersecurity, there will be many job opportunities in the future for “ethical hackers”.

2.2 Stakeholders’ network

What is a stakeholders’ network? In many contexts (e.g., ethics research, corporate social responsibility studies, etc.) it is a tool for stimulating a reflection on the relationships (the *arcs* connecting the *nodes*, or the stakeholders). The visual representation (a *network of nodes and arcs*) facilitates the investigation about the relationships: are they balanced? Or are they based on power relationships? For example, a company developing profit is it ready to share the value created with the stakeholders that contributed to its creation? (Freeman, 2010).

According to some researchers the Anthropocene era started on 16 July 1945, at 5:29 a.m. when the Trinity Test took place, the first nuclear test of history, in Alamogordo, desert of Jornada del Muerto, New Mexico, USA, in the context of the Manhattan Project (Monastersky, 2015). In this era the future of the planet, as a consequence of many critical global problems, starting from the *climate change*, is in the hands of *homo sapiens*. According to this view, a new stakeholder should appear in all stakeholders’ network, including the digital technologies stakeholders’ network:

the *planet Earth*. Designers and engineers of digital systems need to introduce this important stakeholder and the concept of *limits* of technology. Digital technologies need a growing number of materials for their construction (including *rare earths* and minerals), a growing amount of electric power (during their functioning for powering data centers, networks, devices, etc.) and are producing a growing amount of *e-waste*. A good engineer of complex digital system should have a *systemic view* that includes all these aspects. The analysis of this network, in particular the relationships, can be useful for the identification of potential ethical issues, and for cultivating the *ars interrogandi* of students. Questions like *what kind of relationship is there?* Arise. To qualify these relationships it is necessary to analyse them in depth: are they symmetrical? What kind of *meanings and values* are embedded in this relationship? (in terms of power, dimension, role, values, and desires). A fruitful approach like the *comparative approach* used by humans' mind when comparing two entities in general (Patrignani and Kavathatzopoulos, 2021).

2.3 Psychology of technologists

Investigating ethics in technology education imply also to look into the personal way of thinking of researchers. A very interesting example of this view is the evolutionary path of thinking of one of the most important scientists of XX century: Richard Feynman (1918-1988). In his life he participated in two main events: the *Manhattan project* (1942-1946) and the investigation following the *Space Shuttle* disaster (1986). When people put to Feynman the difficult question about his participation in the development of the atomic bomb, he honestly replied: "*I simply didn't think, ok?*" and this demonstrates the risks involved in de-contextualising science and technology development and the so-called curiosity-oriented scientific activities. During the investigation of the *Space Shuttle* disaster, Feynman's rigorous approach provided the possibility to find the truth of the consequences of big technoscientific enterprises and the risk of personal ethics, responsibility and integrity (Benessia, 2021). The awareness of the context represents a fundamental aspect of the way of thinking of researchers, scientists, engineers, designers, all people that in their life reach a high level of concentration capability, but what does it mean the honest admission Richard Feynman "*I simply didn't think, ok?*" How can we stimulate a responsible research and innovation? How can we improve the ethical competence of young researchers, computer professionals? In a simple way: is it possible to learn "how to think"? (Arendt, 2003; Plato, 2004).

Concentrating on the skills and competences of designers of digital technologies implies also to investigate also their "decision-making" processes, that is a philosophical and psychological research. Only with this knowledge, it is possible to develop tools and suitable methods for improving ethical competence (Kavathatzopoulos, 2015).

A good exercise for students is to analyse the stakeholders' network from different point of views, a kind of "role-game" where they are invited to split into different groups of stakeholders, for example: *the developers, the users, and the policy-makers*.

Then the discussion about a real case involving a complex stakeholders' network can start with a simple brainstorming, then to identify the possible ethical issues, etc. The key step is to invite the students to *shift* their roles, *the users* become *the developers*, *the policy-makers* become *the users*, and *the developers* become *the policy-makers*, etc. This allows the students to exercise their capability to see outside the *comfortzone*, to cultivate their ethical skills.

3 Conclusions

The multiple strategy suggested in this paper is an approach currently used in “computer ethics” courses for PhD students in our universities with positive results. First of all, the students are engaged and involved by the real stories about engineers with a deep deontological background, “ethical hackers” (pneumatophores). Second, they are able to start with a simple stakeholders' network (*policy makers*, *technology developers*, and *users*) and then to increase the number of nodes (stakeholders) up to twelve-twenty on average. As a consequence, also the analysis of relationships becomes more rich.

Third, they are able to reflect on their ethical skills and competences and on the most appropriate *virtue* for an engineer: the capability to act in complex situations, with a sensitivity about what is needed in a given situation, *thinking and acting* in complex situations (“*phronesis*”, Aristotle, 1975).

This approach also requires several further studies in particular the need to start *thinking in the long-term* in the digital community, including scientist, educators, professionals, users, policy makers. The main problem is that this requires a kind of “future ethics” (Birnbacher, 2006; Sollie, 2007), a *long-term* perspective. On the other side *homo sapiens* developed in an era where the need to protect from predators was dominant, so quick and short-term reactions were needed. The dramatic challenges in front of humanity (like *climate change* and *loss of biodiversity*) require a big leap towards an *ethics for the future*, an ethics no more limited to humans but including also *animals* and the *planet Earth*. Digital technologies, from one side can help in “decarbonize” many activities (“*using bits for consuming less*”). On the other side, digital technologies are having a significant impact on the environment, in terms of materials needed (including “*rare earths*”), in terms of energy needed to power the gigantic data centers of cloud computing, and in terms of the need to recycle and repair them for relieving the problem of *e-waste* (“*consuming less for using bits*”). Unfortunately, up to now the ICT world has been dominated by the Silicon Valley “mantra”: *disrupt first, ask questions later*, now humans are realizing that “later” could be *too late*. A very dangerous short-term view.

On the contrary ICT designers need to try to “anticipate” problems, and this requires time, even to slow down the design process (Patrignani and Whitehouse, 2018) for involving all the stakeholders with their needs and values.

How can we evolve this way of thinking in the Anthropocene era? How can we introduce a long-term view, a “future ethics” in digital technology education?

References

- ACM (2010). *Advocate of Computer Ethics Honored by International Ethics Society*, <http://www.acm.org/press-room/news-releases/2010/inseit-award-2010/>
- ACM (2018). *ACM Code of Ethics and Professional Conduct*, <https://www.acm.org/code-ofethics>
- Arendt, H. (2003). *Responsibility and judgment*. Schocken Books
- Aristotle (1975). *Ἠθικά Νικομάχεια (Nicomachean ethics)*, Papyros
- Benessia, A. (2021). “I simply didn't think, ok?” Some reflections on the quality of scientific research, *Visions for sustainability*, September, www.ojs.unito.it/index.php/visions
- Birnbacher, D. (2006). *What motivates us to care for the (distant) future?*, Working Papers, N°04/2006. Iddri Seminar on Sustainable Development
- Bruemmer, B.H. (1994). *An Interview with Severo Ornstein and Laura Gould*, Charles Babbage Institute, Center for the History of Information Processing, University of Minnesota.
- Crutzen, P.J. and Stoermer, E.F. (2000). The “Anthropocene”, *Global Change Newsletter* (41): 17–18
- Da Vinci, L. (1506). *Code Leicester* (f.15A-22v). In Da Vinci L. (2002), *The Notebooks of Leonardo Da Vinci*, Konecky & Konecky, 2002
- Davis A. (2015). *A History of Hacking*, The Institute – IEEE
- Ellul, J. (1954). *La technique ou l'enjeu du siècle*, Armand
- Ellul, J. (2014). *Théologie et technique. Pour une éthique de la non-puissance*, Labor et Fides
- Floridi, L. (2001). Ethics in the Infosphere, *The Philosophers' Magazine* 16 (2001):18-19
- Freeman, R.E. (2010). Managing for Stakeholders: Trade-offs or Value Creation, *Journal of Business Ethics*, Vol. 96, 7-9
- Friedman B., Winograd T. (1990), Computing and social responsibility: a collection of course syllabi, *ACM Digital Library*, cpsr.org
- Gillmor, D. (2013, 13 Jan). Remember Aaron Swartz by working against government abuses, *The Guardian*
- Huff, C., Barnard, L. (2009). Good Computing. Moral Exemplars in the Computing Profession, *IEEE Technology and Society Magazine*, Fall 2009
- ICRC (2022). *International Committee of the Red Cross (ICRC) position on autonomous weapon systems*: ICRC position and background paper No. 915 January 2022
- Johnson, D.G. (2020). *Engineering Ethics. Contemporary and Enduring Debates*, Yale University Press
- Kahneman, D. (2011). *Thinking, fast and slow*, MacMillan
- Kavathatzopoulos, I. (2015). ICT and sustainability: Skills and methods for dialogue and policy making, *Journal of Information Communication and Ethics in Society*, 13(1):13-18

- Knappenberger, B. (2014). *The Internet's Own Boy: The Story of Aaron Swartz*, <https://www.youtube.com/watch?v=gQLIodJVbz8>
- Monastersky, R. (2015). First atomic blast proposed as start of Anthropocene, *Nature*. <https://doi.org/10.1038/nature.2015.16739>
- Mumford L. (1934). *Technics and Civilization*, Harcourt
- Ornstein, S., Gould, L. (1994, November 17). *An interview with Severo Ornstein and Laura Gould* (B. H. Bruemmer, Interviewer), Charles Babbage Institute Archives, University of Minnesota Libraries
- Patrignani, N. (2020). *Teaching Computer Ethics: Steps towards Slow Tech, a Good, Clean, and Fair ICT*, Uppsala University
- Patrignani, N., Kavathatzopoulos, I. (2017). On the Difficult Task of Teaching Computer Ethics to Engineers, *ORBIT Journal* 1(1)
- Patrignani, N. Whitehouse, D. (2018). *Slow Tech Slow Tech and ICT. A Responsible, Sustainable and Ethical Approach*, Palgrave-MacMillan.
- Patrignani, N. Kavathatzopoulos, I. (2021). Teaching of Technology IS Teaching of Ethics. But How? In (eds) Koskinen, J., Rantanen, M., Tuikka, A., Knaapi-Junnila, S. (2021) *Tethics2021 - Proceedings of the Conference on Technology Ethics*, Turku, Finland, October, 20-22, 2021
- Plato (2004). *Πρωταγόρας (Protagoras)*, I. Zacharopoulos
- Snowden, E. (2019). *Permanent record*, Pan Mac Millan
- Sollie, P. (2007). Ethics, technology development and uncertainty: an outline for any future ethics of technology, *Journal of Information, Communication and Ethics in Society*, Vol.5, N.4, pp.293-306
- UN (2021, 8 March). Final report of the Panel of Experts on Libya established pursuant to Security Council resolution 1973 (2011), S/2021/229 , United Nations, Security Council
- Weizenbaum, J. (1976). *Computer Power and Human Reason: From Judgment To Calculation*, Freeman
- Wiener, N. (1947). A Scientist Rebels, *Atlantic Monthly*, January
- Wiener, N. (1948). *Cybernetics, or, Control and communication in the animal and the machine*, The Technology Press

Prospects of equity in post-pandemic technology education

Alvarez, Isabel ¹[0000-0003-4381-6328]¹ and Silva, Nuno ²[0000-0003-0157-0710]²

¹Istec / Autonomia TechLab / Comegi

²Universidade Lusíada de Lisboa

¹alvarez@edu.ulusiada.pt, ²nsas@lis.ulusiada.pt

Abstract. The digital transformation has reached all societal areas and not all stakeholders are prepared for these rapid changes. After two years of the coronavirus (COVID-19) pandemic global health crisis which had a profound effect on people's lives – how they learn, work and live - we must reflect on the inequities that occur in the use of technology not only in learning institutions but also and mainly in the population in general. The possibilities available to individuals and institutions by the latest technological advancements are so important that it is no more possible to refrain using these new assets in our daily life. The use of information and communication technology has grown at an unprecedented rate, and it is recognised that the use of various digital devices together with cloud computing allows for scenarios never considered. This paper examines the impact of the digital transition in Portugal before and during the pandemic times, as well as prospected post-pandemic time.

Keywords: Technology, Education, Equity, COVID-19, Portugal

1 Introduction

Most professionals assume they lack the technological skills needed for the future. It is generally recognised that there is a growing world crisis in terms of digital competences urging some course of action to solve this need.

However, technological interventions in the work, home, and public environments, increasingly complicate these socio-technical contexts, affecting our laws, norms, beliefs, and values. Whether good or bad, an empirical understanding of these impacts leads to innovative ethical problem-solving for ICT solutions and readiness to confront emerging challenges and move towards effective ethics and positive change. Technological advances and increasing concerns around pervasive surveillance and the role and unintended consequences of algorithms, provide a rich picture of fears and realities.

Although it is recognised that education may not influence the use of computer technological devices such as sales terminals for routine tasks, it may enable professionals to carry out other higher order tasks.

There will be a need to define the resources found beneficial, assess needs for the future, innovate and implement transparency in alternative computational systems. Novel computer systems impose substantial education issues in the use of their technology and technology education found beneficial for all stakeholders must be defined in a more equitable and ethical way.

Empirical understanding of these impacts leads to innovative ethical problem-solving for ICT solutions and readiness to confront emerging challenges and move towards effective ethics and positive change. Technology is closing the gap between humans and machines which is a new challenge that needs to be faced.

The quick digital transition which almost every country is trying to reach did not prepare the population and left outside old people, those with digital illiteracy and the most vulnerable ones who are likely to face additional barriers being more at risk of increased vulnerability and less likely to receive the support and extra services they need (OECD, 2020). In effect, the pandemic only highlighted and deepened the main digital equity concerns that we deal in this paper.

Alvarez, Silva, and Correia (2015), debated the importance of cyber education which should be enhanced to guarantee a more equitable global society. However, there are generation gaps that need to be taken in consideration in terms of equal competence and digital literacy and which may cause an inter-generational inequity. Finally, balance between cultural and ethical issues should be provided.

2 Methodology

This paper followed a systematic literature review on digital transition and technology education and, to provide a comprehensive study concerning the difficulties encountered by the population on the way technology was used to control the pandemic situation and do the vaccination to the Portuguese society, a case study was done wherefrom some proper suggestions came out on how the actual situation concerning the improvement of technology education of the population should be done. Given the short timeframe since the start of the different lockdown periods that occurred from March 2020 to March 2022, the collection of data was not as comprehensive as in other studies.

3 Digital Transition

The digital transition from analogue to digital is a long process (Irniger, 2017).

The Portuguese society has begun to experience rapid changes as a digitally transforming culture with both the civil society and the state under a strong decentralizing process.

The digital revolution is transforming society; not only from the way we work, but also the way we connect with friends and family and entertain ourselves. These social

and cultural changes are enabled by the increased speed and reach of human communication and other computer technological advances. Diverse industries, from software to financial services to media to pharmaceuticals, are being forced to adapt to this new reality to stay relevant, and so, too, must our democratic institutions.

We argue that this transition implies that citizens are required to possess a new type of competence that contains both financial and digital skills. Citizens who are unable to develop or acquire such competence are likely to be disadvantaged by the services. Lastly, we argue that these developments pose significant challenges for public administrations to ensure the overall quality of the public services (Breit & Salomon, 2015).

However, a key question is whether digital self-service systems can translate complex rules and legislations into a service that is understandable, meaningful, and productive for citizens.

We argue that this transition requires citizens to possess a new kind of combined digital and financial competence to be able to handle the obstacles. Citizens who do not have this competence or are unable to develop or otherwise acquire it (e.g., using private consultants), are likely to be disadvantaged by the selfservice process.

Digital services are often constructed in such a way that they require self-service, for instance as citizens are expected go online to plan their health treatment, apply for schools or kindergartens, or acquire information to aid their retirement pension decisions (Breit & Salomon, 2015).

Despite these benefits, there are five phases that citizens go through as they use the digital services: interest, access, comprehension, reflection, and support. While these phases do not capture all aspects of the delivery of digital services, citizens are likely to have difficulties to face in this process (Breit & Salomon, 2015).

4 Technology Education

First, technology education should not be confused with educational technology. There are technology learning mechanisms focused on how people develop knowledge and competencies and technology learning systems focused on institutional practices. Anyway, educational technology and technology education are intertwined because one cannot subsist without the other. In addition, a new era of teaching and digital learning emerged from COVID-19.

While the introduction of technology education into primary subjects is viewed by supporters to nurture skills, attitudes, and competencies in early learners (McLaughlin, 2021), there is a lack of research focused on older learners and equity on digital citizenship.

Education is a fundamental sector for the cultural and critical formation of society, and for the individual enhancement of its constituents. Culture and learning seem to be interwoven and inseparable. According to Goodfellow and Hewling (2005), cultural dilemmas in e-Learning are related to two main issues: inequities arising from dominant cultural values embodied in content, and potential for miscommunication amongst participants arising from cultural particularities. Thus, it constitutes an

intrinsic market that is not exhausted in historical epochs, but suffers influences of a cultural, technical, and political nature. Other approaches to learning must be considered like the Stem Education that uses science, technology, engineering, the arts and mathematics as access points for guiding student inquiry, dialogue, and critical thinking.

The European Commission therefore lists “Digital Competence” as one of the key competences for lifelong learning (EEA, 2022). Key competences include knowledge, skills, and attitudes needed by all for personal fulfilment and development, employability, social inclusion and active citizenship. The proposed approach is to promote key competences by providing high-quality education, training and lifelong learning for all, supporting educational staff in implementing competence-based teaching and learning approaches and encouraging a variety of learning approaches and contexts for continued learning.

If we add to this that, in the field of Information and Communication Technologies (ICT), one of the most important challenges that currently exists is to cover the often-insufficient demand for professionals trained in scientific-technological careers, and that this demand is forecast to remain high in the coming years.

The technological environment is accelerating because of COVID-19, since, in relation to education, it involved the shift from the face-to-face model to the remote model. The criticality of technological platforms has increased. Video conferencing tools and computer equipment in the home, have meant that technology platforms have, and need to be improved. It is often recommended to create content for a range of devices, such as iPads or mobile phones, and they should make sure they are optimised for the best learning experience. However, the intensification of the use of technologies has enabled the development of different forms of communication and interaction, transforming society and culture. It is therefore essential that teachers discuss and understand the digital culture in which they are inserted and interact in it, and continuing education is one of the dimensions that can contribute to solving the problems of lack of training (Martins et al., 2020).

According to Ferri, Grifoni and Guzzo (2021), teachers should be trained to increase digital and other specific e-learning skills to properly plan and implement an innovative pedagogical programme. Students are generally familiar with the use of digital devices, but they may not be prepared to receive remote learning and it is quite difficult to capture their attention. As for parents, they may not have the necessary experience in terms of digital skills for e-learning. Thus, in addition to teacher training, parent training is also an important factor for the success of e-learning, and one that has been totally ignored. The school has not been mobilised in this sense, and the providers in a passive way put on online channels, namely YouTube, recordings that can be seen by the public of webinars, seminars, etc.

Modernising the initial and in-service training of teachers and trainers so that their knowledge and skills respond to society's changes and expectations and are adapted to the different groups they are targeting is one of the main challenges that education and training systems will have to address in the short term.

In Portugal, most teachers obtained their diploma 25 years ago or more and, in some cases, the updating of their skills has not kept pace with the changes inherent to

an information and knowledge society (Santos, 2018). For example, teachers were already using the school publishers' digital platforms before COVID19, but it was mentioned that they often did work with links mentioned in the books that students then did not have access to. This was one of the reasons why they stopped using it.

There are also other main questions like the failure of the infrastructure for data networks, mobile equipment, and guaranteed access to the internet. We are facing severe equity problems, where regional, social, and even basic survival needs are intertwined.

5 Equity

The adverse equity impact of the pandemic has been tremendous. While the disruptions caused by the

Covid-19 pandemic have affected both rich and poor countries, the impact on students from vulnerable groups has been greater than for the average student population. In low-income groups, students from under-represented groups (low-income students, girls, members of minority groups, students living in remote areas, and special needs students) have been hit especially hard, suffering economic hardship, encountering connection difficulties, and living in emotional distress (UNESCO, 2021).

In spite of this, ICT plays a crucial role in a knowledge society for the personal and social development and senior citizens cannot afford being excluded from ICT innovations. However, the developments of the knowledge and information society also created new social fault lines and inequalities.

According to Stahl, Timmermans, and Flick (2017), digital divides are also related to fairness and equity: while some individuals and groups will be able to better communicate with one another, different availability of technologies and diverging abilities to use them may erect barriers to communication in some cases; also uncertainty can arise due to 'function creep', when data collected or technology designed for a specific purpose may, over time, become used for other (e.g. monitors helping older citizens digitally disabled). Understanding older adults' perceptions of technology (lack of clarity in, instructions and support) is an important factor to facilitate independent living.

Seniors face twofold exclusion as they are both underrepresented in (adult) education and technology use. This can have serious consequences for social inclusion and segregation on the macro level and their everyday lives and quality of life on a micro level (for example not being able to buy a bus ticket anymore). Accessibility to ICT based services is an important ethical issue, independent on where you live, your economic situation or your knowledge.

To promote dignity includes to respect older peoples' independency, integrity, participation, safety and security, the benefit of the service for the individual, and equity and justice.

According to Pérez-Escolar and Canet (2022), digital exclusion is especially acute among vulnerable groups like older adults and disabled people. It is important and

more effective to first identify the problems and issues of vulnerable people, and then select what technological innovations can help solve those obstacles and help people to acquire the digital competencies and skills needed to use this technology. There are two forms of technological stratification - digital divide and knowledge gap. This digital divide has led to the second form, a knowledge gap, which is, as it sounds, an ongoing and increasing gap in information for those who have less access to technology.

To promote equal opportunities for participation in the digital age (equity and access), some problems arise in clusters and schools regarding the digital divide, differences in training, generational differences, and even computer illiteracy. However, other issues have arisen, such as partnerships and private negotiations with suppliers, which creates a problem of inequality in obtaining IT resources. The free e-learning platform Moodle is generally used as a repository of content for traditional classes, maintaining additional values related to access, privacy, and security. However, a more in-depth approach to improving student learning is under several constraints, including technical factors, equipment, content, interactivity, and instructional design. The potential implications are the independent use of the Moodle e-learning platform, and its impact on content sharing issues.

Also, unfortunately, the rate of technological renewal does not necessarily imply the rate of implementation in schools, and thus the guarantee of equity. However, the platforms with cloud computing technology that have spread because of COVID-19, such as Microsoft Office365 (including Teams) and the online platforms of publishers, may be beneficial in this regard.

Several news reports also give an end to equity problems: Parents' representatives consider that the distance learning model implemented to replace as face-to-face classes during the COVID-19 pandemic created and accentuated inequalities among students, but that it was the possible solution (lifestyle.sapo.pt, 2021).

Recently, a group of constitutionalists came to remind us that the principle of equality must be preserved by the distance learning modality. Let us start then with this criterion: which model of distance education provides the best guarantees of equality and universality in the right to education? (ionline.sapo.pt, 2021)

5.1 Equity of collaboration

The leadership drives an organisational culture and cultural awareness designed to avoid failure and reconcile the inconsistencies between different local policies and different online equity proposals. For example, the curricula equity and the uncompromising defence of quality in teaching assume special relevance in terms of the moral and ethical stance. On the other hand, with the teacher specifically as the focus, the provision of formal and informal training, was widely used specially to improve digital competencies (e.g., FCT/NAU - <https://www.nau.edu.pt>).

Over the pandemic times, emerged in YouTube a lot of collaboration and training videos, and it is also possible to see and review testimonies and recorded sessions, namely webinars, seminars, digital meeting cycles, tutorials, etc. where it is possible

to see the impact of these short videos, given the number of views that are recorded, and added over time.

6 Case study

Several participatory notes, documents and documentaries were systematically reviewed to answer the question: What was the impact of COVID-19 in digital competencies in Portugal?

It is important to understand how this pandemic has reinforced and aggravated new and old social inequalities. The people most affected by this crisis are also the most vulnerable from the point of view of social inequalities, those with less social, economic, cultural and symbolic capital (Tavares & Cândido, 2020). Furthermore, the pandemic exposed the extent of the digital divide and the socio-economic inequalities that perpetuate glaring gaps among nations, citizens, higher education institutions and the students themselves (UNESCO, 2021). We have analysed the digital transition case in Portugal.

The process which had already started before 2020, accelerated during the pandemic times for the whole Portuguese society.

6.1 Education

In what regards education, probably the major test for the Portuguese Government among the most important challenges created by COVID-19 is how to adapt the actual system of education built around physical schools (OECD, 2020). The COVID-19 pandemic has provided all stakeholders with an opportunity to pave the way for introducing digital learning and it is generally recognized that there is a need to innovate and implement alternative educational and assessment strategies (Dhawan, 2020).

In the first year of the pandemic, everyone was caught off-guard and the priority was to distribute computers to the student households where there was no equipment, and internet to the places without network. The main concern was to keep all primary and secondary students with a minimum connection to the schools and teachers and rehearse a remote learning never tested before with this dimension. Public and private organisations provided laptops and internet access to some students from disadvantaged backgrounds with little access to such tools and requiring further attention and support. When this was not possible, in cooperation with the Post Office Services and the National Scouts Group, a mechanism was implemented allowing students who lived far from schools, had no computer or without access to the Internet, to receive hard copies lessons and tasks from their teachers. Deliveries of homework/assignments on paper to students and the following collection and return to the teachers were also organised (OECD, 2020).

The virtual classroom platforms like videoconferencing (Google Hangouts Meet, Zoom, Microsoft Teams and Cisco WebEx) and customizable cloud-based learning management platforms such as Moodle and Class365 were increasingly being used.

School closures and confinement measures mean more families have been relying on technology and digital solutions to keep children engaged in learning, entertained and connected to the outside world, but not all children have the necessary knowledge, skills and resources to keep themselves safe online. In Portugal, elementary and secondary students from rural villages may have illiterate parents. Most of these students may not have access to computers or smartphones or TV at home in addition to poor Internet connectivity. Part of the population lost their jobs or their usual income. Continuous access to the internet is expensive and not affordable for the poorer families and face-to-face online classes consume more data packages (OECD, 2020).

Approximately 800 schools across the country hosted children from elementary schools whose parents worked in essential services, as well as provided food support to students from disadvantaged economic backgrounds which most of the time the only meal they have during the day is at school. Schools also reinforced their articulation with the Resource Centres for Inclusion, to ensure the continuity of their specialised support services for students (OECD, 2020).

Concerning the students with special needs, the role of parents is crucial in online learning to help on general or specific study questions, while special education teachers can run on distance learning to clarify sessions for parents. Portugal, meanwhile, provides a rare example of where disadvantaged children performed just as well as others because “inclusion” is a core part of its educational framework (OECD, 2021).

In most of the cases during the pandemic both parents were also at home teleworking. This brought the problem of the number of computers available for all the members of the family and the issues around physical workspaces.

The important thing is to guarantee that learners participate and learn in the actual context, and general society touch digital competencies. But how can this be guaranteed? There are a lack of readiness felt by the students mainly those with greater socio-economic difficulties and older people concerning their digital skills. The alternative is an ethics of care which directs our attention to the need for responsiveness in relationships (paying attention, listening, and responding) (Martel et al., 2021).

Equity means that all students, independently of their personal and social identifiers, have equal access to the educational resources and opportunities they need (Ezra et al, 2021). As we say above, the shift of the Emergency Remote Teaching during this pandemic brought a decrease in educational equity. So, the implementation of e-learning technology must provide an equitable and inclusive environment with transparency and truthfulness. The success of digital education requires not only available technical infrastructure at a student’s home, but, more importantly, the right digital pedagogical skills and effective student-parent teacher communication. In Portugal, a large group of teachers will retire in 2022, which will cause an inability to provide adequate or formal training for new teachers in the numbers needed to replace those who are retiring.

6.2 Digital citizenship issues

For the population in general, any government activity serving the Portuguese citizens started being done digitally. This quick digital transition for all the governmental services done in the country left outside the people who do not have the technical knowledge or equipment to access these institutions or anyone to help them. While in what concerns the education sector, there was a convergence of efforts to help to solve the problem, for the rest of the population with digital needs, there was nothing thought or prepared to help the digital transition. There should have been a better alignment of the digital transition and the preparation of the needs and expectations of the Portuguese society where the quick digital transition left outside the older and those with digital illiteracy.

Also, during these pandemic times, it was through technology that the population was medically assisted or summoned to be vaccinated. However, part of the population does not use computers or a mobile phone, or if they do, they lack the knowledge of using it, which caused a failure in the purpose of the system. This recent experience brings the recognition of the fact that there is a need to increase technology education to the population in general, not only re-skilling from disappearing jobs but also for staying updated within the actual digital environment. But this leads us to a debate on how this can be done.

Whatever the technological intervention, it was obvious from the earliest days that these changes would have enormous equity implications—at the campus and national levels and at the global level. While many trumpeted the equity benefits of technology in higher education delivery, in practice, the digital divide will plague societies rich and poor for a generation of students. Policymakers and education leader must acknowledge this up front and continue to evolve policy interventions to minimise the impact of those most at-risk. In terms of global equity, societies with robust infrastructure (ICT and energy but also housing security and social safety nets) were more quickly and comprehensively able to pivot to online and remote learning. The learning loss in terms of time and content will have been minimised as much as possible. On the other hand, countries with low internet penetration, unstable electricity, expensive ICT and mobile networks, and housing instability or crowding have experienced larger break in learning, if any continued delivery was possible at all (UNESCO, 2021).

In the context of digitization, citizens' digital skills play a crucial role. Over the past two decades, there has been a growing attention on the so-called digital divide, fairness, and equity. Whereas much of the prior focus has been on a binary classification of physical access, research has increasingly paid attention to the social, psychological, and cultural factors that impact citizens' willingness and ability to use technology.

7 Proposed Solution

This serious public health crisis and the economic and social problems associated evidence that the Government interference is crucial particularly to restrain or reduce growing social inequalities (Costa, 2020).

Although in what concerns enterprises and other institutions, there was already for long a law that every employee should have at least one training course per year on any subject, most of this training was done on IT training courses only focusing on simple IT subjects like using the different Office tools or the Internet. Never on how to install hardware or software, how to solve security problems, or the better use of mobiles – how to install apps or to do a secure payment. If this training should be more directed to the real digital transition of the country, focusing on the real needs of the population, in this case the employees, the problem that arose during the pandemic times that made people alone at their homes do computer actions never done before, like installing and configuring new software, hardware and communications, would have had less consequences. One of the good things that these pandemic times brought was that in whatever sector, the digital transition should be strongly supported by government interference. Based on this recent experience, new measures have been approved and started being put in practice by the new government plan for digital transition.

7.1 Adults' Digital Inclusion Programme

Development of an educational project for the digital inclusion of one million info-excluded adults during the legislature, based on a national network of 10,000 young volunteers and 950 training centres. The basic training content covered in the programme involves, among others, creating and managing an email account, online search capabilities, consulting and using digital public services, accessing services such as home banking or accessing social networks. The initiative comprises the following activities: - Development and monitoring of a National Network with 950 centres (secondary schools, universities, polytechnics, Private Social Solidarity Institutions, senior universities, Qualifica Centres, among others); - Development and monitoring of a National Network of 10,000 young volunteers; - Programme communication actions (Advertising, events and media), online presence (centres' website, volunteer App and social networks) and overall programme coordination.

Expected benefits: This measure will actively contribute to the training of 1,000,000 info-excluded adults until 2023, in basic digital skills, thus reducing the percentage of the Portuguese population that does not enjoy the benefits of digitalisation in various areas, among which are communications, access to information and the use of digital public services.

7.2 Social Tariff for Access to Internet Services

Creation of a social tariff for access to Internet services, allowing a more generalised use of this resource, to promote inclusion and digital literacy among the most disadvantaged sections of the population.

Consultation and use of digital public services; Access to home banking; E-mail account management Expected Benefits: This measure will actively contribute to promoting inclusion and digital literacy among the most disadvantaged sections of the population and reduce the percentage of citizens who do not use Internet.

7.3 Telework

In the labour environment, teleworking appears, another effect of COVID-19, in which the Government has established new rules, which should update employment contracts in the context of the digital skills required for this purpose. In some sense, digital citizenship was also updated due to the e-government duties.

On the one hand, the general working population had a forced and untrained digital boost to which they had to adapt through resilience and group, or family supports. Even so, initiatives have emerged for online courses promoted for informal training and self-awareness of issues such as computer security (CNCS, 2022).

7.4 Education

Alves e Cabral (2021) suggest that more should be reflected in the learning of those who are far away, than the simple and generous information on the potentiation of distance learning consummated in training offers, tips, and warnings shared in community. Students can carry out active learning using a computer, tablet, or mobile phone. The potential of using games and simulators, augmented reality, the internet of things, didactic robotics, remote laboratories, and artificial intelligence capable of assisting the teacher in chatting, correcting work, etc., is more evident.

However, Koehler, Mishra and Cain (2013) highlight the importance of balancing three dimensions of teacher knowledge for integrating technologies into teaching: content, pedagogy and technology. In this sense also Alves and Cabral (2021) reinforce that the pedagogical project should be strengthened by technology and not replaced by it. These concepts must be considered in order to envisage an opportunity for change, and e-learning not to be seen as an insurmountable obstacle.

The implementation and effective enhancement of e-learning is generally very diversified, being guaranteed by the availability of free technologies or paid platforms provided by the School Publishers. On the one hand, the projects using the free Moodle platform allowed the dissemination and enhancement of distance learning throughout primary and secondary education (CNE, 2021).

Portugal is showing rapid progress in improving baseline qualifications, including STEM education, according to OECD reports, called Steam-IT Project (projeto-steam-it, 2022) whereby bringing together a community of experts, tools and

guidelines will be created, applicable and enhancing the training of future citizens more able to face challenges in society, in a collaborative, critical and efficient way.

In the context of primary and secondary education, the Digital Education Resolution of the Council of

Ministers no. 30/2020 (in Portugal) allows to identify at which level of digital proficiency teachers at the associated schools are and, in this way, better organize the training. Focuses on the different dimensions of school organisation: technological and digital, pedagogical, and organisational.

The training of teachers will take place at two levels: (1) participation in accredited training in digital skills; (2) participation in complementary training and other initiatives, according to the School's strategic plan.

For educational institutions, Law No. 72/2017 (in Portugal) should also be recalled, which promotes the development and generalisation of the dematerialisation of the various educational resources and facilitates the potentiation of e-learning (or teleworking for teachers). And, in the scope of training, the Digitalisation

Programme for Schools, and the Action Plan for Digital Transition, of 21 April 2020 (Resolution of the Council of Ministers no. 30/2020), contemplate the Plan for Digital Qualification of Teachers (PCDD) which aims to ensure the development of digital skills.

The European programme DigCompEdu proposes six areas of competences falling into three major categories: professional competences of educators, pedagogical competences of educators and competences of students. However, it is silent on the training of the population in general, and especially of the elderly. Moreover, no provision is made for the training of monitors to help the elderly overcome digital difficulties and the respective impacts on social equity.

The Research and Intervention Network for Digital Literacy and Inclusion (Rede ObLID) brings together a community of collective and individual agents and is dedicated to Research and Intervention in the field of Citizenship and Digital Participation.

According to (Aires et al., 2018): (1) The instability in the constitution of the teams of monitors is a factor that can put in question the sustainability of the network; (2) In this new mission, the instrumental dimension of technologies is replaced by the technology-mediated human development perspective of citizens; (3) Linking digital inclusion to the development of digital knowledge; Internet access places that integrate the ObLID Network? (4) The network does not appear to be accompanying neither the changes in the digital ecology of the country nor the challenges that have arisen in relation to increasing the digital participation and citizenship of the more vulnerable public and in more remote areas; (5) The improvement of digital skills in citizenship can contribute effectively to a better personal and social development of the individual and is a determinant to achieve equality and social inclusion. (6) As for future training, monitors who are interested in attending training refer the following preferences: Internet protection and privacy, software installation and use, use of portals and public services, hardware maintenance and information technologies skills (e.g., programming languages).

8 Conclusion

This article reviewed the essential concepts related to the effect of the digital transition in the Portuguese population.

Following the multilevel perspective, the COVID-19 crisis became a landscape development that placed pressure on the socio-technical regime. The results of this investigation showed that the pandemic and the policies of ‘stay at home’ opened, in a way, a window of opportunity for a better digital transition for the whole society.

The findings in this article illustrate some of the theoretical and practical implications of digital self-services in the context of an entire population. We have highlighted the several phases that citizens go through as they use the digital services. While these phases do not capture all aspects of the delivery process of digital services, they reveal obstacles and barriers that citizens are likely to face in this process.

There is no one who can keep up with technological evolution at a training level, nor withstand the errors that come with it, either in ordinary life or in professional life situations. And this is independent of the age of each citizen.

Finally, the power of innovation is one of the characteristics of the human being, which in nature in general is manifested by alterations, changes, etc., which we call evolution. Teaching has undergone changes that may be unreturnable, not least because of synchronous digital forms of communication. The use of mobile phones is one of the most promising areas, because it can solve many equity problems. A close collaboration between telecom providers and publishers could give rise to joint products and marketing actions of great added value.

References:

- Aires, L., Santos, R., Guardia, J., Lima, C., & Correia, J. (2018). Mediating towards digital inclusion: the monitors of internet access places. *The Observatory*, 12, 196-213.
- Alvarez, I., Silva, N. & Correia, L. (2015). Cyber Education: towards a pedagogical and heuristic learning. *Computers & Society*, 45(3).
- Alves, J. M., & Cabral, I. (2021). Ensino Remoto de Emergência. Perspetivas Pedagógicas para a Ação. *Serviços de Apoio à Melhoria da Educação*. Faculdade de Educação e Psicologia, Porto.
- Breit, E., Salomon, R. (2015). Making the Technological Transition – Citizens’ Encounters with Digital Pension Services. *Social Policy and Administration*, 49(3), 299-315, ISSN 0144 – 5596 DOI: 10.1111/spol.12093
- Costa, A. F. (2020). Desigualdades Sociais e Pandemia. In Carmo, R.M.; Tavares, I. & Candido, A. F. (Eds), *Um Olhar Sociológico sobre a crise COVID-19*. Observatório das Desigualdades, CIES-ISCTE. <https://doi.org/10.15847/CIESO D2020covid19>

- CNCS (2020). Centro Nacional de Cibersegurança. <https://www.cncs.gov.pt/>
- Dhawan, S. (2020). Online Learning: A Panacea in the Time of COVID-19 Crisis. *Journal of Educational Technology Systems*, 49(1), 5–22.
- EEA (2022), European Education Area. <https://education.ec.europa.eu/focus-topics/improving-quality/about/keycompetences>.
- Ferri, F., Grifoni, P., & Guzzo, T. (2020). Online Learning and Emergency Remote Teaching: Opportunities and Challenges in Emergency Situations. *Societies*, 10(4), 86. <https://doi.org/10.3390/soc10040086>
- Goodfellow, R., & Hewling, A. (2005). Reconceptualising Culture in Virtual Learning Environments: From an 'Essentialist' to a 'Negotiated' Perspective. *E-Learning and Digital Media*, 2(4), 355–367. <https://doi.org/10.2304/elea.2005.2.4.355>
- Irniger, A. (2017, November 29). Difference between Digitization, Digitalization and Digital Transformation. <https://www.coresystems.net/blog/differencebetween-digitization-digitalization-and-digital-transformation>
- Martel, S., Rourke, S., Wade, S., & Watters, M. (2021). Simulating “Normalcy” in a Global Pandemic: Synchronous e-Learning and the Ethics of Care in Teaching. PS: *Political Science & Politics*, 54(1), 173–175. doi:10.1017/S1049096520001493
- Martins, S., Santos, G., Rufato, J., Brito, G. (2020). As Tecnologias na Educação em Tempos de Pandemia. Uma Discussão (Im)pertinente. *Interações. Educação Online em Tempos de Pandemia: Desafios e Oportunidades para Professores e Alunos*, 16(55), 6–27.
- McLaughlin, Jr., C. H. (2021). Book Review: The impact of technology education: International insights. *Journal of Technology Education*, 32(2), 60–67. DOI: <http://doi.org/10.21061/jte.v32i2.a.5>
- OECD. (2020). The impact of COVID-19 on student equity and inclusion: supporting vulnerable students during school closures and school re-openings. *Tackling Coronavirus (Covid-19): Contributions to a global Effort*.
- OECD (2021). Equity in Education after COVID-19: Tackling the challenges ahead. <https://www.oecdforum.org/posts/equity-in-education-after-covid-19-tackling-the-challenges-ahead>
- Pérez-Escolar, M., Canet, F. (2022). Research on vulnerable people and digital inclusion: toward a consolidated taxonomical framework. *Univ Access Inf Soc*. <https://doi.org/10.1007/s10209-022-00867-x>
- Santos, M. I. (2018). A integração de plataformas de e-learning em contexto educativo: Modelo Bietápico de Formação
- Contínua de Professores. [Tese de doutoramento. FPCE, Universidade de Coimbra]. <https://eg.uc.pt/handle/10316/80682>

- Stahl, B., Timmermans, J., & Flick, C. (2017). Ethics of Emerging Information and Communication Technologies: On the implementation of responsible research and innovation. *Science and Public Policy*, 44(3), 369–381, <https://doi.org/10.1093/scipol/scw069>
- UNESCO (2021). Perspectives on the challenges to access and equity in Higher Education across the world in the context of COVID. <https://www.iesalc.unesco.org/wp-content/uploads/2021/09/DRAFT-WAHED-24thSep-2021-2.pdf>.

Securing Meetings in D2D IoT Systems

Sabina Szymoniak¹[0000-0003-1148-5691] and Olga Siedlecka - Lamch¹[0000-0001-9820-6629]

¹ Department of Computer Science, Czestochowa University of Technology, Czestochowa, Poland

sabina.szymoniak@icis.pcz.pl, olga.siedlecka@icis.pcz.pl

Abstract. Cybersecurity is a collection of techniques, methods, and practices that aim to secure different types of Internet resources. Thanks to these techniques, our data should be secure during their transport via the network. One such technique is security protocols. Security protocols secure the transmitted data. Unfortunately, all Internet resources are exposed to the activities of dishonest users. These users are not acting ethically. They try to steal our data, disrupt the operation of applications or servers or turn off these resources. If a dishonest user guessed the key used for messages encryption or decryption, the entire protocol will be compromised. Also, he will know all data transmitted via this protocol. In this paper, we present a new protocol for agreeing on the meetings. This protocol is dedicated to device-to-device (D2D) IoT systems. Our protocol is lightweight and portable. We can use it on various system and hardware platforms.

Keywords: security protocols, cybersecurity, meetings security, sensors

1 Introduction

Every day we have the ability to use the Internet of Things (IoT) technology. This is a technology that focuses on intelligent electronic devices such as cleaning robots, smartwatches or washing machines. Mentioned devices have their computing capability. They can collect data via specially designed sensors. Also, they automatically communicate with each other and exchange data over a network. Such activities can be performed without human intervention (Kamil & Ogundoyin, 2021).

During each network communication, we transmit a lot of data. These data may contain important personal information such as access passwords, bank account numbers or document numbers. We can divide all network users into two groups: honest users and dishonest users. The first group contains these network users that operate ethically and respect all defined security rules. Also, they want that such data remain secure during their travel via a network. The second group contains these network users whose behaviour is unethical. They try to steal and use honest users' data and disturb the operation of applications or servers (Edris et al., 2021).

The communication in the network (both users – devices, and device – devices) executes with short programs called security protocols. The main task of these programs is to secure users' communication. Each protocol consists of a few steps.

During these steps, users can confirm their identity, exchange data or agreed on the session keys. All protocol executions should provide an appropriate level of security. The protocol security depends on the security of the keys used to encrypt and decrypt messages. If such a key will be intercepted, the entire protocol will be compromised. For this purpose, it is possible to verify the protocol correctness via many implemented methods and tools, such as (He et al., 2019), (Siedlecka-Lamch et al., 2019), (Siedlecka, 2020), (Szymoniak, 2021a) (Szymoniak, 2021d) or (Zbrzezny et al., 2020).

We can find many protocols dedicated to sensors and IoT systems. In (Jain, 2020), the authors presented an efficient key distribution protocol. Xu et al. proposed a protocol with a secure authentication mechanism (Xu et al., 2020). In (Ganesh et al., 2021), the authors proposed a protocol for Vehicular Ad Hoc Networks. In ((Szymoniak, 2021a), (Szymoniak, 2021c)), the author presented a protocol that protects users against false links. Khan et al. proposed a protocol of Smart Charging Pile (Khan et al., 2021). Similar protocols we can find in (Ko et al., 2021), (Kwon et al., 2021), (Maheswari et al., 2021) and (Moreno-Cruz et al., 2020).

In this paper, we propose a new security protocol for agreeing on the meetings. This protocol is dedicated to D2D IoT systems, but also it can be used on various system and hardware platforms. The proposed protocol is lightweight and portable.

The rest of this paper is organized as follows. In Section 2, we will provide ethical considerations and implications for sensors and IoT systems. The next Section describes a proposed protocol. In Section 3, we present the experimental results of the proposed protocol verification. The last Section consists of a summary of this paper, our conclusions and future works.

2 Ethical Considerations and Implications

Internet communication always relates to problems with the ethical usage of communication channels. As mentioned in the Introduction, we can find both honest and dishonest users in networks. Considering the activities of dishonest users, we can indicate many situations that aim to steal the data (Karale, 2021).

Dishonest users try to exist in networks in order to intercept and eavesdrop on transmitted data. Many times, they have also the possibility to break intercepted ciphertexts for example using the brute force method. If they achieve success, they can cheat other users, use their identity and also try to steal their personal data. Consequently, honest users may lose their money (Wood, 2021).

Such situations emphasize the role of security protocols. Security protocols must protect our data, identity and whole communication process. Also, they should protect us against dishonest users and provide us that no one can break into our communication. In case of that, it is necessary to systematically verify the protocol's correctness and security.

From an ethical point of view, security protocols implement AAA logic (Authentication, Authorization, Accounting). Mentioned logic refers to a dedicated security structure that mediates access to the network and applications. The AAA logic takes into account three safety aspects:

- the way of user's identification (authentication),
- enforcing of the users' rules (authorization),
- recording of session statistics and user's information (accounting) (Galinec et al., 2019), (Steingartner et al., 2021), (Bartłomiejczyk et al., 2022).

In the era of wide development of intelligent devices usage, securing Internet communication plays a significant role in our life. We should consider new securing methods against unethical activities of dishonest users.

3 Proposed Security Protocol

This section will present a new security protocol for agreeing on the meetings.

3.1 Assumptions

The proposed protocol works in a system composed of two layers. The trusted server works in the first layer. The trusted server has high computing power and resources. The server must generate symmetric keys for devices and also store the credentials of each device. The server must guarantee a secure communication channel for sending authentication keys to devices. Also, the server must provide the authenticity and confidentiality of the transmitted information.

The second layer contains all devices. Devices can communicate with the server and also with each other.

3.2 Security Protocol

The proposed security protocol aims to secure the adjusting of the meeting place and time for several people. The adjusting of the meeting place is based on current users' locations and also their preferences related to the purpose of the meeting or meals. The protocol provides mutual authentication of devices and their users. Also, the protocol transfers information about the geographical location of devices and additional information about users' preferences. Based on this information, protocol suggests the meeting point to users.

We assumed that the number of devices in the system depends on the decision of the user who proposes the meeting. The user who proposes the meeting must execute two activities. The first of them is the establishment of the trusted server. The second activity is the establishment of communication between devices of people with whom he would like to meet. Invited users must answer to the meeting initiator. When he received the answers from them, he can execute calculations and then propose a meeting place. During calculations, the meeting initiator takes into account both the current location of users and their preferences.

We divided the proposed protocol into three stages. The first of them is the preliminary stage. During the execution of this part, all devices must confirm their identity to the server. Also, they must agree on a symmetric key with it. If any device wants to take part in further stages of the protocol, it must first agree on the key with the server. The agreed symmetric keys enable the transfer of information between the devices and the server.

The scheme of the first protocol stage in Alice-Bob notation was presented in Figure 1.

$$\begin{aligned}
 \alpha_1 \quad D &\rightarrow S : \quad \{i(D), T_D\}_{K_S^+} \\
 \alpha_2 \quad S &\rightarrow D : \quad \{K_{DS}, i(D), T_S\}_{K_D^+} \\
 \alpha_3 \quad D &\rightarrow S : \quad \{T_S\}_{K_{DS}}
 \end{aligned}$$

Fig. 1. A scheme of first protocol part in Alice – Bob notation.

In this notation:

- D – means a device that wants to establish a symmetric key with the trusted server,
- S – means a trusted server,
- $\{m\}_k$ – means message m encrypted by key k ,
- $i(u)$ – means an user's identifier, for example, $i(D)$ is identifier of user D ,
- T_u – means user's timestamp, for example, T_D is the timestamp of user D ,
- K_u^+ – means a public key of user u , for example, is a K_S^+ server's public key,
- K_{XY} – means a symmetric key shared between users assigned as X and Y , for example, K_{DS} is a symmetric key shared between user D and a trusted server S .

$$\begin{aligned}
 \alpha_1 \quad D_{MI} &\rightarrow S : \quad \{i(D_{MI}), T_{D_{MI}}, \{i(D_1), i(D_2), \\
 &\quad \dots, i(D_n)\}_{K_{D_{MI}S}}\}_{K_{D_{MI}S}} \\
 \alpha_2 \quad S &\rightarrow D_{MI} : \quad \{T_S, \{i(D_1), K_{D_{MI}D_1}\}_{K_{D_{MI}S}}, \\
 &\quad \{i(D_2), K_{D_{MI}D_2}\}_{K_{D_{MI}S}}, \dots, \\
 &\quad \{i(D_n), K_{D_{MI}D_n}\}_{K_{D_{MI}S}}\}_{K_{D_{MI}S}} \\
 \alpha_3 \quad S &\rightarrow D_i : \quad \{T_S, i(D_{MI}), K_{D_{MI}D_i}\}_{K_{D_iS}} \\
 \alpha_4 \quad D_i &\rightarrow S : \quad \{T_S\}_{K_{D_iS}}
 \end{aligned}$$

Fig. 2. A scheme of second protocol part in Alice – Bob notation.

In the first step, device D wants to authenticate to a trusted server and establish a shared symmetric key with it. In this case, it generates its timestamp and sends it to the server with D 's identifier. D encrypts this message with the server's public key. After receiving this message, The trusted server prepares a symmetric key that will be shared between them and device D . Also, it must generate its timestamp. Next, it sends a message with a new key, timestamp T_s and D 's identifier to D . Next, D must confirm that it knows a new symmetric key. In this case, it sends the server's timestamp to the server in the message encrypted by a new shared key K_{Ds} .

During the second stage, all devices receive a new session key. First, the meeting initiator sends to the server a list of devices. Next, the server generates a session key for current communication for each device. Also, the server prepares a package of messages that includes the device's identifier and a session key. The server sends this package to the meeting initiator. Also, the server sends the keys to individual devices. The rest of the communication will take place between devices only.

The scheme of the second protocol stage in Alice-Bob notation was presented in Figure 2.

$$\begin{aligned}
 \alpha_1 \quad D_{MI} &\rightarrow D_i : \quad \{i(D_{MI}), T_{D_{MI}}, \{GC_{MI}, \\
 &\quad P_{MI}\}_{K_{D_{MI}D_i}}\}_{K_{D_{MI}D_i}} \\
 \alpha_2 \quad D_i &\rightarrow D_{MI} : \quad \{i(D_i), GC_i, P_i\}_{K_{D_{MI}D_i}} \\
 \alpha_3 \quad D_{MI} &\rightarrow D_i : \quad \{i(D_{MI}), GC_i^2\}_{K_{D_{MI}D_i}} \\
 \alpha_4 \quad D_i &\rightarrow D_{MI} : \quad \{i(D_i), T_i\}_{K_{D_{MI}D_i}}
 \end{aligned}$$

Fig. 3. A scheme of third protocol part in Alice – Bob notation.

Figure 3 shows the third protocol stage in Alice-Bob notation. In this stage, devices communicate with each other. In the first step, the initiator sends the meeting proposal to the selected devices. The message sent to each of the devices contains an initiator identifier, initiator timestamp, a message with a meeting proposal. Initiator encrypts this message by session key, which is shared between him and the device. The component message (with meeting proposal) consists of the meeting initiator's geographical coordinates and also his meeting preferences. Also, the initiator encrypts this message by session key, which is shared between him and the device.

In the second step, each device must respond to this proposal. Each device can confirm or reject the proposal. If the proposal is confirmed, the device sends its identifier, geographic coordinates and also optional user preferences. The device encrypts this message by session key, which is shared between him and the device. If the proposal is rejected, the device sends back the geographic coordinates of the initiator's device to him. In both cases, the device encrypts this message by session key, which is shared between him and the device. The devices that rejected the meeting proposal do not participate in the rest of the protocol.

After the third step, the initiator's device must calculate the potential meeting points based on the information obtained (geographic coordinates and users' preferences). Next, the initiator sends to each device the proposed meeting (geographic coordinates) with his identifier. Also, the initiator encrypts this message by session key, which is shared between him and the device.

In the fourth step, devices must accept or decline the proposal again. If the device's user confirms the proposal, the device must generate the device's timestamp. Next, the device sends this timestamp with the device's identifier to the meeting initiator. If the device's user rejects the proposal, the device prepares a message with the device's identifier and with the meeting initiator's timestamp. Next, the device sends this message to the meeting initiator. Also, in both cases, the device encrypts this message by session key, which is shared between him and the device.

The calculations of potential meeting places will take into account the geographic coordinates of devices, the purpose of the meeting and user preferences. The initiator's device must propose an appointment at a location that is as far as possible from the location where each device is located. If there will be a need for seating for the meeting in a non-smoking place or with a specific kitchen, the algorithm should take into account the individual preferences and tastes of users too.

4 Experimental Results

We conducted a preliminary security study of our protocol. We used the model and tools described in (Szymoniak, 2021). According to the mentioned methodology, we analyzed the times of protocol execution and time of protocol simulations in a computer network. Our analysis took into account the presence of a dishonest user called an Intruder, who wants to intercept the confidential data of other users (Dolev et al, 1983). We performed our analysis using a computer with Linux Ubuntu operating system, Intel Core i5 processor and 16 GB RAM. We used an abstract time unit ([tu]) to determine the time.

First, we assumed that the Intruder could impersonate only honest participants in the protocol, not the trusted server. Next, the tool combinatorically generated executions for each part of the proposed protocol. For the first part, there were four executions:

- honest execution (between D and S),
- execution with an Intruder instead of D (he uses D 's identity and his cryptographic objects during communication),
- two executions where the Intruder impersonates D (he uses D 's identity and cryptographic objects along with his own during communication).

Next, we assumed values of following timed parameters for so-called timed analysis of protocols executions times:

- $D_{min} = 1$ [tu] (minimal delay in the network),
- $D_{max} = 4$ [tu] (maximal delay in the network),
- $T_e = 4$ [tu] (encryption time for symmetric algorithm),
- $T_d = 4$ [tu] (decryption time for symmetric algorithm),

- $T_e = 6$ [tu] (encryption time for asymmetric algorithm),
- $T_d = 6$ [tu] (decryption time for asymmetric algorithm), $T_g = 2$ [tu] (time of generating confidential information), $T_c = 1$ [tu] (time of composing the message).

After that, the tool calculated lifetimes for each step: 63 [tu] (1st step), 34 [tu] (2nd step), 13 [tu] (3rd step). Also, the tool calculated the minimal (44 [tu]) and maximal (63 [tu]) session times for this protocol part. Next, the tool executed the timed analysis of executions times.

Table 1. Summary of timed analysis of first proposed protocol part.

Step	T_C	T_G	T_E	D	T_D	T_{Smin}	T_{Smax}
α_1	1	2	6	<1,4>	6	16	19
α_2	1	4	6	<1,4>	6	18	21
α_3	1	0	4	<1,4>	4	10	13

Table 1 presented a summary of timed analysis for the first proposed protocol part. We assigned all time parameters which are included in individual steps and also minimal and maximal step time. For delay in the network parameter, we assigned a range of values. The timed analysis was executed using $D = 1$ [tu].

We observed that only the first two executions were possible executed. The rest two executions were impossible to execute. The Intruder did not have enough time and opportunities to acquire the appropriate knowledge to execute them. This means that on the first part of our protocol, there is no possible attack. This protocol part is secure.

To simplify the result presentation, for the second part of our protocol, we assumed that the meeting initiator tries to connect with one device only. For the second protocol part there were six executions:

- honest execution (between DMI , S and D),
- execution in which an Intruder instead of D or DMI (he uses D 's or DMI 's identities and his cryptographic objects during communication),
- three executions where the Intruder impersonates D or DMI (he uses D 's or DMI 's identities and cryptographic objects along with his own during communication).

We assumed identical values of timed parameters as in the first protocol part. Then, the tool calculated lifetimes for each step: 72 [tu] (1st step), 51 [tu] (2nd step), 28 [tu] (3rd step), 13 [tu] (4th step). Also, the tool calculated the minimal (60 [tu]) and maximal (72 [tu]) session times for this protocol part. After that, the tool executed the timed analysis of executions times. Table 2 presented a summary of timed analysis for the second proposed protocol part.

Table 2. Summary of timed analysis of second proposed protocol part.

Step	T_C	T_G	T_E	D	T_D	T_{Smin}	T_{Smax}
α_1	1	0	4 + 4	<1,4>	4 + 4	18	21

α_2	1	2	4 + 4	<1,4>	4 + 4	20	23
α_3	1	2	4	<1,4>	4	12	15
α_4	1	0	4	<1,4>	4	10	13

We observed that only the first three executions were possible to execute. The rest three executions were impossible to execute. The Intruder did not have enough time and opportunities to acquire the appropriate knowledge to execute them. This means that on the second part of our protocol, there is no possible attack. This protocol part is secure.

Next, we perform an analysis of the third protocol part. For this part, the tool generated nine following executions:

- honest execution (between *DMI* and *D*),
- execution in which an Intruder instead of *D* (he uses *D*'s identity and his cryptographic objects during communication),
- execution in which an Intruder instead of *DMI* (he uses *DMI*'s identity and his cryptographic objects during communication),
- three executions in which an Intruder impersonates *D* (he uses *D*'s identity and cryptographic objects along with his own during communication),
- three executions in which an Intruder impersonates *DMI* (he uses *DMI*'s identity and cryptographic objects along with his own during communication),

We assumed identical values of timed parameters as in the first and second protocol part. Then, the tool calculated lifetimes for each step: 64 [tu] (1st step), 41 [tu] (2nd step), 28 [tu] (3rd step), 15 [tu] (4th step). Also, the tool calculated the minimal (52 [tu]) and maximal (64 [tu]) session times for this protocol part. After that, the tool executed the timed analysis of executions times. Table 3 presented a summary of timed analysis for the third proposed protocol part.

Table 3. Summary of timed analysis of third proposed protocol part.

Step	T_C	T_G	T_E	D	T_D	T_{Smin}	T_{Smax}
α_1	1	2	4 + 4	<1,4>	4 + 4	20	23
α_2	1	0	4	<1,4>	4	10	13
α_3	1	0	4	<1,4>	4	10	13
α_4	1	2	4	<1,4>	4	12	15

We observed that only the first three executions were possible to execute. The rest six executions were impossible to execute. The Intruder did not have enough time and opportunities to acquire the appropriate knowledge to execute them. This means that on the third part of our protocol, there is no possible attack. Also, this protocol part is secure.

5 Conclusions

Internet communication always relates to problems with the ethical usage of communication channels. All network users are exposed to dishonest activities like intercepting and eavesdropping on transmitted data. Procedures and methods securing electronic communication are a significant stage in our communication via the network. Thanks to them, our data are protected against unfair and unethical activities. Also, we can feel secure. If such methods provide an appropriate security level, any dishonest user cannot steal data or compromise the protocol. Security protocols are one of the methods that aim to secure data.

In this paper, we presented a new security protocol. We dedicated this protocol to D2D IoT systems. The proposed protocol is lightweight and portable. It can be used on various system and hardware platforms. Our protocol consists of three stages (initial, session key establishment and communication stage). In the first part, users and their devices can authenticate with a trusted server. In the second part, the meeting initiator handshakes the session keys for all devices. In the last stage, devices communicate with the meeting initiator to meeting arrangements.

We executed a preliminary verification of the proposed protocol. We provided test series with a timed analysis of our protocol. The obtained results are promising. We did not find an attack for our protocol. In future work, we will test our protocol using other tools for verification of security protocol. Also, we will try to implement this protocol in a real IoT system, including the implementation of a lightweight algorithm for session keys generation.

Acknowledgements

The project financed under the program of the Polish Minister of Science and Higher Education under the name "Regional Initiative of Excellence" in the years 2019-2022 project number 020/RID/2018/19, the amount of financing 12,000,000.00 PLN.

References

- Bartłomiejczyk, M., Fray, I. E., Kurkowski, M., Szymoniak, S., & Siedlecka-Lamch, O. (2022). User Authentication Protocol Based on the Location Factor for a Mobile Environment (Vol. 10). Institute of Electrical and Electronics Engineers (IEEE). <https://doi.org/10.1109/access.2022.3148537>
- Dolev, D. & Yao, A. (1983). On the security of public key protocols. In: IEEE Transactions on Information Theory, 29(2).
- Edris E.K.K., Aiash M., Loo J. (2021). Security in Network Services Delivery for 5G Enabled D2D Communications: Challenges and Solutions. In: Montasari R., Jahankhani H., Al-Khateeb H. (eds) Challenges in the IoT and Smart Environments. Advanced Sciences and Technologies for Security Applications. Springer, Cham. https://doi.org/10.1007/978-3-030-87166-6_1

- Galinec, D., Steingartner, W., & Zebic, V. (2019). Cyber Rapid Response Team: An Option within Hybrid Threats. 2019 IEEE 15th International Scientific Conference on Informatics. <https://doi.org/10.1109/informatics47936.2019.9119292>
- Ganesh, A., Ayyasamy, S., & Kumar, N. M. S. (2021). Performance and Analysis of Advanced and Enhanced Security Protocol for Vehicular Ad Hoc Networks (VANETs). *Wireless Personal Communications*, 121(4), 3163–3183. <https://doi.org/10.1007/s11277-021-08868-4>
- He, X., Liu, J., Huang, C. T., Wang, D., & Meng, B. (2019). A Security Analysis Method of Security Protocol Implementation Based on Unpurified Security Protocol Trace and Security Protocol Implementation Ontology. *IEEE Access*, 7, 131050–131067. <https://doi.org/10.1109/access.2019.2940512>
- Jain, P. (2020). “Sec-keyd” an efficient key distribution protocol for critical infrastructures. *CSI Transactions on ICT*, 8(4), 385–394. <https://doi.org/10.1007/s40012-020-00314-3>
- Kamil, I. and Ogundoyin, S., 2021. A lightweight mutual authentication and key agreement protocol for remote surgery application in Tactile Internet environment. *Computer Communications*, 170, pp.1-18.
- Karale, A. (2021). The Challenges of IoT Addressing Security, Ethics, Privacy, and Laws. *Internet of Things*, 15, 100420. <https://doi.org/10.1016/j.iot.2021.100420>
- Khan, S., Alvi, A. N., Javed, M. A., Al-Otaibi, Y. D., & Bashir, A. K. (2021). An efficient medium access control protocol for RF Energy harvesting based IOT devices. *Computer Communications*, 171, 28–38. <https://doi.org/10.1016/j.comcom.2021.02.011>
- Ko, Y., Kim, J., Duguma, D.G., Astillo, P.V., You, I. & Pau, G. (2021). Drone Secure Communication Protocol for Future Sensitive Applications in Military Zone. *Sensors*.
- Kwon, D.K., Yu, S.J., Lee, J.Y., Son, S.H. & Park, Y.H. (2021). WSN-SLAP: Secure and Lightweight Mutual Authentication Protocol for Wireless Sensor Networks. *Sensors*.
- Maheswari, M., & Karthika, R.A. (2021). A Novel QoS Based Secure Unequal Clustering Protocol with Intrusion Detection System in Wireless Sensor Networks. *Wireless Pers Commun*.
- Moreno-Cruz, F., Toral-López, V., Escobar-Molero, A., Ruíz, V.U., Rivadeneyra, A. & Morales, D.P. (2020). treNch: Ultra-Low Power Wireless Communication Protocol for IoT and Energy Harvesting. *Sensors*, 20, 6156.
- Siedlecka-Lamch, O. (2020). Probabilistic and timed analysis of security protocols, In proceeding of the 13th International Conference on Computational Intelligence in Security for Information Systems CISIS 2020, 16-18 September 2020, Burgos, Spain; paper 24.
- Siedlecka-Lamch, O., Szymoniak, S. & Kurkowski, M. (2019) A fast method for security protocols verification, *Computer Information Systems and Industrial Management*, Springer.
- Steingartner, W., Galinec, D., & Kozina, A. (2021). Threat Defense: Cyber Deception Approach and Education for Resilience in Hybrid Threats Model. *Symmetry*, 13(4), 597. <https://doi.org/10.3390/sym13040597>

- Szymoniak, S. (2021a). Amelia—a new security protocol for protection against false links. *Computer Communications*, 179, 73–81.
<https://doi.org/10.1016/j.comcom.2021.07.030>
- Szymoniak S. (2021b). Time Influence on Security Protocol, *Proceedings of the 16th International Conference on Evaluation of Novel Approaches to Software Engineering, ENASE 2021*, p. 181-188.
- Szymoniak, S. (2021c). Using A Security Protocol To Protect Against False Links, *Moving technology ethics at the forefront of society, organisations and governments ETHICOMP Book Series*, Eds. Jorge Pelegrín Borondo, Mario Arias Oliva, Kiyoshi Murata, Ana María Lara Palma, p. 513-525.
- Szymoniak, S. (2021d). Security protocols analysis including various time parameters. *Mathematical Biosciences and Engineering*, 18(2), 1136–1153.
<https://doi.org/10.3934/mbe.2021061>
- Wood P. (2021). Socio-technical Security: User Behaviour, Profiling and Modelling and Privacy by Design. In: Montasari R., Jahankhani H., Al-Khateeb H. (eds) *Challenges in the IoT and Smart Environments. Advanced Sciences and Technologies for Security Applications*. Springer, Cham. https://doi.org/10.1007/978-3-030-87166-6_4
- Xu, J., Yu, X., Tian, L., Wang, J., & Liu, X. (2020). Security Analysis and protection for charging protocol of smart charging pile. *Proceedings of the 9th International Conference on Computer Engineering and Networks*, 963–970
https://doi.org/10.1007/978-981-15-3753-0_95
- Zbrzezny, A. M., Zbrzezny, A., Szymoniak, S., Siedlecka-Lamch, O., Kurkowski, M., (2020). VerSecTis - An Agent based Model Checker for Security Protocols, *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020)*, B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand, p. 2123-2125.

Intelligent Identity Authentication, Using Face and Behavior Analysis

Mariusz Kubanek^{1[0000-0001-9651-9525]}, Janusz Bobulski^{1[0000-0003-3345-604X]} and Łukasz Karbowski^{1[0000-0003-4389-5548]}

¹ Department of Computer and Information Sciences, Czestochowa University of Technology, Dabrowskiego Street 69, Czestochowa 42-201, Poland
mariusz.kubanek@icis.pcz.pl

Abstract. User identification is one of the most critical issues in the digital world today. Currently, we are participants in a continuous authentication process at almost every step. Whether it is online banking or online sales, our digital profile is responsible for the correctness of transactions, and the adequate protection of our data is responsible for security. There are many methods of adequate identity verification, but such methods are not always applicable and are not always completely secure. We have proposed an intelligent identity authentication method that uses the image of the face and specific user behavior. We related the user's particular behavior to the simultaneous analysis of the facial appearance and the specific position of the hands within the camera's field of view. Additionally, we used the Google Authenticator to increase the security of logging in to protected resources. The Google Authenticator tool allows us to generate one-time access codes, one of the final authentication levels. The entire system is resistant to intrusion attempts by using the user's photo, and the obtained results confirm the high level of protection of sensitive data.

Keywords: identity authentication, behavioral analysis, cybersecurity

1 Introduction

Nowadays, the identification and verification of people is essential and is of great importance for society because the elements of identity in the form of identity cards, passwords, or even a social security number are not sufficient. Each form of identification or verification can be stolen, falsified, or forgotten. Therefore, reliable identification and authentication become essential. Reliable identification of people is associated with the use of biometrics.

Biometrics is the study of identifying and verifying people based on specific physical and behavioral characteristics. These features build a biometric system based on fingerprints, hand geometry, or signature. The purpose of the biometric system is to replace the previously used forms of identification (social security cards, passwords). Biometric systems are more effective and efficient than currently used identification and verification methods, for example, in the case of banking

transactions where there is a risk of card forgery or PIN interception. In this case, using a biometric system carries a lower risk of fraud.

Although biometric systems are better than the currently used methods of adequate identification and verification, some have disadvantages, e.g., a given quality is not present in all people. Another major drawback of biometric systems is that a given biometric cannot be measured and requires complex and expensive equipment.

The fact is that highly effective identification and user verification methods are based on biometric features. It can be distinguished elements fetched based on the iris image (Omidiora et al., 2015) and features contained in the hand (Chandana et al., 2015). All of them require some user-system cooperation or a person to measure it. A non-invasive example of a biometric system can be a system that executes the verification and identification based on facial features.

Face recognition methods are used in systems requiring data security mainly because of non-invasiveness. A user does not usually take part in the process of data consumption and often does not need to know about that. Face recognition can be defined as searching for all faces in a still image or a video sequence. The identification or the authentication can be made by comparing found faces with those in the database.

The facial recognition system must be fully automatic. Such a face recognition system combines face localization and extraction of structural characteristics of a face, such as eyes and nose shape and size, face shape, face color, and other distinguishable features, and holistic methods. By using the structural characteristics of the face, it is possible to search for new ways of identification and verification of identity. Dependencies used as the primary recognition factor are imperceptible at first; these are, e.g., features associated with facial asymmetry (Bobulski, 2017, Mitra et al., 2007). These methods use differences in geometrical features of a look. During extraction of these features, the important thing is the choice of asymmetry line, and there are few ways to do this. One way is to mark the middle of the line between the corners of the eyes. Next, the asymmetry measurement is calculated and used for identification.

An alternative to all of these defects in fingerprint, iris, and facial imaging biometric systems is the use of a hand blood vessel model. Biometrics of blood vessels is present in all people and is unique to each person. The feature recognition process does not require collaboration or complex devices. Moreover, the venous system does not change throughout life apart from its size. Using a blood vessel biometric system means it's hard to fake. Unfortunately, data acquisition equipment is not widely used, limiting the scope of such methods' operation.

It must be primarily ethical to act when using human biometric features for authentication. Because features that unambiguously identify a given person are related to the description of parts of their body, their collection and use violate personal autonomy and human dignity. The main moral problem is the registration phase, i.e., creating and storing a unique biometric template identifying a specific person. Creating biometric templates means converting a person's unique physical characteristics into digital data, leading to ratification, the tendency to develop more and more data, and to rely on algorithms. Because features that unambiguously

identify a given person are related to the description of parts of their body, their collection and use violate personal autonomy and human dignity.

If the trait templates are stored, anyone who obtains them in the future will have the ability to track and recognize an individual anywhere in the world and potentially for any purpose. The entity cannot protect itself from this as it cannot change its biometric identifiers. By also considering data collection and storage security issues, biometric templates have a significant potential for harm. In addition, ethical issues related to using biometric identification methods in public spaces may relate to large-scale human surveillance.

Ethical issues may relate to the specific purpose of the categorization, the degree of sensitivity of the collected data and the conclusions are drawn, the system's accuracy, human surveillance, the potential irreversibility of system effects and storage, and further data processing. On the other hand, in the case of biometric categorization (e.g., assignment to a risk group in the airport security system or evaluation of job applicants), ethical issues related to the definition of the category, related assumptions, and conclusions, or system responses lead to discrimination and stigmatization of individuals. Other threats include manipulation and exploitation of vulnerabilities.

The problem of data identification and retention should be well regulated by law. Therefore, meeting the requirements related to ethics should be the primary goal of creating such systems. Accordingly, our proposal assumes only the user's authorization who voluntarily wants to use modern methods.

We proposed an intelligent identity authentication method that uses the facial image and specific user behavior. We linked the user's particular behavior with the simultaneous analysis of the appearance of the face and the particular positioning of the hands in the field of view of the camera. In addition, we used Google Authentication to increase the security of the login process to protect resources. The Google Authenticator tool allows us to generate one-time access codes, one of the final levels of authentication. Our method provides for intelligent identity authentication without the need to use personal characteristics that may be considered unethical in many cultures.

2 The structure of the proposed method

2.1 Related work

By analyzing the research conducted, one can find identity verification systems based on various characteristics perceived as unique, unchangeable, and unmistakable. However, in almost all of these systems, the verification data can be falsified, or the system's cost is too high. A secure system should check whether the verification applies to a specific person, i.e., it can, for example, analyze the distribution of the hand's blood vessels and thus check whether the hand is a natural, living hand, or just a photo. An important aspect of secure identity verification methods is making the system immune to hacking attempts, for example, using a picture of the verified

person. It is equally essential to ensure the security of verification systems where the authentication process occurs remotely. An important aspect here is the security of communication protocols (Szymoniak, 2021a, Szymoniak 2021b, Szymoniak 2021c), allowing verification processes in electronic banking systems.

Due to their reliability, methods that consider the arrangement of blood vessels are of great interest, as mentioned before. In the works (Park, 2011, Al-Juboori et al., 2014), one can find a description of the construction of identity verification systems based on the vascular distribution of the hand, which uses the wavelet transform to extract the features. The papers (Miura et al., 2005, Zhou et al., 2014) describe Gaussian functions to remove vessel features from the hand. In many studies, various methods of hand veins have been proposed e.g. (Kumar et al., 2009, Yuksel et al., 2011, Zhou et al., 2011). The hand blood vessel pattern is similar to a fingerprint. This similarity can be seen in the work (Wang et al., 2008), which uses trifles to code the system of hand blood vessels.

Many different solutions work pretty well with authentication. However, only assessing a personality trait and its truthfulness can guarantee the system's proper functioning. Hence we propose to use multi-factor authentication to analyze the test person's behavior.

2.2 The Convolutional Neural Networks

We used Convolutional Neural Networks (CNN) for facial recognition and hand gestures to authenticate the identity. Convolutional Neural Networks are artificial neural networks designed to classify visual objects based on their observed shape. Convolutional Neural Networks consider all recognizable features such as corners and fine edges. Then the network, by analyzing the mutual position of the detailed elements, tries to create more complex shapes, up to the complete forms of recognizable objects. Convolutional Neural Networks proved to be effective, especially in recognizing objects in the image (Wang et al., 2015, Zhong-Qiu et al., 2019, Kubanek et al., 2019). Convolutional neural networks can be an effective tool for intelligent identity authentication due to the overuse of pattern recognition. Convolutional Neural Networks explicitly assume that the inputs are images and reflect them in their architecture, suggesting that even small attempts at fraudulent authentication contain anomalies.

Convolutional Neural Networks usually consist of convolution layers, pool layers, and a fully interconnected layer. The convolution and pool layers are stacked on top of each other, and the fully connected layers at the end of the network generate class probabilities. A convolutional neural network consists of neurons with learnable weights and biases as a feed-forward artificial neural network. Network neurons still contain activation functions, and the whole network expresses a single differentiable score function. The position of the pixel matters in comparison with a multilayer perceptron. It receives 3-dimensional space input. The convolutional and pooling layers are locally connected to the previous layer's outputs, recognizing or magnifying local patterns in the image. The pooling layer is usually put after the convolutional layer. This pair of layers are repeatedly stacked upon each other, followed by the fully

connected layers at the top. The positions of neurons in the previous layer are always considered without stride as inputs for the last layers. The fully connected layer is connected to all outputs of the final pooling layer. The results of the last pooling layer should already represent complex structures and shapes. The fully connected layer usually follows another or two layers, finally outputting the class scores.

2.3 Applied neural network

To use the method of recognizing the user's face, it was necessary to select a neural network that would meet our expectations quickly and efficiently. We used a particular type of Convolutional Neural Network as it is the best proposition for image recognition. We chose the Large version of MobileNetV3.

Mobile models have been built on increasingly more efficient building blocks. MobileNetV1 introduced depthwise separable convolutions as an efficient replacement for traditional convolution layers. Depthwise separable convolutions effectively factorize conventional convolution by separating spatial filtering from the feature generation mechanism. Two layers define Depthwise separable convolutions: lightweight depthwise convolution for spatial filtering and heavier 1x1 pointwise convolutions for feature generation (Howard et al., 2017, Howard et al., 2019).

MobileNetV2 introduced the linear bottleneck and inverted residual structure to make even more efficient layer structures by leveraging the low-rank nature of the problem. This structure is defined by a 1x1 expansion convolution followed by depthwise convolutions and a 1x1 projection layer. This structure maintains a compact representation of the information and the output while expanding to a higherdimensional feature space internally to increase the expressiveness of nonlinear per channel transformations. The input and output are connected with a residual connection if they have the same number of channels (Howard et al., 2017, Howard et al., 2019).

MnasNet built upon the MobileNetV2 structure by introducing lightweight attention modules based on squeeze and excitation into the bottleneck structure. The module is placed after the depthwise filters in the expansion for attention to the most significant representation. Note that the squeeze and excitation modules are integrated into a different location than ResNet-based modules (Howard et al., 2017, Howard et al., 2019).

MobileNetV3 used a combination of these layers as building blocks to build the most effective models. Layers are also upgraded with modified "swish" nonlinearities. Both squeeze and excitation and the "swish" nonlinearity use the sigmoid, which can be inefficient to compute and challenge maintaining accuracy in fixed-point arithmetic. Hence, it was replaced with the hard sigmoid (Howard et al., 2017, Howard et al., 2019). The construction of the MobileNetV3 network is shown in Figure 1.

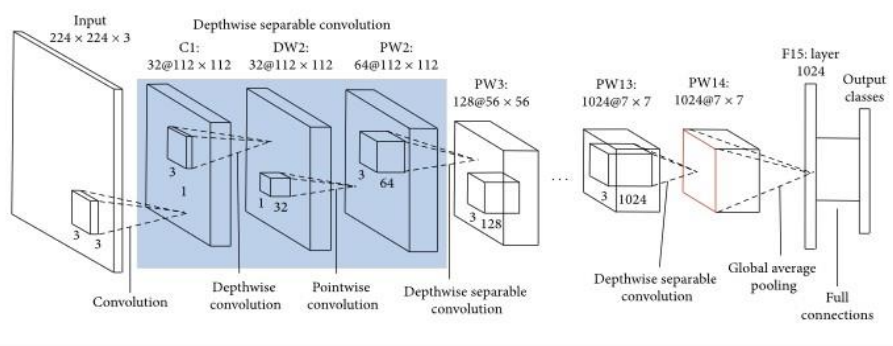


Fig. 1. The construction of the MobileNetV3 network (Pujara, 2020)

2.4 The structure of the proposed method

The viability of the verification feature can also be checked by analyzing the specific behavior of the test subject. For example, you can see if a person who wants to verify their identity based on an image of their face provides a vivid picture of their face where they naturally move their eyeballs and blink occasionally. While recognizing only the user's face, it is possible to falsify by using a photo; tracking the movement of the eyeballs or the mouth allows eliminating this type of fraud attempt. This type of approach requires tools that enable tracking selected image elements in real-time. You can also check other characteristics that indicate that the system is dealing with a live user. We have proposed a method that takes a face image and a specific hand position within the face image during user registration (See Fig. 2.).

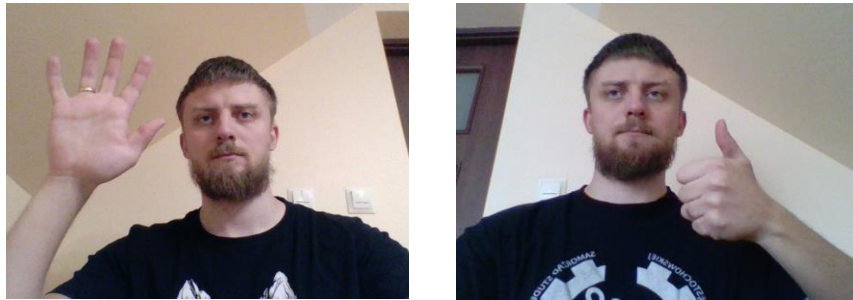


Fig. 2. An example of the face and hand analysis in the user registration process

Tracking the hand's position can be carried out by tracing selected points on the hand. In contrast, the recognition of a hand gesture, unique to the tested user, can be made by training the neural network to recognize the entire scene. In this case, the background must be devoid of other people who may interfere with the learning and verification process. Of course, the web should not learn only the appearance of the user's face and a specific gesture in a stationary position, so it is necessary to train the sequence of hand positioning with the user's face.

After registering the face and hands, the system generates a QR code, used for a twostep identity verification process. After recording the user, you can proceed to the stage of identity verification. The system recognizes faces and, at the same time, asks the user to arrange their hands according to the arrangement of the registration process. If there is a correct match, then in the next step, the system asks for the code generated by Google Authenticator for the user assigned to this identifier (see fig. 3). Providing the correct code confirms the identity and provides the user with protected resources.

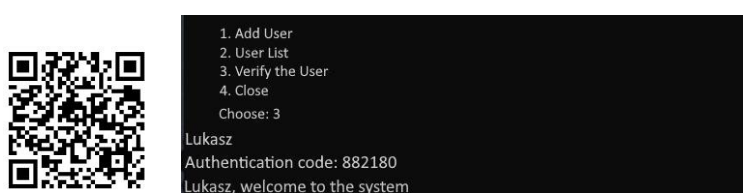


Fig. 3. Generated QR codes for a registered user to access Google Authenticator temporary codes.

3 Research

We chose a group of 40 of different ages, nationalities, and skin colors to conduct the research reliably. The study was divided into three stages. We tested the user verification capabilities using only the face in the first stage. Of course, we assumed the possibility of using user photos for identity authentication. In the second stage, apart from recognizing the user's face, we also added static hand gesture recognition. A static hand gesture means that we realize only one scene with a visible face and hand. We also assumed the possibility of using user photos for identity authentication at this stage. An attempted hacking into such a system may involve suspecting a user's handprint registered in the design, taking a fake photo with the user's face and hand, and then using such a photo for authentication. In the third stage, we assumed that the system is to recognize the sequence of pictures within the scene with the tested user. At this stage, placing photos of the set was completely unacceptable. One could only try to put a mask with a printed image of the user's face on a person trying to break into the system and read a specific hand gesture.

The use of Google Authenticator makes sense only for the third stage of research, so we performed additional tests only for this stage. Of course, it is always possible to suspect the code generated for the tested user. Still, we assumed that the user would not allow any stranger to read the unique number during the identity verification process. After all, if a user does not take care of his protected resources, then no system will be safe for such a user. We have determined the accuracy and the False Acceptance Rate (FAR) and False Rejection Rate (FRR). The results of the experiment are shown in Table 1.

Table 1. The result of the authentication process for all stages.

Stages	Users/Tests	Accuracy [%]	FAR [%]	FRR [%]
Stage 1	40/2000	91,27	23,74	0,03
Stage 2	40/2000	93,95	17,23	1,02
Stage 3	40/860	99,99	3,03	1,26
Stage 3 + Google Authenticator	40/860	99,99	0,01	3,07

Each registered user made multiple attempts to log in to their account and the version of other selected users. Despite only 40 registered users, we performed over 2,000 tests. Based on the results obtained, it can be concluded that our proposal is a valuable tool for secure identity authentication. In the first stage, we received an accuracy of 91.27%. This is quite a good result, but the analysis of the FAR and FRR indicators showed that using the user's photo allows for a reasonably easy hacking into protected resources, and there are few false rejections. In the case of the second stage of the research, we also obtained a sufficiently high level of accuracy (93.95%), but also, in this case, the attempt to break into the system was quite simple. It was only necessary to assemble the appropriate photos, and the system allowed for false acceptance. The third stage of the research has shown that using hand positioning within the camera's field of view, along with the face recognition stage, is a very effective method for intrusion attempts. The false rejection rate increased, but not enough to cause significant problems for users who want to verify themselves correctly.

On the other hand, the false acceptance rate dropped significantly, making the cases of hacking into his account sporadic and resulting mainly from poor lighting of the user's face. Combining the third stage with the Google Authenticator tool eliminated any possibility of breaking into the system. It is safe to say that multi-step intelligent authentication is a genuinely secure system.

4 Conclusion

The method of intelligent identity verification we proposed turned out to be very safe and effective. First of all, during the tests, it was not possible to break into the system for stage 3 with the Google Authenticator tool used. Currently, there are approximately 40 users registered in the system, but we will gradually increase this number. Only a significant increase in the number of users will allow us to answer the question of how safe our method is. Considering the obtained results, we can be expected that the result will be good. The system turned out to be resistant to hacking attempts, mainly when an unauthorized person tried to enter the system by presenting

the user's photo. Our system immediately detected a forgery attempt. We plan to include artificial intelligence algorithms that analyze the user's facial expressions in further work. This other approach will increase the system's protection and provide additional features, specific and individual within the registered user.

Acknowledgements

The project financed under the program of the Polish Minister of Science and Higher Education under the name "Regional Initiative of Excellence" in the years 2019{2022 project number 020/RID/2018/19, the amount of financing 12,000,000 PLN.

References

- Al-Juboori, A.M., Bu, W. Wu, X., Zhao, Q. (2014). Palm vein verification using multiple features and locality preserving projections. *The Scientific World Journal*, vol. 2014
- Bobulski J. (2017). Face recognition with 3D face asymmetry. *Advances in Intelligent Systems and Computing*, 525, pp. 5360
- Chandana, C., Surendra, Y., Manish, M. (2015). Fingerprint Recognition based on Minutiae Information. *International Journal of Computers and Applications* 120(10), pp. 39-42
- Howard A., Sandler M., Chu G., Chen L., Chen B., Tan M., Wang W., Zhu Y., Pang R., Vasudevan V., Le Q., Hartwig A. (2019). Searching for MobileNetV3, *Computer Vision and Pattern Recognition*, arXiv:1905.02244, pp. 11
- Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Hartwig, A. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint arXiv:1704.04861*, pp. 9
- Kubanek, M., Bobulski, J., Kulawik, J. (2019). A method of speech coding for speech recognition using a convolutional neural network. *Symmetry*, vol. 11, number 9, 1185
- Kumar, A., Prathyusha, K. V. (2009). Personal authentication using hand vein triangulation and knuckle shape. *IEEE Transactions on Image Processing*, vol. 18, pp. 2127–2136
- Mitra, S., Lazar, N.A., Liu, Y. (2007). Understanding the Role of Facial Asymmetry in Human Face Identification. *Journal Statistics and Computing*, 17 (1), pp. 57-70
- Miura, N., Nagasaka, A., Miyatake, T. (2005). Extraction of finger vein patterns using maximum curvature points in image profiles. *Conference on Machine Vision Application*
- Omidiora, E. O., Adegoke, B. O., Falohun, S. A., Ojo, D. A. (2015). Iris Recognition Systems: Technical Overview. *International Journal of Research in Engineering & Technology*, Vol. 3, Issue 6, pp. 63-72
- Park, K. R. (2011). Finger vein recognition by combining global and local features based on SVM, *Computing and Informatics*, vol. 30, pp. 295–309
- Pujara A. (2020). Image Classification With MobileNet. *Analytics Vidhya*

- Szymoniak, S. (2021a). Amelia—a new security protocol for protection against false links. *Computer Communications*, 179, pp. 73–81
- Szymoniak, S. (2021b). Using A Security Protocol To Protect Against False Links, Moving technology ethics at the forefront of society, organizations, and governments
ETHICOMP Book Series, Eds. Jorge Pelegrín Borondo, Mario Arias Oliva, Kiyoshi Murata, Ana María Lara Palma, pp. 513-525
- Szymoniak, S. (2021c). Security protocols analysis, including various time parameters. *Mathematical Biosciences and Engineering*, 18(2), pp. 1136–1153
- Wang, L., Graham, G., Cho, D.S.Y. (2008). Minutiae feature analysis for infrared hand vein pattern biometrics. *The Journal of The Pattern Recognition Society*, vol.41, pp.920–929
- Wang, L., Ouyang, W., Wang, X., Lu, H. (2015). Visual tracking with fully convolutional networks. *Proceedings of International Conference on Computer Vision (ICCV)*, pp. 3119-3127
- Yuksel, A., Akarun, L., Sankur, B. (2011). Hand vein biometry is based on geometry and appearance method. *Computer Vision*, vol. 5, pp. 398–406
- Zhong-Qiu, Z., Peng, Z., Shou-tao, X., Xindong, W. (2019). Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1-21
- Zhou, Y., Kumar, A., (2011). Human identification using palm-vein images. *IEEE Transactions on Information Forensics and Security*, vol. 6, pp. 1259–1274
- Zhou, Y., Liu, Y., Feng, Q., Yang, F., Huang, J., Nie, Y. (2014). Palm-vein classification based on principal orientation features. *PLOS ONE* 9(12)

Zero Trust Model and the Shift in the Ethos of Cybersecurity – Towards the Deleuzian Society of Control

Jukka Vuorinen¹ and Ville Uusitupa²

¹ University of Jyväskylä, Jyväskylä, Finland

² University of Jyväskylä, Jyväskylä, Finland

jukka.a.vuorinen@jyu.fi

Abstract. In this paper, we analyse a shift in ethos that has taken place in terms of cybersecurity models. The conventional network perimeter model has been replaced by more and more fluid models such as Zero Trust Architecture, in which the traditional division into the trusted inside of network and untrusted outside has been blurred. We examine the features of the two models and compare them to Michel Foucault's disciplinary society and Deleuze's Control society. The two models organise space and treat their users in different manner producing distinctive information security subjects. In the disciplinary society and in the network perimeter model, the subject becomes ready and complete in the processes of institutions and information security machinery, whereas in Zero Trust model and in the control society the subject remains in the constant state of becoming and resides in the constantly altering space accompanied by doubt.

Keywords: Cybersecurity, Disciplinary society, Information security, Network security, Perimeters, Zero Trust

1 Introduction

In this paper, we analyse a shift in the ethos of cybersecurity by analysing two security models, more conventional network perimeter model and more recent Zero Trust model. We use the term “ethos” to refer to the central characteristics of security model through which its subjects – security, user, and organisation – are constituted. In other words, a security model comes with a particular focus in terms of what should be protected and how the processes of protection should take place. The model is a security architecture that organises its environment and defines the terms by which of connections between different entities can emerge. Indeed, this affects the subjects of information security, i.e., users and organisations that apply the model. The two security models that we focus on are, firstly, the conventional perimeter network security model – in which an organisation has a network with clear boundaries and devices that are defended – and, secondly, more recent Zero Trust

model, which has no clear boundaries, and it welcomes remote connections and “bring your own device” policy (Rose et. al. 2020). In order to analyse ethos, we compare the two models in relation to Michel Foucault’s (2007) disciplinary society and Gilles Deleuze’s (2017) control society. In other words, the two security models are compared to the two theoretical formations of societies in order to track the movement in the ethos of information security. Briefly, in terms of our current western societies and security – especially during the Covid-19 pandemic – we have adopted more and more features of the control society, which include a constant doubt towards its subjects. For example, in an epidemic environment, healthy human beings are constantly under a continuous threat of being infected; being healthy is not permanent status but very prone to a sudden change. In such society, everything is about risk, uncertainties, and assessment of the probabilities (see e.g., Beck, 1992). We examine whether we are able to find similar changes in the field of information security which initially is about controlling risks (e.g., Shin & Lowry, 2020). After all, the change in society is produced also in the field of information security.

Especially, we are interested in the position of individual user. For a long time, there has been a tendency of blaming the user in terms of information security issues (e.g., Evans et al. 2019; Gupta & Sharman, 2009; also cf., Pieters, 2013; Schmidt, 2019), even though the system could be blamed instead. For example, in case of phishing, the user is required to visit a malicious site and it is easy to blame the ignorant user for clicking a suspicious link. However, we could alternatively ask, why does our system allow such an action for the user that becomes duped so easily. Our purpose is not to argue which side should be blamed – that would be a different analysis. Rather, we analyse the way how the individual “user” and their environment is approached in the field information security in terms of the two information security models. Analysis requires a lens through which we examine the object, i.e., the user and their environment. We borrow the lens of examination from Foucault (2007) and Deleuze (2017) and their analysis of society.

We start with the theoretical section by going through the concepts of disciplinary society and control society. We seek to spell out the concepts in extend that is necessary for our purposes. This is to say that the concepts and their connections to other concepts are vast, and we can only cover them partially. Then we use the concept as a theoretical lens and examine the two information security models. In the end, we draw the conclusions.

2 From discipline to control

Let us begin with Michel Foucault’s disciplinary society. Foucault (2005; 2007; 2013), was a French philosopher who examined the systems of knowledge and power – or simply analysed the conditions in which a certain type of thinking emerges. Especially, he was interested in the changes and ruptures of epistemes and discourses. Particularly in the beginning of his career, different discursive systems (constituted by statements) resided at the heart of his examination (Foucault, 2013). However, later, the emergence of individual subject – how subject becomes subjected by itself – got

attention as well (Foucault, 2012). Foucault's (2007) analysis on Bentham's panopticon is well known and has been widely used in different security studies (e.g., Lyon 2006). In addition, Foucault is no stranger in the field of information systems science either (Wall & Buche 2017; Willcocks & Mingers, 2004).

In Foucault's (2007) disciplinary society, power relations organise surrounding space and time in a special way in different institutional environments such as schools, military barracks, factories, hospitals, and monasteries. The arrangement of spaces endorse surveillance and maintenance of a particular order. For example, each institution comes with schedules and architectural solutions by which activities are organised. In the organisation of space (e.g., a classroom with a fixed seating order or long prison hallways with cells), it is easy to spot if a student, a private, an inmate, or an employee does not follow the schedule or other rules. In such a case, they can be corrected (disciplined). Correction – making the situation right again – means that the subjects of the system are made to follow the desired order. In other words, incompatible traits and features – e.g., body positions, postures are eliminated – and the space is made to pulsate with the order (e.g., a rush of pupils emerge between classes, inmates leaving their cells for work, a change of shift in an assembly line). As institutions organise space temporally and spatially, they create an order that promotes internalisation of rules and norms, which they seek to teach.

The system is open to imposition of any desired norms. No school or barrack comes with the readymade rules but rather they are wild cards – order machines (Vuorinen & Tetri 2012) – which are open to be filled with a desired value (Serres, 2007). For examples, schools can be made to serve different ideologies, but in addition, they can have different micro-rules that relate to moulding the looks of its subjects (e.g., short cut hair and uniforms of military). There is always a goal towards which the subjects are mould. Thus, the will to correct springs from the social. In other words, it depends on currently dominating discourses that define, what is wanted out of those masses that are educated (corrected and disciplined) in the facilities. Nevertheless, although the content is open, the possibility of movement and its direction are pre-set in the disciplinary society. Importantly, the institutional space of the disciplinary society is a one-way street (Foucault, 2007; Deleuze, 2017). It means that the individual starts from the beginning (enters a school) and moves through the space that is organised by the disciplinary society (Deleuze 2017). The institutions are passable, as they occupy a physical space with a temporal dimension. If a subject enters an institution of disciplinary society, they go through that institution – a school, barrack, factory, and hospital – in a certain temporal period and they get degrees, military ranks, pension, and get cured (if lucky). The disciplinary machinery is finite and processes its subjects in a scheduled manner. Thus, the subjects become ready, complete, and disciplined. The sites and operations of disciplinary institutions do not stick physically with the individual, but they let go – “they are done with you, off you go”.

Gilles Deleuze, another French philosopher, was thrilled about Foucault's work and commented it eagerly (Deleuze, 2006). Later, in his peculiar reading of Foucault's oeuvre, he developed the Foucault's idea of control society (Deleuze 2017), which follows or co-exists with the disciplinary discourse (Foucault, 2007). At

this point, it should be emphasised that the two societies should be understood as a discursive ideal type, meaning that they are a collection of features that form an ethos. They should not be understood in terms of absolute conditions, in the sense that they would form either a pure disciplinary or control society, but rather these are dimensions which actualise with different intensity in different times.

The society of control differs from the disciplinary society in terms of how the desired order is sought to be imposed. The controls are not tied to a particular place (such as a school) but follow or stick with their subjects through different spaces. The connection to the spatiality is abandoned and replaced by fluid controls. In the disciplinary society, the individual goes to and through an institution of norms. In the control society in turn, the controls move just as if the institution had become the mobile order machine that seeks and correct incompatibilities. (See Deleuze, 2007.) We define the control here as a recurrent check, which seeks to find incompatibilities and stop them.

The controls are limited by time in a different way than in the disciplinary society. They are not permanent but rather permanently recurring. For example, education of the control society does not refer to a passable school but constant re-education for a lifetime. There is no chance for a subject to become ready and complete. A control is never done with the subject. “Off you do not go – we’re never done with you.” To translate Deleuzian controls into the language of Nietzsche’s “eternal return”, it is no longer the eternal return of the same, but eternal return of re-occurring check on the subject (see Deleuze 1994). Controls actualise in different forms. For example, in the world of information security, there is an endless stream of regulatory requirements, which demands (re-)education. Thus, the life cycle of information security training is short as there is a constant demand for a new training package. The information subject – an individual carrying out and being subjected to information security procedures – is never ready; subjection becomes recurring.

3 The two security models

3.1 Network perimeter model

In the course of the Covid-19 pandemic, the use of company networks changed from location-based (office-based use) to remote use. The remote use of digital resources and networks have become an integral part of working life. However, the concept of remote work and digital remote work tools – such as VPN, Zoom, or Microsoft Office Teams – are nothing brand new, but the Covid-19 pandemic forced majority of organisations to transit to the mode of remote work. This change in our societies – in the West – has affected information security solutions as well. Just as the concept of remote work has been there for some time now, so have the two security models – the network perimeter model and Zero Trust model – that we analyse. Both models seek to establish controlled spaces of use, but they differ significantly in their way of carrying out controls in relation to the subject. We argue that the shift from the network perimeter model towards Zero Trust has similarities with the move from the

disciplinary society towards the control society. Moreover, the way in which an information security model works, alters the way of how the model treats its subject – the user. All the models subject their users with different procedures creating a specific information security subject. For example, a two phased identification authentication process subjects the user to extra work creating a diligent information security subject.

In the conventional network perimeter model, once users become authenticated, they are treated as insiders that can access resources in the entire network by default (Rose et al., 2020). The model produces durable insiders in terms of trust. In addition, the network perimeter security model is tied to a physical and logical space – i.e., perimeters of a network – such as a local office network. The controls of model are local but also utilize logical and physical stability. For example, the network perimeter model can use local information for authentication: if user's network address resides within office network, it is not necessary to authenticate and authorise the user again if an access to resources of the same network is required. So, let's assume that a user at the office has signed on the local network. As the user is within the perimeter, access rights can be given by default. In other words, identity authentication for the resources can be partly done by locational information. Users are known by and through the logical and physical space. This resembles the physical and temporal stability of disciplinary society. The institutions of disciplinary society are local and durable. The subjects they produce are durable as the educational degrees last and give the subjects status of being ready. The subject's process of becoming has an end point – for example, an educational degree – and as it is once reached, the subject becomes permanently ready. In the similar manner the network perimeter model grants durable and overarching access rights.

It should be noted that interruption is an integral part of information security – or “information securing” to emphasise its processual nature (Vuorinen, 2014). Fences, walls, and obstacles are inventions that tend to interrupt relations between users and objects – to hinder movement that Luciano Floridi (2005, 106) would call ontological friction in the infosphere. However, information security is not a matter of mere neutral friction – slowing down – but in case of information security the user becomes fully paralysed in case access is denied. For example, a requirement of credentials means a stoppage of the system for outsiders but also for its legitimate users. To access something that is mine, first, I become interrupted by security. Information securing is aggressive and total, which – of course – is a requirement for information to remain safe. Interruption is a price to be paid. In this sense, the cybersecurity that utilizes perimetric information seeks to diminish interruptive activity towards the user. It uses devices and information for authentication, decides and lets the subject go.

3.2 Zero Trust model

The development of a more recent model, Zero Trust Architecture, started with the idea of de-perimeterisation (Rose et al., 2020; see also Jericho Forum, 2008). Compared to the network perimeter model, Zero Trust is more fluent and mobile and

does not bound its activity to a local network space but rather it assumes all traffic as if it came from outside: thus, it treats the users in a very different manner (Rose et al. 2020). In Zero Trust architecture, the inside region is revealed in a partial way in the sense of Just-inTime and Just-Enough. In other words, the access is granted for a period. Zero Trust is able to deal with changing network topology, because it does not assume a perimeter that would create an inside; it is able to work with “bring your own devices” and remote work, as it treats the local network as if its assets and users were from the outside world, i.e., they under a constant doubt (Rose et al. 2020).

In Zero Trust model, the requirements for authentication are continuous whereas in network perimeter model, in principle, all the user access rights are given by default. In the world of continuous authentication, the context and behaviour of the users are monitored constantly. If exceptions to the defined baseline are noted, the user is requested to identify and prove themselves again. The “inside region” – if there even is an inside – is trusted as much as outside space. As if it was the eternal return of authentication and authorisation. Zero Trust model does not ignore logical and physical information. However, the significant difference – compared to the network perimeter model – is that logical evidence – such as an IP-address – does not work as a foundation of trust but as a source of doubt. Any change in physical and logical condition sparks suspicion. Zero Trust model increases the complexity of information security and increases the need for energy and resources that are consumed for the security analysis. Analysis, in turn, is as continuous as in Deleuze’s (2017) control society. As information security is essentially interruptive, this means that either the user is interrupted constantly (to re-check that the analysis is correct) or to protect the user, the other parameters are needed for continuous authentication. Although the energy consumption of such systems is mild, we can claim that network perimeter security is more straightforward and requires less energy. The energy, which we refer to, comes from the system that is protected (Vuorinen and Tetri 2012). There simply is no other source.

Society has moved towards control society – institutions have changed, and Covid19 pandemic has made them mobile, the information security models follow that by moving from the conventional network perimeter security towards Zero Trust model that seeks to tackle the problem of mobility and fluid perimeters. However, this emphasises continuous authentication and subjects the user to constant analysis. This, of course, makes security work much more complicated. The locations have been abandoned and replaced by moving users. As Zero Trust model considers all the systems to be infiltrated, it, in fact, makes the system a different place; in the control society, there are no safe places but constant doubt – just what Zero Trust model delivers.

3.3 The two information security subjects

The functions of disciplinary society are tied to the place just as is in case of the perimeter security model. The role of the subject is always clear. The same locality comes with the information security subject. As information security models are employed, they become connected with ethical demands which relate to the correct –

desired – behaviour. For example, an information security policy states how users should behave – what they can and cannot do. The desire of proper conduct – i.e., what a policy wants from the user – does not differ between models. The models are open for different policies. Slightly differently, the two models seek to achieve the goal the proper way of conduct, which is open for (re)definition.

Both models seek to provide a clean, a proper space with a particular order. They try to organise the space and the users within (Vuorinen & Tetri 2012). “Clean” equals to specific order and all the incompatible actors in relation to the order are sought to be excluded. The result of purified space is sought by both models. However, when they seek to achieve that total purity and compatibility, the models generate conditions for emergence of different information security subjects.

Hitherto, by the term “information security subject”, we have referred to the conditions and pressure of demands surrounding the user and how the environment initiates a process of subjectification. For example, demands of information security policy partially seek to form an information security subject i.e., certainly behaving subject that is compliant with information security policy (e.g., Amankwa, Looock & Kritzing 2018; Moody, Siponen & Pahlila, 2018). However, drawing on Foucault (2012), we are able to add the dimension of self-relation into the concept of information security subject. In this subjectification – emergence of information subject that is facilitated by the information security policy code – there are three important aspects.

Firstly, there is the “self”. This refers to the individual user as a conscious actor. Secondly, there is the ethical or moral code, by which the self is subjected by itself – i.e., self creates a relation to the self through the code. For example, a simple piece of advice “do not use the same password in multiple services” is a security code that requires a certain behavioural pattern (don’t use the same password in multiple services). These are order words that seek to organise the world (Deleuze & Guattari, 1988). An individual, who faces this kind of requirement, subject themselves through this requirement. However, the attitude what is employed before such a piece of advice differs. How important such a requirement is thought to be differs. No matter what the significance of such a statement is at the level of individual, still all the activities that are launched by such order words, can be considered as creation of a relation to the self. What am I going to do in order to remember such a statement? What kind of a subject I am supposed to be – and what I am going to be. These define the information security subject – the subjected self by the code. However, there is the attitude towards the code. What am I going to do in order to carry out that code, to remember it? Foucault (2012) calls this as the relation to self.

Both security models provide an environment in which the information security subjects can emerge – but the pressure (subjectivation) is different. In other words, the code of conduct that is taught is the same but the environments, in which the code is imposed, are entirely different in terms of pressure: the models provide unique spaces for subjectification. The perimeter model that springs from the disciplinary society systems of thoughts relies on the thought of single entity; user is a point of activity, which then is authenticated once. The process of securing becomes ready and complete which then in turn functions as a trusted point as part of the secured inside.

The trusted user is compatible with the order of model. A single authentication brings the access to various space. Information security subject becomes ready as well. The space, on which the user has once signed on, provides a permanent status of identified, authenticated, and authorized user.

The situation is totally different with Zero trust model, in which the subject never becomes ready. There is no single activity through which user would be authenticated and authorised permanently, but the process of control is constant – it is dynamic. “Access to resources is determined by dynamic policy—including the observable state of client identity, application/service, and the requesting asset—and may include other behavioral and environmental attributes” (Rose et al., 2020, p.6). User is trusted momentarily but if there is a change in the behavioural pattern, the user must be reauthenticated and re-authorised. There is a constant re-evaluation going on with recurring multiple policy decision and enforcement points, whereas in the network perimeter model there is only a single point (Rose et al., 2020). The multiple points are needed as the status of authenticated withers away.

However, in terms of subjectification, there are no continuous new requirements for the subject to adapt a new attitude. There are no order words for the user towards which they could subject themselves to. Instead, there is a constant analysis on where the user is and what the user does. The identity of user is not authenticated by a stable information about location, but the identity and authentication are built on numerous factors. Thus, a single user is not an individual but “dividual” just as in Deleuze’s (2017) society of control. Individual becomes dividual as they become divided in the slices of control: IP-address, geolocation, device, contents of their network traffic, operation system version, and software used – all possible data should be collected according to Zero Trust model (see. Rose et. al., 2020).

The process of identification, authentication, and authorisation does not produce a permanently authorised user, but the process continues constantly as there is no permanent trust. There are peaks of trust that fade away. This certainly follows the conventions of the control society. Nevertheless, there are no constant requirements of subjectification (i.e., identify, form, and contemplate yourself as an information security subject) but continuity relates to the processes of never-ending observation. In other words, subject is not interrupted but monitored in terms of abnormalities. *Abnormal attracts doubt*. Normal and reoccurring patterns of behaviour work as an authentic individual fingerprint.

Yet, we have to remember that all the requirements that apply to individual users in the perimeter security model, they also apply to users under the Zero Trust model. Intriguingly, regarding subjectivation and the relation to the self, Zero Trust model in fact seeks to bypass the relation to the self as it assumes user not to be trusted. *At the theoretical level, Zero Trust Model cannot blame the user as it does not trust the user in the first place*. It authenticates the user not by the essence of user (what user is) but through user’s effects on other actors i.e., in what way the user uses the network. More precisely, the Zero Trust model does not assume an absolute user in the sense that the user behaviour is universal, general; rather users are unique (individual and dividual) and there is no essence of a user but “userity” is built upon on user’s behavioural history. Behaviour is defined by the external relations. The approach

resembles the ontological assumptions of actor-network theory, which does not reduce actor to essence but to its effect on other users (e.g., Latour, 2007; also Harman, 2009). User is authenticated against its own history, of its behavioural profile. In this sense, Zero Trust “favours” stability, repetition of the same instead of surprises and new. Zero Trust does not interrupt the user if it repeats the same. However, the moment user diverges from the baseline, interrupts the cycle of the same – i.e., becomes different – then Zero Trust model steps in and halts the user. The control is ready to correct the user. Zero Trust is appalled by the new, by generative activity (see Zittrain, 2008). Indeed, the security model carries out the Floridi’s (2005) “ontological friction” by sticking in the same.

4 Conclusions

Zero Trust model signifies a shift towards Deleuze’s society of control, in which individuals are subjected to continuous cycles of re-education. In terms of Zero Trust information security model this means a process of continuous authentication. In the societies of control, power relations are mobile as the controls follow their subjects. No process of identification, authentication and authorization becomes complete and final but is temporary (just for a certain time, just for the resources that are enough); the controls follow the user is through different (cyber)spaces ready to interrupt the user if necessary. Zero trust model has no perimeter-based space, but the space is opened solely by individual requests. Thus, the user becomes rather a wandering nomad than a subject that passes through a single space as it would be the case in the societies of discipline – or in the network perimeter model.

The two models subject their users differently. The perimeter model does not potentially interrupt the user as frequently as the user in Zero Trust model because authentication and authorisation processes are static instead of being dynamic. Zero Trust model is ready to monitor the user and doubt it constantly. In terms of authentication, Zero Trust turns away from mere what user has (token) or claims to know (password) – claims to be – as it turns its analysis on the user and users’ effect on its environment, i.e., surrounding actors (such as network traffic, assets requested, resources accessed, software used). This way Zero Trust model moves away from essential identity (what the user is), towards connective and behavioural identity. It seeks to individualise the activities related to the user. In disciplinary society an educational degree is seen to construct the authenticity of the subject, in control society it is only a part of authentication data. There is a clear shift in ethos; information security subject is not about the relation to the self but its effect to other actors.

In terms of user being the weakest link, the Zero Trust model distributes the agency of security. It no longer lies on the “essence” of user, but it seeks to grasp the user by probing the users’ environment. In terms of information security subject, no security model will have access to the relation of self, but they are bound to impact the code by which the user subjects themselves. However, Zero Trust seeks to bypass the relation to self by observing what takes place in the environment of a user. It seeks

to define the user by analysing user's effect on their behavioural environment and context. The shift in ethos moves the source of "good" and "authentic" from the essence of user to the users' effects on the environment.

References

- Amankwa, E., Loock, M., & Kritzinger, E. (2018). Establishing information security policy compliance culture in organizations. *Information & Computer Security*, 26(4), 420–436.
- Beck, U. (1992). *Risk society: Towards a new modernity*. London: Sage Publications.
- Deleuze, G. (1994). *Difference and repetition*. Columbia University Press.
- Deleuze, G. (2006). *Foucault. Continuum*.
- Deleuze, G. (2017). *Postscript on the Societies of Control*. Routledge.
- Deleuze, G., & Guattari, F. (1988). *A thousand plateaus: Capitalism and schizophrenia*. Bloomsbury Publishing.
- Evans, M., He, Y., Maglaras, L., & Janicke, H. (2019). HEART-IS: A novel technique for evaluating human error-related information security incidents. *Computers & Security*, 80, 74-89.
- Floridi, L. (2005). The ontological interpretation of informational privacy. *Ethics and information technology*, 7(4), 185-200.
- Foucault, M. (2005). *The order of things*. Routledge.
- Foucault, M. (2007). *Discipline and punish: The birth of the prison*. Duke University Press.
- Foucault, M. (2012). *The history of sexuality, vol. 2: The use of pleasure*. Vintage.
- Foucault, M. (2013). *Archaeology of knowledge*. Routledge.
- Gupta, M. & Sharman, R. (eds.). (2009). *Social and human elements of information security: Emerging trends and countermeasures*. Information Science Reference/IGI Global, 2.
- Harman, G. (2009). *Prince of networks: Bruno Latour and metaphysics*. re. press.
- Jericho Forum. (2008) *Collaboration Oriented Architectures*. Position paper. Available at https://collaboration.opengroup.org/jericho/COA_v1.0.pdf (viewed on April 26th 2022).
- Latour, B. (2007). *Reassembling the social: An introduction to actor-network-theory*. Oup Oxford.
- Lyon, D. (2006). The search for surveillance theories. In Lyon, D. (Ed.). *Theorizing surveillance*. (pp. 17-34). Routledge.
- Moody, G. D., Siponen, M., & Pahnla, S. (2018). Toward a unified model of information security policy compliance. *MIS quarterly*, 42(1), 285–311.

- Pieters, W. (2013, December). Defining" the weakest link" comparative security in complex systems of systems. In 2013 IEEE 5th International Conference on Cloud Computing Technology and Science (Vol. 2, pp. 39-44). IEEE.
- Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). Zero trust architecture (No. NIST Special Publication (SP) 800-207). National Institute of Standards and Technology. Available at <https://www.nist.gov/publications/zero-trust-architecture>
- Schmidt, A. (2019). Don't blame the user: toward means for usable and practical authentication. *Interactions*, 26(3), 73-75.
- Shin, B., & Lowry, P. B. (2020). A review and theoretical explanation of the 'CyberthreatIntelligence (CTI) capability'that needs to be fostered in information security practitioners and how this can be accomplished. *Computers & Security*, 92, 101761.
- Serres, M. (2007). *The parasite*. University of Minnesota Press.
- Vuorinen, J., & Tetri, P. (2012). The order machine – The ontology of information security. *Journal of the Association for Information Systems*, 13(9), 695–713.
- Vuorinen, J. (2014). *Parasitic Order Machine: A Sociology and Ontology of Information Securing*. Annales Universitatis Turkuensis.
- Wall, J. D., & Buche, M. W. (2017). To fear or not to fear? A critical review and analysis of fear appeals in the information security context. *Communications of the Association for Information Systems*, 41(1), 13.
- Willcocks, Leslie P., and John Mingers. *Social theory and philosophy for information systems*. John Wiley & Sons, 2004.
- Zittrain, J. (2008). *The future of the internet--and how to stop it*. Yale University Press.

Building the Learning Environment for Sustainable Development: a Co-creation approach

Ewa Duda^[0000-0003-4535-6388]

Maria Grzegorzewska University, Warsaw, Poland

eduda@aps.edu.pl

Abstract. Education for sustainable development supports the improvement of knowledge, skills, attitudes and behaviors related to global challenges such as climate change, global warming and environmental degradation, among others. It is increasingly taking place through projects based on information and communication technologies. The effectiveness of the actions taken depends not only on the quality of the project activities or the sophistication of the innovative tools used. Social commitment also depends on the beliefs and moral judgements manifested by potential recipients of educational activities on environmental issues. This study aimed to identify the beliefs and moral judgements that may facilitate or hinder the implementation of educational activities based on information and communication technology, shaping proenvironmental attitudes and behavior among city dwellers. Based on the cocreation workshops conducted, five general categories emerged: responsibility, sense of empowerment, local leadership, real eco-approach, and ecoknowledge. The research findings may contribute to the design of educational activities dedicated to shaping the pro-environmental behavior of city dwellers.

Keywords: Adult Education for Sustainable Development, Learning environment, Moral judgements, Pro-environmental attitudes, Proenvironmental behaviors, Urban education

1 Introduction

One of the biggest challenges of recent years has become the growing need to slow down the direction of human-caused climate change. Issues such as environmental degradation, negative consequences of globalization, social inequality and poverty are increasingly noticeable. We are beginning to understand we should be more concerned about the quality of the environment, but above all, be aware of how not to destroy what nature offers us and give the next generation a chance to (worthily) live. Transformations, both systemic and individual, are necessary. Awareness of the negative impacts of our acts plays a significant role in changing current unsuitable habits. It is essential to learn how to anticipate the consequences of future actions and start planning them more carefully.

One of the tools to support human behavior change is the creation of proenvironmental policies based on education. In particular, education for sustainable development (ESD) (Vare and Scott, 2007; Leicht, Heiss and Byun, 2018; Scott and Vare, 2020) is becoming more widely recognized and implemented in everyday life. ESD is much more than environmental education. It has a broader, multidisciplinary context, including sociological, pedagogical, economic, political, and cultural aspects (Arbuthnott, 2009; Venkataraman, 2010). Education for sustainable development refers not only to environmental knowledge but also to relevant skills, attitudes, and beliefs. Values such as sustainable society, responsibility for the Earth, and responsibility for a shared future are becoming increasingly important. However, it should be noted that they are still not sufficiently reflected in our daily routine. A solid education on a global scale seems to be necessary.

ESD is constantly evolving, covering more and more issues and dimensions. The approach applies at different levels, from early childhood education (Hedefalk et al., 2015; Siraj-Blatchford et al., 2016), school education (Mogren and Gericke, 2017; Hallinger and Nguyen, 2020), higher education (Mulà et al., 2017; Hallinger and Chatpinyakoo, 2019), but also within adult learning (Noguchi, 2019). The potential of this last target group seems underestimated and not fully explored. Moreover, the main focus of ESD is on formal education, leaving behind the non-formal and informal learning that fills most of our lives. Without any doubt, more attention should be paid to adult education for sustainable development.

Educational pro-environmental activities based on information and communication technology (ICT) are one example of the ESD concept implementation. Projects aimed at enhancing people's environment-friendly behavior attempt to involve adults in the learning process in innovative ways. Such projects adopt the principles of encouragement through a system of incentives rather than penalties. The growing number of community-based programs not only enables learners to develop their everyday educational practices but also provides an opportunity to deepen research in the field of a learning environment. Researchers are expanding their analysis by addressing issues of pro-environmental behavior, sustainability interventions (Ro, 2017), social learning (Kaaronen & Strelkovskii, 2020), and environmental knowledge (Pothitou et al., 2016). The conducted research allows gaining knowledge about people's habits, attitudes, perceptions, and values regarding ecology issues.

The presented article reflects on research to construct the scientific background for an emerging transdisciplinary project designing an interactive ICT system to promote pro-environmental behavior among urban residents. It focuses on identifying the existing beliefs and moral judgements of people involved in a participatory learning process based on a co-creation approach. Their recognition and understanding will contribute to a better implementation of the process of co-creating an effective lifelong learning environment for climate improvement. The consecutive paragraph will briefly present the theoretical background and literature review on moral judgements. The third paragraph presents the study methodology, the fourth one shows the results obtained during the co-creation workshop, and the fifth one discusses the findings and conclusions. The article closes with references.

2 Theoretical background

2.1 Co-creation

Many researchers and practitioners consider the co-creation method as a valuable tool for involving future audiences in the process of designing products (Roberts & Darler, 2017), services (Jaakkola et al., 2015), and values (Farr, 2015). In particular, the potential of using co-creation is widely considered in building ICT-based solutions (Johansson et al., 2012). Breidbach et al. (2013) found that people's motivation their actions play a more critical role in the ICT-enabled co-creation process than the technology itself. Technology acts as a facilitator rather than a moderator of change. The social collocations created and the decisions and actions taken by people play a pivotal role in this process.

In the field of urban policies, studies on dwellers' participation in the co-creation of solutions for cities show an increase in engagement due to both residents and the government side. As a result of collaboration, authorities become more open to the needs and expectations of residents, their decisions become more effective and efficient, while city dwellers become more satisfied and accepting of government actions regarding the functioning of the city (Agusti et al., 2014). Evidently, cocreation is no longer seen as a way of production, but as an expectation of multi-level sustainable development of the whole city, through strengthening urban functioning, urban planning, municipal governance, inhabitants development, and building association among individual residents and officials (Wamsler, 2016).

The connection of these fields has recently found more and more followers. Building ICT-based city solutions facilitated by a co-creation process is thus becoming more popular. Much of it is used by commercial organizations, focused on their own strategic goals, employed to drive the process. The research focuses on identifying factors relevant to the co-creation mechanism, considering the variety of actors, culture, commitment and attitudes (Akterujjaman et al., 2020).

2.2 Moral judgements

According to Rest et al. (1997) "moral judgment is a psychological construct that characterizes the process by which people determine that one course of action in a particular situation is morally right and another course of action is wrong". Moral judgements are a pivot part of our engagement with society. They play a substantial role in how we perceive ourselves, others and the situation where we find ourselves. They allow us to decide whether to engage in an activity or withdraw. Our moral system also largely determines the climate (in)action we take (Peeters et al., 2019) and how we judge the pro-environmental behavior of others. Research indicates that people consider the intentionality of action when assessing its environmental impact. An activity for which both the purpose and effect were pro-environment was rated as less significant than one for which only the result was pro-environment, while the reason was not environmental (Hoogendoorn et al., 2019).

Markowitz (2012) argues that when people view environmental issues through the lens of moral judgment, rather than simply a scientific or technical problem, their proenvironmental attitudes are stronger. These people have a greater willingness to engage in climate change mitigation. They also express a belief that actively responding to climate problems is their ethical obligation. Likewise, Wang (2017) found that peoples' moral attitudes correlate with subjective norms. Moreover, moral attitudes correlate with anticipated guilt. Individuals with collectively higher moral attitudes, higher levels of anticipated guilt, low empathy, and future orientation were more likely to show greater willingness to engage in global warming mitigation behaviors.

3 Methodology

3.1 Research goals

The study serves the effective implementation of the project aimed at creating an educational application based on ICT solutions (called Greencoin application), which is intended to be a tool for shaping and strengthening the pro-environmental behavior of city residents. The study methodology, regarding the consecutive phases of the project, follows the steps presented in Figure 1. The first phase is to define the main objective of the project: create a mechanism based on an interaction between government, private sector and citizens to bring sustainability to our cities. The proposed system will be implemented and tested in the Gdańsk and then can be transferred into any other city.

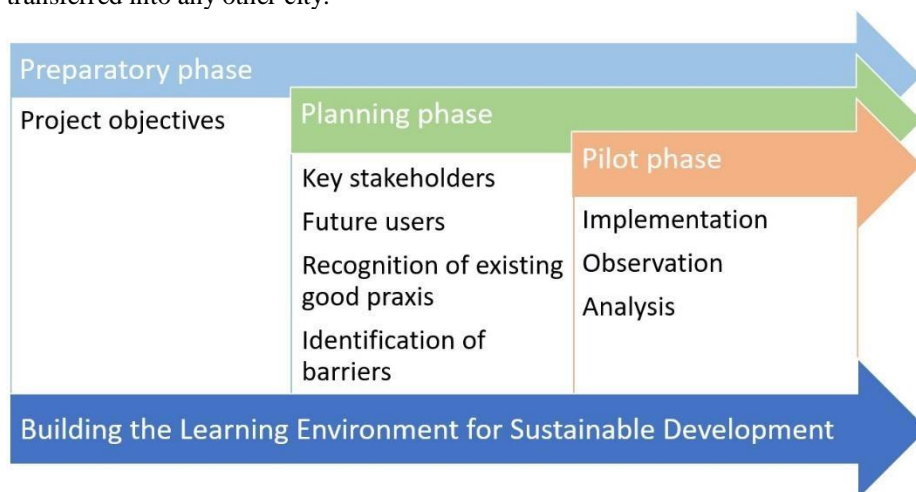


Figure 1. The study framework. Author's own elaboration

In order to achieve the project objective, it was planned to involve a wide range of stakeholders in the participatory process of active creation of solutions initiating pro

environmental behavior among city dwellers. Workshops for crucial stakeholders and future recipients of project activities will enable the establishment and maintenance of an active dialogue with them at every stage of the project flow. A direct relationship will facilitate the creation of a product that meets not only the needs of the environment but also the current needs of city dwellers.

The purpose of the current phase of the project, and thus of the presented study, is twofold:

- (1) to identify existing beliefs and moral judgments that may facilitate the implementation of ICT-based educational activities that shape pro-environmental attitudes and behaviors among urban residents;
- (2) to identify beliefs and moral judgments that may hinder the implementation of ICT-based educational activities that shape pro-environmental attitudes and behaviors among urban residents.

3.2 Participants

A co-creation methodology was used to achieve the objectives of the study. The second phase of the project employed a series of participatory workshops. The first one-day workshop was held stationary (Figure 2). Due to restrictions related to the state of the epidemic, subsequent workshops were held remotely. The cycle of workshops was conducted between October and November 2021. The stationary workshop took place in the city of Gdańsk. The next ones in remote form allowed the participation of residents of other Polish cities. The workshops were attended by local stakeholders active in different thematic areas and occupying different socioprofessional functions. About 50 participants took part in the entire cycle of workshops.



Figure 2. Photo from the co-creation workshop held onsite

3.3 Workshops

The workshop aimed to provide information on the issue of implementing an alternative currency as a solution to support the development of informal networks and services between local community residents. The workshop was held in five sessions, face-to-face for one hour each over one day, remote for 20 minutes each, over one afternoon. The main issues of the workshop included activities aimed at identifying difficulties and challenges faced by city dwellers, identifying wellfunctioning solutions in Poland and the world, reflecting on ways to solve current challenges of cities, in response to the determined challenges reflecting on the system of the functionality of the planned application and the set of rewards supporting it, analyzing the strengths and weaknesses of the potential mechanism for supporting pro-environmental behavior for city dwellers.

The workshop structure consisted of one introductory lecture, presented with the help of a power-point presentation, and other practical sessions, carried out in a stationary form in groups of several people, at tables, using sheets of paper, stickers and colorful stationery. Participants in the remote workshops used the equivalent tools offered by the Jambord application. The workshop was also attended by members of the project team. Four of them joined each group of workshop participants, one acting as moderator, one as a facilitator, the other two as observers. They observed, listened and noted down the participants' perceptions and behaviors accompanying the discussion of the workshop topics, including uttered beliefs, moral judgements and comments.

3.4 Data analysis

The data gathered included material collected by six observers (members of the project team) during the onsite and two observers during the remote workshops. The material consisted of notes taken directly during the workshops and after the workshop during the project team's closing discussion. Data analysis was conducted using a general inductive approach strategy (Thomas, 2006). According to the procedure adopted, coding was done based on the raw data. The categories were taken from the dominant and accented phrases uttered by the workshop participants. The advantage of this approach is that “although the findings are influenced by the evaluation objectives or questions outlined by the researcher, the findings arise directly from the analysis of the raw data, not from a priori expectations or models” (Thomas, 2006, p. 239). The codes selected reflect the actual content of the respondents' statements. The analysis carried out was aimed at developing the detailed beliefs, moral judgements reported directly by the workshop participants.

4 Results

4.1 Responsibility

The workshop participants' statements indicate a divergence in beliefs and moral judgements of various stakeholder groups. The category that emerged most often was the attribution of responsibility for caring about the quality of the urban environment. The example cited was waste management. Municipal officials undertake various policies and top-down remedial measures to improve the situation in the city, but these do not have the expected, noticeable effect. They take the form of regulation through appropriate legislation specifying waste collection and disposal fees. Despite the necessity of waste segregation into particular fractions, the city does not have a good way of verifying how this process works. It is particularly evident concerning housing estates, where there are no mechanisms to control whether and how residents separate waste. Officials limit their actions to establishing appropriate guidelines but do not take successful measures for their implementation, indicating that the responsibility for segregation lies with the residents themselves.

On the other hand, the residents feel responsible for proper waste segregation to a limited degree, shifting the responsibility onto companies collecting waste from their properties. Workshop participants pointed out that the city authorities do not carry out appropriate educational activities to raise the inhabitants' awareness of both the correct methods of waste segregation and the consequences of such actions for the environment but also the economic aspects of the city's operations. If the city managers were responsible for their waste management policy, implementing solutions facilitating its efficient implementation, the inhabitants would be more motivated to cooperate for the common good. Another aspect is the inadequate selective waste collection system in public spaces and companies/institutions. The users of these spaces do not feel responsible for looking for the right bin to dispose of a given fraction. They use those bins which are closest.

Other issue is the responsibility for costs. Despite the implementation of European Union directives, inhabitants are faced with the dilemma of who should bear the costs of producing excessive amounts of waste. Workshop participants noted that the trend of producing packaging that is disproportionately to the size or needs of the product, overuse of pre-packaging, or multiple packaging of the same product, serving de facto to manipulate consumer behavior, is growing. They blamed the authorities for the lack of an effective law forcing producers to use returnable packaging. On the other hand, the customer does not have much influence on how the product is packaged. Workshop participants expressed the belief that the producer should bear the costs of waste utilization, as he decides on the form of packaging produced, which soon becomes rubbish hard to recycle.

Urban transport is another example of the blurring of the responsibility for adopting pro-environmental behavior. Workshop participants representing city residents pointed out that the main reason for poor mobility habits is poor city management. The location of strategic services is mainly in the central districts. The granting of building permits for numerous new apartment blocks without the

simultaneous construction of schools and kindergartens means that residents are forced to commute constantly to the city center - resulting in heavy traffic jams. The inflexible public transport fare system exacerbates the situation. The city authorities are unable to carry out an effective process of integrating urban transport services offered by dispersed providers. Overcrowded public transport discourages people from using public transport.

4.2 Sense of empowerment

Another category of beliefs that emerged from the workshops was that of empowerment. Representatives of city residents repeatedly expressed the belief that they have little influence on the quality of the environment. They perceived their activities as insignificant. They said they were not very motivated, for example, to save water in their flats, seeing that other neighbors in the block of flats did not do so. In addition, water consumption for shared purposes, the costs of which are distributed among the inhabitants of individual flats (washing of staircases, washing of the nearest neighborhood belonging to the block of flats, watering the greenery around the block of flats by the managers) generates high charges, concerning which individual savings are not satisfactory.

Home heating was another example. Residents of detached houses spoke about how difficult it was for them to give up wood-burning, especially in their fireplaces. They claim that they have good quality fireplaces certified by law and do not emit as much smoke as the industrial plants in the city. They do not use them constantly, often once or twice a week, especially in the early heating season. The reason for using fireplaces is for comfort purposes and aesthetics rather than for actual heating needs, which creates the impression of a negligible global impact on the environment. Despite the interest in environmental issues, this behavior was not seen as being unecological, especially as wood is still considered by the population as a renewable resource.

4.3 Local leadership

The next category that emerged was the attribution of blame for the lack of action to an insufficient number of community leaders. Workshop participants felt that they would like to do something to improve the quality of the environment in which they live, but they do not know what exactly to do and how to do it. According to them, their closest environment lacks initiators of social change. In particular, the fact that their employers are not such initiators was highlighted. If, for example, employers support their employees through a commuting subsidy scheme, they would be more likely to use city transport or cycle. Workshop participants also cited examples of shared commuting across borders. Several people use their means of transport to reach an agreed point, where they change to one car and travel together to their workplace in the city center.

Another example is the lack of positive change leaders among the managers of large corporations or workplaces, who do not try to be truly eco-friendly, often using

sham measures. Positive solutions developed in the pandemic period are abandoned in the post-pandemic period. An example of this was the circulation of documents. At the time of the pandemic, most workplaces introduced an electronic workflow confirmed by an electronic signature. Employees started to use it in their day-to-day work, and it seemed that due to the numerous perceived advantages (saving time, paper, toner) this system would become permanent. However, with the return to stationary work many employers began to require paper documents again. According to the workshop participants, if the employer is not an initiator or an eco-leader, the employees will not implement environmentally friendly solutions in their daily work either. The potential of employers is untapped in this area.

4.4 Real eco-approach

The next category indicated by workshop participants was insufficient social acceptance of eco-behavior. Without positive experiences related to the benefits of displaying pro-environmental attitudes and behaviors, city dwellers have little motivation to demonstrate them in everyday life. Respondents clearly expressed the belief that there is a lack of consistency in the media approach to environmental problems. On the one hand, the media publish many films, documentaries, and social actions aimed at raising the ecological awareness of their audiences and encouraging them to implement positive changes to improve the quality of the environment. On the other hand, media flood their audiences with aggressive advertisements tending to increasingly consumerist lifestyles. What is more, these advertisements are often targeted at the youngest viewers.

Workshop participants were concerned about environmental issues but felt that their environmental awareness was limited and vulnerable to manipulation. They gave the example of a recent advertising campaign promoted in the media. A chain of shops encouraged people to buy water under the slogan "save wolves", declaring that they would donate a certain amount from the sale of each bottle of water to this cause. However, while helping nature in this way, they were also harming it because not only did they encourage people to buy bottled water instead of promoting drinking tap water, but they were also selling it in bottles made of plastic.

During group discussions, participants expressed the opinion that they would like to live a greener lifestyle, but it takes too much time. They also felt that the costs of being green are passed on to consumers, which does not motivate behavior change. Individuals represented different areas and professional positions, which was reflected in what they saw as the main environmental challenges in the city. Officials expected residents to be more involved in the actions they promoted or implemented, while residents expected the apparent removal of barriers limiting the implementation of green behaviors. Participants suggested that the benefits of implementing new solutions would be higher if we take care to co-create greater social acceptance of pro-environmental approach.

4.5 Eco-knowledge

The last category that emerged most frequently was adequate knowledge of environmental issues. It was perceived two fold. Firstly, there was a discussion on the difficulty of building de facto environmentally friendly solutions. Participants pointed that the basis for creating such solutions is in-depth knowledge of the course of individual processes. An application that could monitor mobility behavior was used as an example. Creating a system that controls whether a user earns points for actually riding a bicycle (and not, for example, a scooter or riding a bus that moves slowly because of a traffic jam) requires the implementation of solutions based on a network of Global Positioning System (GPS) transmitters, software, and devices that consume large amounts of energy. Expert knowledge is required to create a system whose operating costs based on excessive energy consumption, the transmission of large amounts of data, and network traffic consumption will not outweigh the profits from users' ecological behavior.

In-depth knowledge is also needed for the optimal inclusion of stakeholders in the created cooperation network. Participants discussed how to assess whether a potential project partner is environmentally friendly, while it may be a big polluter and have a negative environmental impact. City dwellers often have difficulty confirming what is green behavior. An example cited in this regard was electric cars. In the media space, various reports are provided on whether electric vehicles are really green because the electricity to power them can come from burning coal or gas. So, by choosing to buy an electric car and then powering it with electricity of unknown origin, are citizens contributing to environmental degradation?

Workshop participants also pointed out the threat of the growing phenomenon of greenwashing. Companies, driven by the desire to increase profits, deliberately mislead their customers by suggesting that the solutions dedicated to them or the products offered are ecological. In reality, customers are unable or cannot verify this. Negative experiences cause a lack of trust in subsequent campaigns or offers, and people do not believe that ecological products are really eco-friendly.

5 Discussion and conclusions

Creating a learning environment for sustainable development, of which the ICT-based planned application (called Greencoin) is an example, is a process that requires consideration of many factors that affect its subsequent effective functioning. The selection of active collocations and interactions between stakeholders and future users should consider not only the needs and expectations of particular groups of users but also their beliefs or moral judgments about environmental issues. The conducted workshops for stakeholders and future users of the application gave the representatives of these groups a voice in the co-creating process of environmentally friendly solutions, which have a chance to become technically feasible and socially acceptable at the same time. Taking into account the views and beliefs of city dwellers in connection with the experienced problems of everyday urban life facilitates the creation of a pilot solution that can be replicated. When implemented in

other cities, it will shorten the process of reaching out to people so that further users can be involved more quickly and effectively to improve environmental quality.

The presented research was aimed to identify existing beliefs and moral judgements that can facilitate the implementation of ICT-based educational activities shaping pro-environmental attitudes and behaviors among city dwellers. Conducted workshops revealed the need to recruit local leaders for the implementation process of the application, who will enable to reach the city dwellers and encourage them to join the project activities. This endeavor, considered a socio-economic investment, should bring together, as confirmed by other studies, both leaders at the local (Homsy, 2018), neighborhood (Kretser & Chandler, 2020) and workplace levels (Afsar et al., 2018; Kumar et al., 2022). Workshop participants attributed local leaders as initiators of social change, providing direction and a course of action, motivating people to become active in improving the quality of the environment (Gruber et al., 2017).

At the other extreme of the approach focused on looking for factors that support the process is the approach that looks for barriers preventing the process from taking place. Removing the obstacles identified will allow the process to happen naturally. According to this approach, the second aim of this study was to identify beliefs and moral judgements which may hinder the implementation of ICT-based educational activities shaping pro-environmental attitudes and behaviors among city dwellers. One of the barriers identified during the workshop was the attribution of responsibility for implementing pro-environmental behavior to other people. Residents pointed to city managers as those who should initiate change, both by introducing rational regulations governing the multidimensional functioning of the city and by implementing policies and initiatives for the efficient and ecological functioning of the city. On the other hand, city authorities or officials declared that the solutions they are creating are appropriate, but the inhabitants lack the will to act and the sense of responsibility for the environment in which they live. Just as Eckersley (2016) argues, a complex model of social connections based on a system of structural inequalities leads to a blurring of the notion of co-responsibility for climate change and even to the creation of a belief of collective irresponsibility among representatives of particular groups of social actors.

The findings have raised the question of how to strengthen the sense of responsibility for the environment among urban residents. A co-created Greencoin application could reach people who do not consider environmental care as their daily responsibility or who expect particular actions for environment from others (Cleveland, 2020). The theme of meager sense of empowerment also emerged concerning this matter. Workshop participants impugned the relevance of taking individual action due to its low effectiveness, especially against the background of global anti-environmental endeavors. Questions arose: Is it even worth it to act individually? Do our efforts make sense? Will anyone appreciate our involvement? It is not uncommon that people decide to take pro-environmental actions, but when they see the low effectiveness of those behaviors, it decreases their motivation to act for the benefit of climate, so they abandon the undertaken direction of activity (Abrahamse, 2019).

The results obtained indicate that the co-created functionalities of ICT applications for sustainable development should consider the issue of knowledge transfer, which is the basis for building a pro-environmental approach in society. We live in a world where the problem is not a lack of information but an excess of it. We are dealing with a twofold situation here. Knowledge is developing so fast that it is impossible to be an expert in many areas. People's knowledge about climate and the environment is often too superficial, making them susceptible to commercial manipulation or the negative influence of eco-sceptics (Lukinović & Jovanović, 2019). The lack of in-depth knowledge also causes difficulty in assessing whether the taken actions and the resulting consequences are, in fact, environmentally friendly.

The currently developed Greencoin application, like other similar solutions, should consider the results obtained, representing the beliefs, opinions and judgments held by representatives of the local community. If they are not taken into consideration at the solution development stage, then the whole effort of the project team could turn out to be useless, as the product created will not be adapted to the needs and expectations of the future user. Our responsibility is to develop technology that responds to the challenges of this world and is a tool that does not drive change, but enables the change.

The limitation of the presented study is twofold. Firstly, the study was conducted in Poland as an initial stage of an ongoing project. The results are currently of local nature. Extension of the research scope is planned after the pilot phase. Secondly, the research material was gathered by different observers participating in three independent, parallel groups of workshop participants and therefore is inevitably subject to perceptual bias. Therefore, a follow-up study is envisaged and will be based on a series of in-depth focus interviews. Their analysis will be presented in a future article.

Acknowledgments

This research is supported by €1.9 million in funding received from Iceland, Liechtenstein and Norway under the EEA Funds, grant agreement NOR/IdeaLab/GC/0003/2020-00.

References

- Abrahamse, W. (2019). *Encouraging pro-environmental behaviour: what works, what doesn't, and why*. London: Academic Press.
- Afsar, B., Cheema, S., & Javed, F. (2018). Activating employee's pro-environmental behaviors: The role of CSR, organizational identification, and environmentally specific servant leadership. *Corporate Social Responsibility and Environmental Management*, 25(5), 904-911. <https://doi.org/10.1002/csr.1506>
- Agusti, C. et al. (2014). *Co-Creating Cities. Defining co-creation as a means of citizen engagement*. <https://doi.org/10.13140/RG.2.1.3684.5849>

- Akterujjaman, S. M., Mulder, R., & Kievit, H. (2020). The influence of strategic orientation on co-creation in smart city projects: enjoy the benefits of collaboration. *International Journal of Construction Management*, 1–9. <https://doi.org/10.1080/15623599.2020.1736834>
- Arbuthnott, K.D. (2009). Education for sustainable development beyond attitude change. *International Journal of Sustainability in Higher Education*, 10(2), 152-163. <https://doi.org/10.1108/14676370910945954>
- Breidbach CF, Kolb DG, Srinivasan A. (2013). Connectivity in Service Systems: Does Technology-Enablement Impact the Ability of a Service System to Co-Create Value? *Journal of Service Research*, 16(3), 428-441. <https://doi.org/10.1177/1094670512470869>
- Cleveland, M., Robertson, J. L., & Volk, V. (2020). Helping or hindering: Environmental locus of control, subjective enablers and constraints, and pro-environmental behaviors. *Journal of Cleaner Production*, 249, 119394. <https://doi.org/10.1016/j.jclepro.2019.119394>
- Eckersley, R. (2016). Responsibility for Climate Change as a Structural Injustice. [In] Gabrielson, T., Hall, C., Meyer, J. & Schlosberg, D. (Eds.). *Oxford Handbook of Environmental Political Theory* (pp. 1-17). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199685271.013.37>
- Farr, M. (2016). Co-Production and Value Co-Creation in Outcome-Based Contracting in Public Services. *Public Management Review*, 18(5), 654-672. <https://doi.org/10.1080/14719037.2015.1111661>
- Gruber, J.S., Rhoades, J.L., Simpson, M., Stack, L., Yetka, L., & Wood, R. (2017). Enhancing climate change adaptation: strategies for community engagement and universitycommunity partnerships. *Journal of Environmental Studies and Sciences*, 7, 10–24. <https://doi.org/10.1007/s13412-015-0232-1>
- Hedefalk, M., Almqvist, J. & Östman, L. (2015). Education for sustainable development in early childhood education: a review of the research literature. *Environmental Education Research*, 21(7), 975-990. <https://doi.org/10.1080/13504622.2014.971716>
- Hallinger, P. & Chatpinyakoo, C. (2019). A Bibliometric Review of Research on Higher Education for Sustainable Development, 1998–2018. *Sustainability*, 11(8), 2401. <https://doi.org/10.3390/su11082401>
- Hallinger, P. & Nguyen, V.-T. (2020). Mapping the Landscape and Structure of Research on Education for Sustainable Development: A Bibliometric Review. *Sustainability*, 12(5), 1947. <https://doi.org/10.3390/su12051947>
- Homsy, G.C. (2018). Unlikely pioneers: creative climate change policymaking in smaller U.S. cities. *Journal of Environmental Studies and Sciences*, 8, 121–131. <https://doi.org/10.1007/s13412-018-0483-8>
- Hoogendoorn, G., Sütterlin, B., & Siegrist, M. (2019). When good intentions go bad: The biased perception of the environmental impact of a behavior due to reliance on an actor's behavioral intention. *Journal of Environmental Psychology*, 64, 65-77. <https://doi.org/10.1016/j.jenvp.2019.05.003>

- Jaakkola, E., Helkkula, A. & Aarikka-Stenroos, L. (2015). Service experience co-creation: conceptualization, implications, and future research directions. *Journal of Service Management*, 26(2), 182-205. <https://doi.org/10.1108/JOSM-12-2014-0323>
- Johansson, L.-O., Lund Snis, U., & Svensson, L. (2012). Boundary dialogues in co-creation of ICT-innovations. IRIS 35, Proceedings of the 35th Information Systems Research Seminar in Scandinavia: Designing the Interactive Society.
- Kaaronen, R. O., & Strelkovskii, N. (2020). Cultural Evolution of Sustainable Behaviors: Proenvironmental Tipping Points in an Agent-Based Model. *One Earth*, 2(1), 85–97. <https://doi.org/10.1016/j.oneear.2020.01.003>
- Jennifer Kretser & Katie Chandler (2020) Convening Young Leaders for Climate Resilience. *Journal of Museum Education*, 45:1, 52-63, <https://doi.org/10.1080/10598650.2020.1723994>
- Kumar, S., Panda, T.K. and Pandey, K.K. (2022), The effect of employee's mindfulness on voluntary pro-environment behaviour at the workplace: the mediating role of connectedness to nature. *Benchmarking: An International Journal*. <https://doi.org/10.1108/BIJ-05-2021-0237>
- Leicht, A., Heiss, J. & Byun, W. J. (eds). (2018). *Issues and trends in Education for Sustainable Development*. Paris: United Nations Educational, Scientific and Cultural Organization.
- Lukinović, M., & Jovanović, L. (2019). Greenwashing – fake green/environmental marketing. *Fundamental and Applied Researches in Practice of Leading Scientific Schools*, 33(3), 15-17. <https://doi.org/10.33531/farplss.2019.3.04>
- Markowitz, E.M. (2012). Is climate change an ethical issue? Examining young adults' beliefs about climate and morality. *Climatic Change*, 114, 479-495. <https://doi.org/10.1007/s10584-012-0422-8>
- Mogren, A. & Gericke, N. (2017). ESD implementation at the school organisation level, part 1 – investigating the quality criteria guiding school leaders' work at recognized ESD schools. *Environmental Education Research*, 23(7), 972-992. <https://doi.org/10.1080/13504622.2016.1226265>
- Mulà, I., Tilbury, D., Ryan, A., Mader, M., Dlouhá, J., Mader, C., Benayas, J., Dlouhý, J. & Alba, D. (2017). Catalysing Change in Higher Education for Sustainable Development: A review of professional development initiatives for university educators. *International Journal of Sustainability in Higher Education*, 18(5), 798-820. <https://doi.org/10.1108/IJSHE-03-2017-0043>
- Noguchi, F. (2019). The concept of education for sustainable development in adult learning and education. In U. Gartenschlaeger (Ed.), *Rethinking adult learning and education – Asian perspectives* (pp. 105-114). Bonn: DVV International.
- Peeters, W., Diependaele, L., & Sterckx, S. (2019). Moral Disengagement and the Motivational Gap in Climate Change. *Ethical Theory and Moral Practice*, 22, 425–447. <https://doi.org/10.1007/s10677-019-09995-5>

- Pothitou, M., Hanna, R. F., & Chalvatzis, K. J. (2016). Environmental knowledge, proenvironmental behaviour and energy savings in households: An empirical study. *Applied Energy*, 184, 1217–1229. <https://doi.org/10.1016/j.apenergy.2016.06.017>
- Rest, J., Thoma, S., & Edwards, L. (1997). Designing and validating a measure of moral judgment: Stage preference and stage consistency approaches. *Journal of Educational Psychology*, 89(1), 5–28. doi:10.1037/0022-0663.89.1.5
- Ro, M., Brauer, M., Kuntz, K., Shukla, R., & Bensch, I. (2017). Making Cool Choices for sustainability: Testing the effectiveness of a game-based approach to promoting proenvironmental behaviors. *Journal of Environmental Psychology*, 53, 20–30. <https://doi.org/10.1016/j.jenvp.2017.06.007>
- Roberts DL, Darler W. (2017). Consumer Co-Creation: An Opportunity to Humanise the New Product Development Process. *International Journal of Market Research*, 59(1), 13–33. <https://doi.org/10.2501/IJMR-2017-003>
- Scott, W. & Vare, P. (2020). *Learning, Environment and Sustainable Development. A History of Ideas*. New York: Routledge.
- Siraj-Blatchford, J., Mogharreban, C., Park, E. (eds.). (2016). *International Research on Education for Sustainable Development in Early Childhood*. Switzerland: Springer International Publishing.
- Thomas, D. R. (2006). A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation*, 27(2), 237–246. <https://doi.org/10.1177/1098214005283748>
- Vare, P., & Scott, W. (2007). Learning for a Change. *Journal of Education for Sustainable Development*, 1(2), 191–198. <https://doi.org/10.1177/097340820700100209>
- Venkataraman, B. (2010). Education for Sustainable Development. *Environment: Science and Policy for Sustainable Development*, 51(2), 8–10. <https://doi.org/10.3200/ENV.51.2.08-10>
- Wamsler, C. (2016). From Risk Governance to City–Citizen Collaboration: Capitalizing on individual adaptation to climate change. *Environmental Policy and Governance*, 26(3), 184–204. <https://doi.org/10.1002/eet.1707>
- Wang, X. (2017). Risk perceptions, moral attitudes, and anticipated guilt in US consumers’ climate change behavioral intentions. *Journal of Risk Research*, 20(12), 1554–1567. <https://doi.org/10.1080/13669877.2016.1179213>

Capitalism, Technology and the Elderly: The Ethic of Inclusive Digital Conscience for African Values

Dr Thando Nkohla-Ramunenyiwa^[0000-0003-2224-8341]

University of Pretoria, Pretoria, South Africa

thandonkohla@yahoo.com

Abstract. Capitalism and technology have been the driving forces of modernday society. With the Enlightenment Era providing a kick start to their recognized presence, capitalism and technology have gained confidence in their thriving, which has encouraged their continued existence. This thriving is witnessed in fields such as medical field where private hospitals have used technology to save lives and improve medicine. On the contrary, there are cases where capitalism and technology have contributed to social regress, such as creating and feeding division in society based on socio-economics. Like the proletariat, those who have less materialistically tend to be marginalized, and have lost ownership of the fruits of their very own labor. Further, technologically, the likes of the proletariat are excluded digitally based on the kind of access they have to technology and development. In indigenous and African societies, this is a problem since within the confines of their value system of communitarianism, exclusion is repudiated. With the elderly in particular, their involvement and participation in society includes being the custodians of values systems and knowledge. Hence, their responsibility is to instill these values and knowledge in younger generations. To socially exclude them is to alienate them from their responsibilities as elders. The purpose of this paper is to, through the exploration of anthropocentrism and technocentrism, find means to address this digital exclusion, and other layers of it. Accordingly, this paper aims to propose a digitally inclusive perspective that will re-store African values in the midst of a technologically developing world.

Keywords: Technology, values, capitalism, elderly, ethical

1 Introduction

The discourse on digital inclusion is often times preceded by the discussions concerning digital exclusion. For exclusion to occur, there needs to be a system in place that marginalizes and places to the periphery the vulnerable, when it comes to technology.

In works such as Karl Marx's Communist Manifesto, such a system is intensely criticized, and labelled as Capitalism. Marx's definition of capitalism is articulated by

Amin (2018: 431), who states that the primary premise for this system is that those with little or no financial power, the proletariat, sell their labor to the owners of production who have a lot of financial power, the bourgeoisie. Further, Amin states that the bourgeoisie make a profit by exploiting the proletariat through obtaining excess labor for low wage, (2018:431). In addition, Jaeggi mentions how this is a system that has manifested itself at an economic, political and societal level, persistent in engraving a legacy of the Industrial Revolution. Such a legacy is a drift away from the feudal system, gravitating towards ideals of private ownership, profit, and promoting technological sophistication in industry, (2016: 46). Consequently, the progression of the latter may not be accommodating of anything that does conform to its social order, ushering in a necessary condition for exclusion. Part of the excluded entails the proletariat, who are not the biggest players in the economy. In the grand scheme of things, the proletariat is represented mostly by developing countries, who have a weakened sense of integrity and power when it comes to economic agency.

This (digital) exclusion is multi-layered, where by addressing one layer, another layer is realized, creating a continuous pattern of layered exclusions. Capitalism not only feeds this pattern, but also exposes a division in society which sustains this pattern. This division, which is specifically brought on by technology, is the digital divide. Chen and Wellman (2004:19) define the digital divide as a gap informed by resources, or lack thereof, which enable or disenable an individual's participation in the information era. Although this definition is rather general, it acknowledges the existence of a gap brought on by the digital world. The definition which will be more useful for this paper in particular is mentioned by Lembani (2019:72), who highlights the importance of possessing skills which will enable the user to have a meaningful experience with technology. The internet is a popular feature accessed via Information Communication Technology (ICT) devices, which with reference to the latter mentioned definition, encapsulates the possession of skill(s) required to navigate its unlimited, virtual space.

In South Africa, the adoption of internet is revealed in a study by Deen-Swarrray, who reveals that the gap between those who are literate in English and illiterate when it comes specifically to fluency in reading the language is relatively high compared to other African countries in the study. The gap in percentage form is that there are 39.4 more people who can read English fluently in comparison to those who cannot, (2016:38). This not only reveals persistent inequality in South Africa,

(Orthofer,2016:2), but raises the concern that in percentage form, 39.4 more people in South African have a more meaningful experience with the internet than those who do not. The elderly in the rural areas are part of the population who do not have this experience.

The intersection of age and location play a role here, where with age, Hao et al (2017:972) state that social participation for the elderly in South Africa can be challenged by their physical functionality, (and sometimes mental), due to health. As a result, they may be challenged to maintain existing relations with others or to even start new relationships. As far as location is concerned, Correa and Paves (2016: 248) reveal that lack of technological infrastructure and resources lags rural areas behind when it comes to digital inclusion. Consequently, the elderly in rural areas are easily

excluded when it comes to the digital space. As the custodians of the value system and morals of their society, it is concerning that the rural elders in African communities are excluded from guiding the youth, especially in this digital space which can be a challenging space. This is the research problem that this paper presents. This exclusion of the elderly silences them and hence affects their communal role to part values and knowledge to the youth. To explore this in more depth, this paper will be divided into four sections. The first and second sections will address anthropocentrism and technocentrism, which through capitalism, are avenues which contribute to digital exclusion. The third section will provide an analysis of the discussion. The fourth section which follows will propose a theory which will address the current digital landscape by recentering African societies. Lastly, a conclusion will close the paper.

2 Anthropocentrism

Kopnina identifies a common thread in literature when it comes to defining the term anthropocentrism. A term which is essentially founded on placing human existence at the center of creation. The idea that everything was created for the benefit and betterment of human beings. Further, by virtue of being human are worthy of ethical consideration, (2019:1). Using anthropocentrism as an avenue to elaborate on its contribution to digital exclusion, this definition requires more unpacking. Human beings' supposedly central position of creation can potentially produce one of two outcomes which introduce a further breakdown of the definition, namely weak and strong anthropocentrism. Starting with the former;

2.1 Weak anthropocentrism

Weak anthropocentrism advocates the suppressing of human self-interest with the aim of preserving the environment (Norton, 1984:135). Understanding that human beings are the specie capable of reason requires an agency where, with reason in mind, there is a consideration for the environment. In fact, Norton labels this consideration as a

“considered preference”, where human beings are considered as the center of creation and play the responsible role of governing nature in a manner that is beneficial to both human beings and the ecological system. Hence, understanding that human life is part and parcel of the environment and vice versa. Moreover, Ecologist and feminist Carol Merchant (1993: 270) states that this, is an organic world. view, illustrated in preIndustrial Europe as a feudal system, a system which harmonized the existence of human kind with the environment.

Further, Merchant highlights how this world view is one which carries the narrative of Mother Nature, as the environment takes on the metaphor of a nurturing and providing mother for all living organisms, (1993:270). Consequently, as the center of creation in the anthropocentric sense, the human being in this case understands the relationship with mother nature, and as a rational specie has the

natural ability to maintain a good relationship with her. The human being, because of the “considered preference”, is also able to manage a healthy relationship with mother nature, knowing that all species not only depend on her but are fundamentally a part of her.

Indigenous communities and some rural communities in Africa understand this harmony and dependence that human beings have with mother nature. In Botswana and South Africa, the Khoi San (one of the earliest homo sapiens in Southern African region) illustrate the importance of the dependence and harmony that human beings have when it comes to mother nature. This perception is articulated by a Khoi San participant in an ethnographic research by Chris Low. Low explores the ontological and epistemological experience of Khoisan in their natural settings. The participant states that the wind is breathe from God that human beings inhale and exhale to stay alive. Even in its invisible quality, the wind (like the invisible soul), are essential for life and are a connection to God the Creator, (Low, 2007: 574). This insight provided by the Khoisan participant is quite telling of the spiritual nature of human beings, which is connected to a bigger spiritual system in nature, introducing the Spiritual Creator who breathes this life into this united whole. This anthropocentrism is one which places emphasis on human beings being at the centre and connected to a natural and spiritual system surrounding them, as well as a higher Spiritual Power above them that sets the order of the whole ecological system in constant motion. This resonates with the value systems of African communitarianism of harmony, honouring God as a Higher Power, respecting others, as well as respecting elders (such as mother nature who takes care of humanity) (Ndegwah and Kroesen, 7:2012). Elders are important for the sustenance of harmony within a communitarian system as they are the custodians of its value system, bearing the responsibility of passing down values to the younger generations.

Jirata and Simonsen (2014:136) mention a narrative of an ethnic group in Ethiopia, where storytelling, folktales, songs, proverbs, etc are a means through which the elderly pass down values from generation to generation. In addition, children are perceived as moral agents, hence, passing down values to them is an acknowledgement that as moral agents, they have an important role to play in society especially insofar as the sustenance of their values is concerned, (Jirata and Simonsen, 2014:137). Moral agency here is again emphatic of the centrality of the human species as they have the ability to reason when they navigate society, further emphasizing how they treat and manage the environment as well. The harmonious communitarian values persist, as they again encompass how value systems should be inclusive of all living organisms and the presence of a Higher Spirit that too is in unison with nature, human beings and the environment.

At this point, weak anthropocentrism is illustrating an organic, spiritual (immaterial) unity of Mother Nature where all living organisms are under her maternal care. Moreover, the unity is also under a supernatural order that stems from the Creator, adding a very Spiritual and immaterial dimension to this unity. With the introduction of the capitalism and industrious technology, the organic, immaterial reality starts to gradually drift away, gravitating to a reality that thrives on technological production, division of labour, exploitation, profit and less regard for

nature. Consequently, choosing a pre-industrial reality did not fit into the new mold, consequently presenting the inevitable position of being an immigrant to such world. In South Africa, this inevitability is seen with the elderly who stay in the rural environments and take care of grandchildren whose fathers migrated to make a better life for their families. But the dawn of capitalism in South Africa was ushered in earlier, through agriculture and not migration. This dawn was around the 1930's, where the work model of the owner and proletariat materialized, but it was the South African natives who were the proletariat and not the owners, because of colonization, (Morris, 2013;306). This capitalism was accompanied by migration which disrupted the family structure of rural homes, (Schatz, 2007:148), and created a division.

Prensky would identify this as a digital divide within the family, based on a general digital literacy difference between generations, where in this case, the elders who remain at home in the rural areas have less access to technology, and are therefore strangers or immigrants to the digital world (2001:1). Even though some households in rural homes may use mobile phones to communicate with family members in the city, that access is not enough to classify elders in rural homes as digital natives, (Stork and Schidmt, 2009 :2), or even grant them digital inclusivity. In fact, Stork and Schidmt,

(2009:3) state that an increase in the number of households in Africa accounts only to 54%, and that does not assert to the user being skillful in their ability to use internet effectively on their mobile phone or other ICT devices. This, for the elderly in rural Africa especially, would mean that they are digitally excluded, as the broader skills required here do not grant them the access to a better experience with internet. As

Osborn (2006:86) states: "Literacy is an important consideration in a broader definition of access, and user skills for access imply other kinds of literacy". In fact, considering this statement by Osborn, the importance of the combination of literacy and access when it comes to the use of ICT devices in Africa is concerning, as this could translate to a compromise of other kinds of literacy, (Stork and Schmidt, 2009:3).

A division in the household introduced by the digital divide is also a division that speaks to the breaking away of the harmony intended by weak anthropocentrism. In addition, this broken harmony puts the elderly at the periphery when it comes to the digital space, consequently affecting their essential role of transferring and departing values to the younger digital natives. The periphery the elders find themselves in is a breakaway from the city, from development. This breakaway is one which displaces the elderly who essentially becomes digitally excluded, and more so, removes African youth from the transfer of values, hence away from their natural setting. This division within the family is also an indication of an introduction to a strong anthropocentrism.

2.2 Strong anthropocentrism

Strong anthropocentrism is based on the goal of achieving human ends at all times, even at the cost of nature. It is a position that does not prioritize nature conservation and the right of non-humans, it would do so only if it benefits the interests of human

beings, (Kopnina et al, 2018: 118). Norton states that strong anthropocentrism is associated with a “felt preference”, which is short term, and is fulfilled by a particular experience. This preference is one which views human beings as a specie that should exploit nature or benefit from environment at the expense of other species in nature, (1984:135). In contrast to weak anthropocentrism, strong anthropocentrism is a drift away from an organic, immaterial world view to one which is mechanistic and material, necessary conditions for capitalism to thrive. Within this mechanistic (and material) world view where the harmony with nature is substituted by the domination of nature for human benefit. Dominating nature in this sense translates back to the image of nature as a mother, but this time mother nature being dominated and violated for human benefit, Merchant in (Zimmemramn et al, 1993:270). In a capitalist economy, the more economic power and resources human beings have the more enabled they are to dominate mother nature. The bourgeoisie would be the ones who possess the ability to dominate at a higher extent than the proletariat. Thus, the inequality of human beings in strong anthropocentrism is evident, (Kopnina et al, 2018: 115), emphasizing the move away from the harmony and synergy from weak anthropocentrism.

Within the system of capitalism, the division and inequality can be experienced through alienation, where in the context of Marx’s Communist manifesto will be discussed from the view point of the proletariat (the marginalised in society). The alienation experienced by the proletariat was based on an exploitation, where his/her labour was a commodity sold at a price that will profit the bourgeoisie. To achieve this profit, the proletariat has to put in many hours at work, which isolate him from his support structure (his family), from his interests, goals and even the fruit of his labour, Kalekin-Fishman and Langman (2015:918). The narrative of the proletariat represents that of the elderly in the rural areas. Being at the periphery digitally, geographically and in terms of age, they are isolated from the support structure of their children who work in the city, isolated from social inclusion because of their functionality, and lastly isolated from their important duty of passing down values to their children and grandchildren. Even though some elderly may live with their grandchildren, within the house they are isolated in terms of quality time because of the digital divide within the household. Their digital native grandchildren isolate themselves into their bedrooms as they use their ICT devices usage. The time to sit down together with grandchildren, and through oral culture, share stories with idioms used to depart values of living communally, living in harmony as family members, members of society and also nature as a whole. This division and alienation are accompanied by another enabler of capitalism and digital exclusion: technocentrism

3 Technocentrism

The definition of technocentrism which will be used in this paper is rudimentary, yet useful. Brennan states that technocentrism is understanding or perceiving social phenomenon or events as stemming from technology and its advancements, rather than observing complex phenomenon where technology is situated, (2015:289). With

technology being such a huge part of modern life, technocentrism is not an unexpected lens for humanity living in a technologically developing society. It has become a huge part of how human beings execute tasks. Even Marx contends that technology is an enabler, it allows human beings to perform tasks better than before, but in the process of this enabling also comes a loss and an isolation. A loss of control over the production process as well as an isolation from the process, (Bimber, 1990:345). This narrative is twofold: on part of the proletariat, the loss is his labor which is eventually realized as a commodity. The isolation is that he or she isolated from the fruit of this labor that has become a commodity. This is the narrative of the marginalized when it comes to technology. With the bourgeoisie, the loss and the isolation is being removed from the “ought” in their human action. They are alienated from promoting justice in the workspaces that they own. Instead, according to Bimber (1990:345), they focus on the instrumental aspect of technology that they introduce in the production process, which further enables them to increase their profits at a faster rate and shorter period of time. With so much focus on self-gain and using the proletariat and technology as a means to an end, this overlooks the intrinsic value of the proletariat, outweighed by the instrumentality attained from technology.

An instrumental view of technology is not morally wrong, what brings concern is when the instrumental use comes at the cost of the worth and integrity of a human being. Within the context of the home and the digital divide in the home, the elderly in the rural areas, even though they experience of ICT devices is not as advanced as those who live in the city, the instrumental use of technology is to keep in touch with their children other family members in the city. But within the same household the instrumental concept of technology means different things. Nkohla- Ramunenyiwa (2017) in her PhD conducted interviews with teenage children from South African Pietermaritzburg. In this study, what the researcher in question discovered is that even though some of the teenagers were from single headed homes or homes where they are raised by their grandparents, the instrumental use of technology by these teenagers eventually led them to consumerism. Nkohla-Ramunenyiwa (2017:107, 148) highlights that the teenagers revealed their ‘over-indulgent’ behavior in their use of ICT devices, that there were some respondents who use their ICT devices during class time. Consuming ICT devices to the extent that teenagers struggle with discipline even during class time is quite telling of the amount of time they consume on ICT devices. This consumerism is also translated into the home space, speaking to the idea that a majority of the respondents in Nkohla-Ramunenyiwa mention how they use their phones in the privacy of their rooms, (2017:149).

The elderly who live with teenagers during such a time as this where technology is developing at such a fast pace are faced with this experience in the household. The technocentric experience of their teenagers not only aggravates the digital divide in the house, but also contributes to the exclusion of their elderly family members. The advice and guidance which children could seek from their parents and grandparents is now being googled on the internet. Nicholas et al (2011:29) labels them as the “google generation” born 1993 and above, who consult the internet for knowledge before they consult with anyone [Sometimes it may be their only source of consultation, by choice]. Consequently, technocentrism is at play in the home,

where the google generation allow knowledge attained from their ICT devices to take precedence over advice and guidance could receive from their grandparents. The elderly are digitally and physically excluded by their “google generation” grandchildren. Additionally, the compromised use of the internet experience of the elderly in the rural areas makes it challenging to be a part of the digital experience with their grandchildren

This digital exclusion which has fragmented the reality of the natural harmony brought on by mother nature can be further illustrated by Carol Merchants metaphor about the death of nature’, which she articulates is a shift away from the feudal system to a world that is geared towards technologically developing space that does not respect nature. A space where nature is scientifically broken apart in order to be understood. A place where nature, like the female/mother nature, is taken for granted and destroyed in the process, interrogated through the use of science, (Merchant, 2006 :517). This death of nature into the mechanistic world view is also a representation of the death of harmony and coherence that existed within the family. Where African elders were the first point of call for anyone in the community who needed advice. Alienation from the known has become a new and convenient order for the death of nature to be fully realized. The known is not convenient because it reminds the modern man of the “ought”, and encourage the proletariat to constantly challenge the status quo. To avoid this challenge, society must be isolated from the known (the truth and remainder of injustice) so that they do not challenge the new order. To not know is convenient because ignorance is bliss, the bourgeoisie is not reminded of what they “ought” to do, the proletariat is not reminded of what the “ought” is.

4 Analysis

Marx’s Communist Manifesto is a classic contribution in the political economic sphere and has addressed the current socio-economic landscape of inequality in the most prophetic way. With the introduction of technology, this discourse introduces a nuanced approach as far as the challenges which come with it are concerned. The digital divide is one of the challenges, which in the digital space in particular, divides to exclude, and in excluding alienates. This is in a nutshell the experience of the digitally excluded in society can be elaborated through the avenues of anthropocentrism and technocentrism. The digital divide has prompted the following divisions and exclusions within the home which can also be extended to society at large:

4.1 Clashing worldviews

Organic and mechanistic worldviews illustrated by Merchant are presented in the shift from feudal to capital system. This shift is not only an economic shift but it is also a mind shift. Even if the world has to a large extent changed in terms of appearance into a mechanistic world view, there are still some elderly people whose mind-sets are still set in the organic worldview. Thus, a mind-set is not dependent on an environment

but on will power. Hence, for the elderly who are still set in their ways of a system that is organic, in the case of African or indigenous communities, this organic way of life is what they grew up knowing. Their value system of harmony with nature and others, harmony with the ought, harmony with how a family should function is now either an idea or a fragmented reality for them. Merchant in Zimmerman (1993:270) highlights that the change from one world view came with the change in what is culturally acceptable and unacceptable, it came with a change of attitude that humanity has when it comes to how they perceive and treat nature. Consequently, what lies between weak anthropocentrism and technocentrism are cultural norms, and human perception of the world through their mind-sets and attitudes. Ultimately, the division created by the digital divide within the home is about change in cultural norms, mind-sets and attitudes that the elderly and their digitally exposed and literate grandchildren/children. The elderly are excluded by the modern status quo because they are not removed from the known, the original and natural (the organic world view), which is inconvenient for a system that was designed to alienate humanity from that known.

4.2 Materialism vs Immaterialism

Introducing this paper with Marxian materialism is essential on the grounds of relevance and accessibility. Relevant because all human beings can relate to the fact that they are material and physical, a sufficient condition for embodiment to materially and physically navigate the world. This sufficient condition is further sufficient to access the material world which has socio-economic inequality, accessible digital divide and also an accessible digital exclusion. Merchant on the other hand presents a dualism of materialist/immaterialist. Her portrayal of an Organic Worldview is one which is premised on a unison founded on a spiritual connection that humanity had with the environment. Mother Nature herself is portrayed as a woman with soul and spirit...her body functions with supernatural order. The shift into Mechanistic Worldview is a shift that kills this spirit, which Merchant portrays as the death of nature.

Merchant in Zimmerman (1993:270) mentions "As Western Culture came increasingly mechanized in the 1600's, the female earth and virgin earth were subdued by the machine." When the mechanistic worldview takes over, in comes a world with no spirit but a wake of a science enlightenment in societies. A wake of materialism that is convenient because everyone can access it by virtue of being physical entities. Also convenient because materialism allows for socio-economic inequalities and difference attained by materials. Materials such as the kind of access and ICT device(s) a person has that influences their experience with technology. Materials which exclude those who do not have materials and include those who have.

4.3 Passing down of values compromised

Through the mentioned avenues which promote capitalist ideals, namely strong anthropocentrism and technocentrism, digital exclusion is manifested in the different forms of alienation mentioned in the paper. The alienation that indigenous and or rural African communitarian societies experience from the digital space is premised on space and time. Space being geographical and on the outskirts of the city, outside of development parts of the country. In terms of time, they are more in tune with the organic worldview, where they are in union with nature and the environment. Hence, at a time such as this where the world is developing technologically, attaching to worldviews outside of the mechanical, digital world is in itself an alienating exercise. In addition, being in a place of poverty and lack of development in rural Africa further isolates and digitally excludes. This digital exclusion becomes as personal as inside the household, where different generations are exposed to or have different digital literacies. When it comes to the elderly and their grandchildren in particular, this division has or is compromising the passing down of African values. Moreover, Aminin (2018:170) states that in order for a nation or society to restore its values, there needs to be some form of education or “conscientizing” within the home first, instilled in children whilst they are young at home. Which brings one to the last section that will talk about a specific “conscientizing” that will be suggested for this paper. This is ecological conscience by Aldo Leopold.

5 The way Forward: digital conscience

Before defining ecological conscience, a visit into the historical archives becomes a necessary endeavor. An endeavor which firstly illustrates the importance of Merchants organic and mechanistic worldviews, secondly, the importance of how far back humanity has been constantly conscientizing through their lifestyle way before they could even be called modern. This archival venture goes as far back to one of the first human civilization: Ancient Mesopotamia located in Fertile Crescent. Within this fertile location, agriculture became way of life. Farming knowledge anchored Mesopotamians to one place for long time, (Xianhoau, 2018: 195). Consequently, it is not surprising that agricultural way of life and farming knowledge, which required efficient land/ environmental governance by humanity. Understanding that taking care of land meant taking care of humanity, which complements the harmony of the organic worldview mentioned by Merchant.

This introduces Land Ethic by Aldo Leopold (1949), which is premised on communitarian concept, which is an individual being a member of a community of interdependent parts. Land Ethic aims to reignite human relation to land, animals and plants, (Leopold, 1949:204). This is a concept in line with the unity and order existing between humanity and mother nature. Where human beings are intentional with how they relate with nature, which is manifested through a “considered preference” found in weak anthropocentrism. With modernity and technological development, this preference has become a “felt preference”, which is here Leopold’s concern lies. Leopold argues that with modernity relation has become underdeveloped because

more emphasis has been placed on an economic relation, where ultimately human privilege occurs with no obligations (and accountability) (Leopold, 1949:205). For Leopold, the discussion of human obligations is one which must be discussed in light of the human conscience, as conscience and obligations go hand in hand, (Leopold, 1949:209). It is this point of Leopold's discussion that the ecological conscience is realized. A conscience which should not only stay as an idea but must become a part of everyday life for the modern humanity especially. An ecological conscience is all embracing and all inclusive, understanding that for the ecological system to work all species within the system need to be considered. Exclusion of one specie would be crippling the system as a whole. Inclusion of all species means holistic, coherent and united system.

Mother nature's body needs all her organs to not just function, but to function well together. Mother nature, as much as she is a force, she is "voiceless", so to consider her is to consider the voiceless, the fragile, the excluded like the elderly in rural areas. An ecological conscience considers the voiceless and excluded as well, as it understands the power of harmony and unity. The digital space should learn from the ecological system which functions under an all-inclusive ecological conscience. There should be an all-inclusive digital conscience that will inspire human beings to do internal changes that will change their priorities and help change conversations about the digital space, change actions, change how this space is governed, etc. In the African communitarian societies, an inclusive digital space will allow the elderly to continue their important role of making an important impact into society by instilling values in the youth. Contemporary youth who are entrenched in digital space need this as much, if not more, than the previous youth.

6 Conclusion

It takes a particular system to exist for there to be an attainment and sustenance of inequality. Capitalism has shown to be one of those systems. This paper has demonstrated that through the avenues of anthropocentrism and technocentrism, capitalism not only sustains inequality but also creates divisions in society. Marx's illustration of the division that capitalism brought between the proletariat and the bourgeoisie demonstrates this and how it is sustained. With the introduction of technology, the divisions in society and within the home have also permeated into the digital space. The problem with this division is that in developing countries, the elderly living in rural areas are left at the periphery, and are excluded on the grounds of age, location and digital literacy. As the custodians of the value systems of society, the exclusion brought on by capitalism and technology compromises the role of the elders in their communities. They are challenged to pass down knowledge to their children and grandchildren with whom they have been digitally (and geographically) alienated from.

Further, the alienation is one which holistically affects humanity, who have been alienated from a healthy, considerate relationship with nature. Technological development can challenge the harmony that nature has with humanity, birthing a

mechanistic world. A worldview which prioritizes on the capitalistic avenues of strong anthropocentrism and technocentrism. To remedy this division in society, a lesson needs to be learnt from the harmony of the organic worldview, where communitarianism as one of its values aims to unite all those who live within society. This why Leopold's ecological conscience is so important, as it reminds humanity of the benefits of such harmony. Consequently, a digital conscience is needed to create harmony even in the midst of technological development, and further include those at the periphery, such as the rural elderly in Africa. This way, when integrated into society, they will be able to fulfil their duty and responsibility of passing down values and knowledge to the youth.

References

- Amin, S. (2018). The Communist Manifesto, 170 Years Later. *Sociological Review*, 2, 430-452
<https://doi.org/10.5937/socpreg52-16323>
- Aminin et al. (2018, April 4). Sustaining civic-based moral values: insights from language learning and literature. *International Journal of Civil Engineering and Technology (IJCIET)*, 9(4), 157-174.
https://iaeme.com/MasterAdmin/Journal_uploads/IJCIET/VOLUME_9_ISSUE_4/IJCIET_09_04_018.pdf
- Bimber, B. (1990). Karl Marx and the Three Faces of Technological Determinism. *Social Studies of Science*, 20, 333-551. <https://doi.org/10.1177/030631290020002006>
- Brennan, K. (2015) Beyond Technocentrism: Supporting constructionism in the classroom. *Constructivist foundations*. 10(3), 289-304. <http://constructivist.info/10/3/289>
- Bryan G., Norton, B.G. (1984). Environmental Ethics and Weak Anthropocentrism," *Environmental Ethics*, 6(2), 131-148. <https://doi.org/10.5840/enviroethics19846233>
- Correa, T., Paves, I. (2016). Digital Inclusion in Rural Areas: A Qualitative Exploration of Challenges Faced by People From Isolated Communities. *Journal of Computer Mediated Communication*, 21(2016), 247-263. <https://doi.org/10.1111/jcc4.12154>
- Deen-Swaray, M. (2016, June 10). Toward digital inclusion: Understanding the literacy effect on adoption and use of mobile phones and the Internet in Africa. *Information Technologies & International Development* [Special Issue], 12(2), 29-45. <http://itidjournal.org/index.php/itid/article/view/1504.html>
- Haenssger, J. (2017). The struggle for digital inclusion: Phones, healthcare, and marginalisation in rural India Marco. *World Development*, 104, 358-374. <https://doi.org/10.1016/j.worlddev.2017.12.023>
- Hao et al. (2017). Social participation and perceived depression among elderly population in South Africa. *Clinical interventions in aging*, 12, 971-976.
<https://doi.org/10.2147/cia.s137993>

- Jaeggi, R. (2016). What (if Anything) Is Wrong with Capitalism? Dysfunctionality, Exploitation and Alienation: Three Approaches to the Critique of Capitalism. *The Southern Journal of Philosophy*, 54, 44-65. <https://doi.org/10.1111/sjp.12188>
- Jirata, T. J., Simonsen, J.K. (2014). The roles of Oromo-speaking children in the Storytelling Tradition in Ethiopia. *Research in African Literatures* 45(2), 135-149. <https://doi.org/10.2979/reseafritlite.45.2.135>
- Kopanina et al. (2018). Anthropocentrism: More than Just a Misunderstood Problem. *Environ Ethics*. 31, 109-127. <https://doi.org/10.1007/s10806-018-9711-1>
- Kopanina, H. (2019). Anthropocentrism and Post-Humanism. International Encyclopaedia of Anthropology 1, 1-8. <https://doi.org/10.1002/9781118924396.wbiea2387>
- Lembani et al. (2020). The same course, different access: the digital divide between urban and rural distance education students in South Africa, *Journal of Geography in Higher Education*, 44(1), 70-84. <https://doi.org/10.1080/03098265.2019.1694876>
- Leopold, A. (1949). *A Sand County Almanac and sketches here and there*. Oxford university press:Oxford.
http://www.eebweb.arizona.edu/faculty/Bonine/Leopold1949_GreenLagoons-150158.pdf
- Low, C. (2007). Khoisan wind: hunting and healing. *Journal of the Royal Anthropological*, 13(1), 71-90. <http://doi:10.1111/j.1467-9655.2007.00402.x>
- Ndegwah, D. J. and Kroesen, J. O. (2012, November): *Technology transfer and the soul of Africa*. <http://atpsnet.org/conferences/presentations/index.php>
- Nicholas, D., Rowlands, I., Clark, D. and Williams, P. (2011), "Google Generation II: web behaviour experiments with the BBC", *Aslib Proceedings*, 63(1), 28-45. <https://doi.org/10.1108/00012531111103768>
- Orthofer, A. (2016, June). Wealth Inequality in South Africa: Evidence from survey and tax data. REDI3x3 Working Paper 15, 1-50. <http://www.redi3x3.org/paper/wealth-inequalitysouth-africa-evidence-survey-and-tax-data>
- Stork, C. & Schmidt, J.P. (2009). Towards Evidence Based ICT Policy and Regulation: eSkills. Johannesburg: Research ICT Africa, 1(3), 1-25 (2009).
- Xianhua, W. (2019, March 1). *State and Empire in Early Mesopotamia*. https://www.socionauki.ru/upload/socionauki.ru/journal/seh/2019_1/195-216.pdf

Technology and Ethics for Alternative Medicine

T V Gopal^[0000-0003-2572-2627]

Co-Ordinator, Center for Applied Research in Indic Technologies [CARIT] & Professor
Department of Computer Science and Engineering
College of Engineering, Guindy Campus
Anna University Chennai
- 600 025, INDIA

e-mail: gopal@annauniv.edu; gopal.tadepalli@gmail.com

Abstract. "Alternative", "Integrative" or "Complementary" or "Traditional" Medicine indicate therapeutic methods used instead of the existing ones in medical practice. Medical research studies involving people are called clinical trials. There are two main types of trials - interventional and observational. The former is specific to a given intervention. The latter is generic. This paper revisits the historic use of Cybernetics in Medicine and adapts the computation modelling for Alternative therapies.

Keywords: Integrated Health Care, Cybernetics, Therapeutic Methods, Clinical Trials

1 Introduction

Any methodology that forces one to forge a connection between one's mind and body is a candidate for healing arts. There have been quite a number of such healing arts and today a few of them survived the test of time. A few of them that have been into practice for three generations ideally withing a given family or a given region on the land is said to be a viable "alternative" medicine.

The practice of medicine in industrialised countries is popularly termed scientific medicine. The usage of the term science in this context is not really to diminish the art in the practice but to focus on training that is supposed to teach physicians to apply scientific knowledge to people in a rational way. In this sense medicine is an applied science. The entire process of making the practice of medicine as an applied science has invoked many debates, controversies and more important comparisons with "alternative" medicine [Glymour, 1983]. These ideas are not merely about the terms and terminology. Knowledge and its representation have been at the crux of many crucial differences in the methodologies for the practice of medicine. Knowledge at the fingertips [Sparrow, 2011] gradually began to mean the computer keyboard and the search engines. The "scientific" medicine thus radically differs from the "alternative" medicine that relies mostly on nature and natural aspects of maintaining the health records all over the basic elements of nature. There is no unique

methodology that applies to all the healing arts. Life is precious and hence civilised societies began to rapidly adopt the “scientific” medicine. Schools teaching the practice of medicine became brands in themselves. So did the drugs. While brand name drug refers to the name giving by the producing company, generic drug refers to a drug produced after the active ingredient of the brand name drug. Generic drugs are sold under different brand names, but will contain the same active ingredients as the brand-name drug. “Scientific” Medicine thrives of social acceptability and is built on trust. “Alternative” medicine is invariably based on faith. There are many differences in the practice of these medicinal systems and there is no denial of the fact that “alternative” medicine has produced magical healing effects all over the world. However, most of them are invariably the “practitioner” or “healer” centric thus leaving scope to umpteen biases in every deliberation include patient education and rights.

There are four reasons for proposing the Cybernetic Model for Alternative Medicine. They are as follows.

1. Alignment with the Management Requirements
2. Integration
3. Change Management
4. Reduced Time-to-Market

Medical Cybernetics is the science of control in complex dynamic medical systems. Currently, it is closely connected with medical informatics - the science of obtaining, processing and transmitting medical information. One of the core challenges is that alternative therapies have little or no method of docketing the Patient Health Records. They are usually specific to a given patient and are not likely to be amenable to the interventional clinical trials [Saunders, 2001]. They are also replete with Confirmation Bias, Sampling Bias, and Brilliance Bias. It is useful to note that there are more than 50 cognitive biases reported in the scientific literature. Norbert Wiener was operating with the triad "substance-energy-information" for the Cybernetic Control Systems. This forms a more generic basis for the possible framework to integrate the medicinal systems.

“Virtual Clinical Trials” represent a relatively new method of collecting safety and efficacy data from clinical trial participants, from study start-up through execution to follow-up. These trials take full advantage of technologies [mobile apps, monitoring devices, sensors and so on] and online social engagement platforms to conduct each stage of the clinical trial from the comfort of the patient’s home- - including recruitment, informed consent, patient counselling, through to measuring clinical endpoints and adverse reactions. The usage of proxies is also in vogue due to the Usability Engineering based trials. However, the biology holds the key in the usage of proxies demanding advanced technologies.

The major developments in cybernetics and cyber-biology are occurring when computers are linked directly to the human / animal nervous system. Simple deep brain stimulators and cochlear implants have been around for years, and breakthroughs in this area have been staggering. It is possible to use the nerves of the

arm to control bionic arms or even other machines. Computer-controlled electrodes can be used to stimulate hand muscles by moving the shoulder.

A bio-electronic implant, which is about the size of a pencil eraser, would actually sit behind the retina at the back of the eyeball, and images would be transmitted to the brain via a connector the width of a human hair. One can see images better.

A digital video camera can replace the eyes of a blind man. Electrical signals are processed by a micro-computer and then transmitted to the nerves in the visual cortex by way of electrodes giving the blind man an archaic but effective vision as bright dots, resembling a stadium display.

Cyborgs brought in a wide range of Electronic Implants [Greguric, 2014]. The miniaturization and speed of computer components allows the representation of models and systems of great complexity, with many interacting elements at a variety of scales. The electronic surveillance systems can simultaneously follow and handle millions of people. Each of us has a unique bioelectrical resonance frequency in the brain, just like we have unique fingerprints.

With electro-magnetic frequency (EMF) brain stimulation fully coded, pulsating electromagnetic signals can be sent to the brain, causing the desired voice and visual effects to be experienced by the target. Researchers say the technology is currently available to implant biometric devices in human beings, which can be monitored by software, satellites and utilized by Government and Industry. The automaton needs to be architected in conformance to the Cybernetic Limits.

Statistical thinking uses data to separate variation from special causes and variation from common causes. One can draw patterns and correlations from the data but cannot check validity. A mechanism for a causation theory cannot be found as well. The more rigorous a trial is scientifically, the less generalisable are its findings to real-world settings. The concept of the “Efficacy-Effectiveness Gap” (EEG) has started to challenge the confidence in the decisions pertaining to even the modern medicine.

Computational Intelligence Techniques in smart healthcare systems and personalized healthcare applications, especially in the areas of model-based healthcare treatment, physiological control systems, alternative medical knowledgebased systems are being increasingly deployed.

*"The art of healing comes from nature, not from the physician.
Therefore the physician must start from nature, with an open
mind."*

- Philipus Aureolus Paracelsus, 1536


Nature and harnessing the natural forces is the common basis across all medicinal systems. The methodologies vary significantly and the artistic component integral to every such methodology makes it unique and also dependent on the “healer”. “Scientific” medicine is more amenable to the objective deployment of technology. In other words it facilitates the visualization of the “objective reality”.

2 A Framework for Integration

Norbert Wiener's axiom states that "Information is information, not matter or energy". Prima facie, information can be culled out of matter and energy with a representation that enables its objective processing. However, control is the crux and it is not innate in the axiom of Norbert Wiener [Coming 2001]. Cybernetics has a very limited tool for social sciences. However, cybernetics is vital for assuring "human use of human beings" in the alternative medicinal systems.

Good Architecture for the Medical Cybernetics is an asset that enables: **Speed, Flexibility, Choice and Interoperability** within the framework [Microsoft, 2009] in table 1 given below.

Table 1: A Framework for Integration of Alternative Medicinal Systems

		Healthcare Services for Patients and Society <ul style="list-style-type: none"> • Automatic health monitoring • Prevention through continuous care and lifestyle coaching • Better and faster treatment • Elderly and ill can live at home rather than in a nursing home or hospital • Greater life expectancy • Fewer absences from work • Lower health insurance premiums and lower public spending on healthcare 			
		Primary Characteristics of Integration			
Interfaces and Adapters <ul style="list-style-type: none"> • Effective and Efficient Interfaces • Evolutionary Approaches • Scalability and Mobility. 	Healthcare Portals & Management Healthcare Management Information System	Data and Process Models - Types of Data [Structured, Semi-Structured and Unstructured]; Temporal Constraints; Distributed Databases; Data Analytics;	Products & Tools <ul style="list-style-type: none"> • Clinical Decision Support system • Medical Devices and Implants 	Standard Reference Models & Medical Bibliographic Systems <ul style="list-style-type: none"> • HIPAA • DIMAT 	Ecosystems of the Partners <ul style="list-style-type: none"> • Medical Errors • Healthcare Waste • Healthcare Readiness Indices & Maturity Models • Workflows
Affordable, Available and Assurable Technologies					
Sensing and Data Acquisition: Integration of Sensors	Cloud and Mobile Computing Based Services: <ul style="list-style-type: none"> • Geographic Information Systems • Patient and Hospitality Locator Services 	Mobile Applications: <ul style="list-style-type: none"> • Audio & Video Transmission • Hand-Helds • Social Networks 	Human Computer Interaction: <ul style="list-style-type: none"> • Haptic & Tele-Haptic • Voice Recognition 		

“The rapid advances in device monitoring capabilities have, to some extent, outstripped the ability of many clinics to leverage

the plethora of available diagnostic information. However, many implantable devices now also contain telemetric capability that allows the device to transmit all sensor derived parameter trends from the patient's home directly to the clinic. This is typically achieved through a bedside monitoring / telemetry device that can be programmed by the clinic to transmit automatically based on preset schedule. Alternatively, a transmission of device stored monitoring and diagnostic data can also be triggered based on detected clinical events. For example, some devices can be programmed to alert the clinic directly if the patient experiences the onset of atrial tachyarrhythmias or if the ventricular rate during a sustained atrial tachyarrhythmia exceeds a pre-programmed threshold. Such remote monitoring 'care alerts' may be quite useful to monitor rate and rhythm control strategies. These remote monitoring capabilities also foster the potential to transmit nondevice recorded information automatically, such as weight, blood pressure and associated symptoms back to the managing clinic. Thus, the stage is set for a new paradigm of heart disease management based on continuous remote monitoring rather than intermittent clinic visits."

- Michael Gold, Yong Cho, Tom Bennett and Douglas Hettrick, 2008

3 Quality Indicators

A quality indicators-based framework [Onyebuchi, 2006] is being proposed can serve the long – term strategic needs for integrating various medicinal systems internationally. India has taken the lead in Yoga and other therapies and there is copious traditional knowledge and is also tacit to traditional healers. On a long term the proposed cybernetics based observational framework can systemically support the various medicinal systems that are founded on traditional healthcare. The quality indicators being proposed are given in table 2 below.

Table 2: Quality Indicators for Alternative Medicine

1. Acceptability	8. Clinical Focus
2. Accessibility	9. Efficiency
3. Appropriateness	10. Equity
4. Capacity	11. Patient Centeredness
5. Competence or Capability	12. Safety
6. Continuity	13. Sustainability
7. Effectiveness	14. Timeliness

It is important for the cybernetic model to ensure ethical equilibrium that can be monitored using the quality indicators. Monitoring is tooling or a technical solution that facilitates and objective methodology of watching and understanding the state of the systems based on gathering predefined sets of metrics or logs stemming from the quality indicators. The technologies to measure the quality indicators are rapidly advancing and many of them are affordable.

Observability [van, 2016] is tooling or a technical solution that allows teams to actively debug their systems. “Observability” that includes the “healer –centric” practices needs the Cybernetic ethics [Henry, 2015] as the basis. Cybernetic ethics is a way of viewing the evolution of ethical systems in terms of the informational feedback certain human actions generate.

Knowledge is, after all, what we know. And what we know can't be commodified. Perhaps if we didn't have the word 'knowledge' and were constrained to say 'what I know', the notion of 'knowledge capture' would be seen for what it is - nonsense!

F.J. Miller, 2002

"The web of our life is of a mingled yarn, good and ill together."

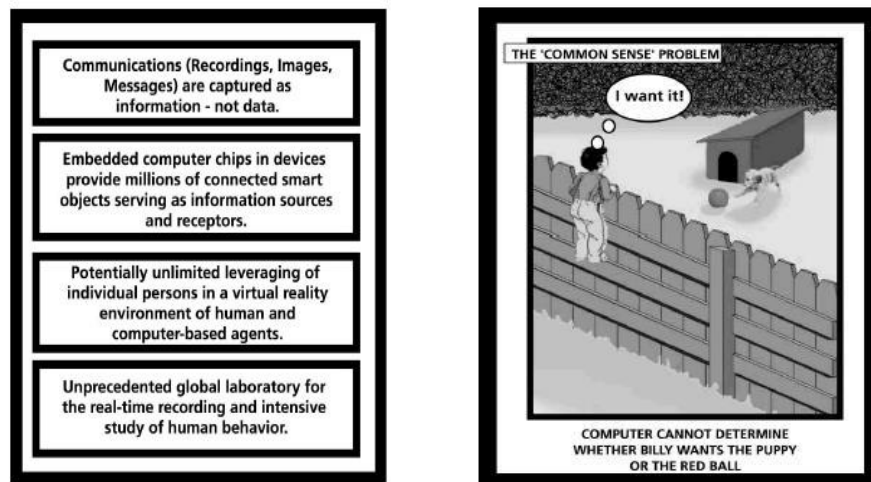
William Shakespeare, The Fourth Part of All's Well that Ends Well act 4, scene 3.

4 Cybernetic Ethics

Cyberspace describes an emerging environment in which most information about physical objects such as manufactured products, buildings, processes, organizations, artifacts, human beings, and dialog between human beings, is accessible on-line through computer-based communication systems [Norbert, 1950]. Further, it is expected that in Cyberspace both the user and the environment will be able to analyse, synthesize and evaluate information in a virtual reality environment that couples to the human senses through, at the very least, sound, speech synthesis, threedimensional space, and animation.

Distinguishing between “information” and “knowledge” is very important. There is no direct relationship between the thing you are talking about and the words you use. It is the processes of comprehension, understanding and learning that go on in the mind and only in the mind of human that lend meaning to the words used. In other words, unless the various symbolisms are mediated through "personal" experience they have no meaning. Thus, information is intrinsically meaningless on its own and remains so unless interpreted by human beings, within some context. It is unfortunate that the wide range of communication devices and systems are inducing us into ascribing intrinsic meaning to the information. The assumption that the impersonal stimuli [machine -to-machine] have minds of their own and can have their own meaning often times makes one unwittingly cross the boundaries of rationality and

enter a bizarre world. The figure 1 given below is an illustration of the need for distinguishing between “information” and “knowledge”.



Cyberspace

The Difficult Problem

Figure 1: Distinguishing between “Information” and “Knowledge” in Cyberspace

The profound belief in many alternative medicinal systems is that the Human Body communicates the ailment to the healer. Such a belief obviates any medical instrument to diagnose. The current technology makes it a reality.

Human Body Communication (HBC) uses human body itself as a communication route to transmit data [Jian, 2017]. It usually operates at dozens of kHz to dozens of MHz, because at these frequencies the propagation loss along the human body is smaller than that through the air. This feature makes it be especially promising in the healthcare, medical diagnoses, consumer electronics and user identification applications, and be also suitable to establish a body area network.

Body Area Network (BAN) uses the surface of the human body as a network transmission path [Carlos, 2021]. Communication starts when the skin comes in contact with a transceiver and ends with physically separation. The system works through shoes and clothing as well. The human body becomes a secure communication channel as shown in the figure 2 below.

A theoretical framework for observability in “alternative” medicinal systems is feasible to enable the integration of medicinal systems. A human being can be modelled as a communicating entity very much like Norbert Wiener modelled the computer as a communication device [Florentin, 2014].

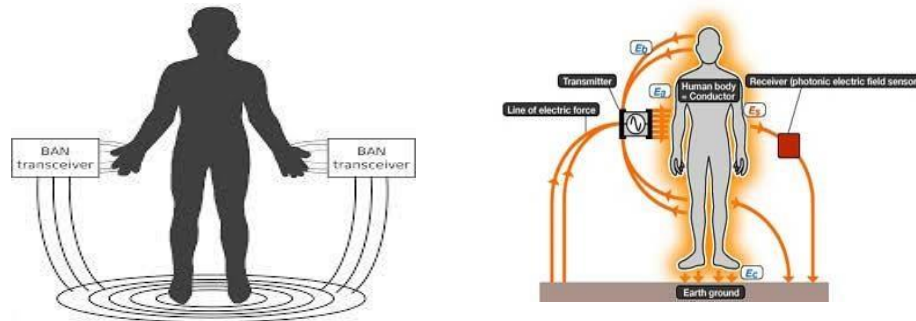


Figure 2: Human Body Communication and the Technology Equivalent

Potential advantages include [Meredith, 2021]:

- services tailored to the individual needs of the user;
- as communication is triggered by natural human actions, there is no need to insert smart cards, connect cables, tune frequencies;
- setup, registration, and configuration information for an user can all be uploaded to a device the instant the device is touched, eliminating the need for the device to be registered or configured in advance;
- tables, walls, floors and chairs can act as conductors and dielectrics, turning furniture and other architectural elements into a new class of transmission medium. For example, you could have instant access to the Internet by placing a laptop onto a conductive tabletop.
- the system could be installed on any locations calling for secure access, such that each secure access could be initiated and authenticated with a simple touch.

Human body that has both digital (= synaptic) and analog (= humoral) parts.

Norbert Wiener insists on a balance of power between analog and digital aspects.

“To what extent is any given man morally responsible for any given act? We do not know.”

- Alexis Carrel, 2010

To enable human beings to reach their full potential and to live a good life, according to Wiener, the society must uphold the following principles.

- **The Principle of Freedom** – Justice requires “the liberty of each human being to develop in his freedom the full measure of the human possibilities embodied in him”.
- **The Principle of Equality** – Justice requires “the equality by which what is just for A and B remains just when the positions of A and B are interchanged”.

- **The Principle of Benevolence** – Justice requires “a good will between man and man that knows no limits short of those of humanity itself”.
- **The Principle of Minimum Infringement of Freedom** - “What compulsion the very existence of the community and the state may demand must be exercised in such a way as to produce no unnecessary infringement of freedom”.

“Law may be defined as the ethical control applied to communication, and to language as a form of communication, especially when this normative aspect is under the control of some authority sufficiently strong to give its decisions an effective social sanction. It is the process of adjusting the “couplings” connecting the behavior of different individuals in such a way that what we call justice may be accomplished, and disputes may be avoided, or at least adjudicated. Thus the theory and practice of the law involves two sets of problems : those of its general purpose, of its conception of justice; and those of the technique by which these concepts of justice can be made effective.”

- Norbert Wiener, 1948

The American Psychological Association published “Varieties of Anomalous Experience,” covering enigmas from near-death experiences to mystical ones. Neurotheology [Christopher, 2002] is stalking bigger game than simply affirming that spiritual feelings leave neural footprints, too. Some of the early results in this field include:

- **Attention:** Linked to concentration, the frontal lobe lights up during meditation.
- **Religious emotions:** The middle temporal lobe is linked to emotional aspects of religious experience, such as joy and awe.
- **Sacred images:** The lower temporal lobe is involved in the process by which images, such as candles or crosses, facilitate prayer and meditation.
- **Response to religious words:** At the juncture of three lobes, this region governs response to language.
- **Cosmic unity:** When the parietal lobes quiet down, a person can feel at one with the universe.

“It is certain that thought may be transmitted from one individual to another, even if they are separated by long distance. These facts, which belong to the new science of metaphysics, must be accepted just as they are” “They express a rare and almost unknown aspect of ourselves”... “What

extraordinary penetration would result from the union of disciplined intelligence and of the telepathic aptitude"

-Alexis Carrel, 2010

The "Healer's Eye" is vital in perceiving the ailment. The third eye (also known as the inner eye) is a mystical concept referring to a speculative invisible eye which provides perception beyond ordinary sight. It is often associated with religious visions, clairvoyance, the ability to observe chakras and auras, precognition, and out-of-body experiences. People who are claimed to have the capacity to utilize their third eyes are at times known as seers. Indic traditions associated healing powers with seers.

C.W. Leadbeater claimed that by extending an "etheric tube" from the third eye, it is possible to develop microscopic and telescopic vision. It has been asserted by Stephen Phillips that the third eye's microscopic vision is capable of observing objects as small as quarks.

A new theory of how the brain works is being called "Neural Transduction Theory". This theory may be crucial for understanding consciousness and the universe itself. Every pertinent aspect of healing is then an induction from the cosmos into the mind of the healer.

5 Conclusions

"Alternative" medicinal systems have always been distinct from the "Scientific" Medicine that is now global. Transpersonal psychology is a field or school of thought in psychology centered on the spiritual aspects of human life. This branch of psychology is very useful at the first level of integration even with the Artificial Intelligence and Machine Learning systems.

"Transpersonal psychologists attempt to integrate timeless wisdom with modern Western psychology and translate spiritual principles into scientifically grounded, contemporary language. Transpersonal psychology addresses the full spectrum of human psychospiritual development—from our deepest wounds and needs to the existential crisis of the human being, to the most transcendent capacities of our consciousness."

- Mariana Caplan, 2009

"The popularised word "togetherness" aptly captures a general notion of human proximity, of meeting and speaking, or dancing together at a festival. Social groups, be they families urban communities Or the older universities, have institutions which promote togetherness; the dining table , a market, or a cafe as the case may be. On more or less ritualized occasions, and in the traditional places, humans converse; either verbally, or by image and gesture. I submit that the conversation which occurs,

debate and sometimes agreement , is the stuff of civilised life and togetherness is essential to it. On the other hand there are also limits upon "togetherness"; too much of it, for example. gives rise to specific symptoms of individual and social malaise. These symptoms typically appear when the communication, allowed by proximity is not conversation. Communication and conversation are distinct, and they do not always go hand in hand."

– Gordon Pask, 1980

References

- Alexis Carrel, (2010), "Man the Unknown", Wilco Publishing House, Mumbai, India
- Carlos A. Tavera, Jesús H. Ortiz, Osamah I. Khalaf, Diego F. Saavedra, Theyazn H. H.
- Aldhyani, (2021), "Wearable Wireless Body Area Networks for Medical Applications", Computational and Mathematical Methods in Medicine, vol. 2021, Article ID 5574376
- Christopher C Hook, (2002) "In Whose Image? Remaking Humanity through Cybernetics and Nanotechnology," Dignity 8, No. 1, Pp 2–3.
- Corning Peter, (2001), "Control information": The Missing Element in Norbert Wiener's Cybernetic Paradigm?", Kybernetes. Vol. 30, Pp 1272-1288
- Florentin Smarandache and Ștefan Vlăduțescu, (2014}, "Communicative Universal Convertibility Matter-Energy Information", Social Sciences and Education Research Review, Vol.1, Pp 44 – 62.
- Glymour C and Stalker D., (1983), "Engineers, Cranks, Physicians, Magicians", New England Journal of Medicine, Vol. 308, Pp 960-964.
- Gordon Pask, (1980) , "The Limits of Togetherness", Proceedings, Invited Keynote address to IFIP, World Congress in Tokyo and Melbourne, Editor, S. Lavington. Amsterdam, New York, Oxford: North Holland Pub. Co, Pp 999-1012
- Greguric, Ivana. (2014) "Ethical issues of human enhancement technologies: Cyborg technology as the extension of human biology." Journal of Information, Communication and Ethics in Society. Vol. 12, No. 2. Pp 133–148.
- Henry, D., & Nielsen, K. (Eds.). (2015). "Bridging the Gap between Aristotle's Science and Ethics". Cambridge University Press, Cambridge, United Kingdom
- Jian Feng Zhao, Xi Mei Chen, Bo Dong Liang, Qiu Xia Chen, (2017), "A Review on Human Body Communication: Signal Propagation Model, Communication Performance, and Experimental Issues", Wireless Communications and Mobile Computing, Vol. 2017, Article ID 5842310.
- Meredith Trejo, Isabel Canfield, Whitney Bash Brooks, Alex Pearlman, and Christi Guerrini, ,

- (2021), “A Cohort of Pirate Ships”: Biomedical Citizen Scientists’ Attitudes Towards Ethical Oversight”, *Citizen Science: Theory & Practice*.
- Michael Gold, Yong Cho, Tom Bennett and Douglas Hettrick, (2008), “Heart Failure Management - Monitoring with Implantable Devices”, *Asian Hospital and Healthcare Management*, Issue 17, Pp 31-35
- Microsoft, (2009), “Connected Health Framework Architecture and Design Blueprint - A Stable Foundation for Agile Health and Social Care”, Part 2 – Business Framework, USA
- Miller F.J, (2002), “ $I = 0$ (Information has no intrinsic meaning)”, *Information Research*, Vol. 8, No. 1. paper no. 140.
- Norbert Wiener, (1948), “Cybernetics: Or Control and Communication in the Animal and the Machine”, MIT Press, Cambridge, MA, USA
- Norbert Wiener, (1950). “The Human Use of Human Beings: Cybernetics and Society”, Houghton Mifflin Co., USA.
- Onyebuchi A Arah, Gert P Westert, Jeremy Hurs and Niek S Klazinga, (2006), “A conceptual framework for the OECD Health Care Quality Indicators Project”, *International Journal for Quality in Health Care*, Pp. 5–13.
- Saunders J. (2001). “The Practice of Clinical Medicine as an Art and as a Science”, *Western Journal of Medicine*, Vol. 174, No. 2, Pp 137–141.
- Sparrow B, Liu J and Wegner DM, (2011), “Google effects on memory: Cognitive consequences of having information at our fingertips.”, *Science*. Vol. 333, Pp 776778.
- van de Poel Ibo. (2016). “An Ethical Framework for Evaluating Experimental Technology”, *Science and Engineering Ethics*, Vol. 22, No. 3, Pp 667–686.

Acknowledgements

The author places on record sincere thanks to the stakeholders in the Canada – India CuRE Healthcare Project Proposal in the year 2012. The stakeholders include McGill University, Canada, IIIT, New Delhi, Apollo Telemedicine Network Foundation [ATNF], Chennai, Tata Consultancy Services, Chennai, C-DAC, Mohali and Computer Society of India. The author was service as the Technical Lead for this proposal and Anna University has signed the pertinent ethics statements.

The author thanks Anna University for all the support.

Endless Forms Most Deceitful

William M. Fleischman¹, Nick Langan¹ and Leah N. Rosenbloom²

¹ Villanova University, Villanova, Pennsylvania, U.S.A.

² Brown University, Providence, Rhode Island, U.S.A.
william.fleischman@villanova.edu

Abstract. We consider the effects on technology and information professionals of working in an environment where they are routinely subject to dishonest or disingenuous communication from managers, and are aware that public pronouncements of corporate spokespersons are similarly lacking in candor.

Keywords: Dishonesty, Facebook, Ethics

1 Introduction

We consider the effects on technology and information professionals of working in an environment where they are routinely subject to dishonest or disingenuous communication from managers, and are aware that public pronouncements of corporate spokespersons are similarly lacking in candor.

Though we are far from asserting that Facebook (now Meta) is unique in this respect, the company nonetheless provides a rich harvest of examples upon which we gratefully draw. We attempt a partial taxonomy of forms of dishonest discourse including instances both inward facing and outward facing.

In the final section of our paper, we lift the focus beyond that of an individual organization to the problem of working in a global informational environment where there is no shared commitment to honest discourse.

2 Endless Forms Most Appalling

2.1 Examples and a Taxonomy

We classify examples as inward-facing (**IF**: Facebook employees are the primary audience) or outward facing (**OF**: the intended audience includes the general public and/or governmental personnel). Inward-facing statements often form part of discussion over policy on Workplace, a network only Facebook staffers can access. Resonance between inward- and outward-facing statements is significant since Facebook employees can compare the company's public assertions with the tenor of often contentious internal debate. Occasionally internal discussions become part of the public record as a result of disclosure by whistleblowers. Internal records sometime "disappear" after disclosure resulting in seriously negative publicity. We further categorize statements as:

- Loss Leader (**LL**): A statement that pleads the enormous effort in service of the public put forth by Facebook to address a problem, absent evidence of effectiveness and present evidence of persistence of the problem.
- Moving Target Admission (**MT**): One of a **series** of acknowledgments of the gravity of one particular problem, that it's being addressed but, in the nature of things, "we need more time to fix it."
- Gaudy but Irrelevant Statistic (**GI**): Given the magnitude of Facebook's wealth of money and personnel, any statement asserting enormous resources are being devoted to a problem is, on its face, unassailable. We reserve this designation for instances in which massive investment of resources is asserted as a mask for resistance to making unpalatable policy changes.
- Environmental Informational Racism (**EIR**): Statements conveying the message that negative effects of Facebook policies and practices on marginalized and non-English speaking populations are not, in the grand scheme of things, significant nor are those populations worthy of a duty of care.
- Delay, Linger, and Wait (**DLW**): A small war of attrition, usually in the form of an exchange of worker concerns and inward-facing delaying and deflecting responses as illustrated in the next subsection.

We begin with a Loss Leader from Facebook CEO Mark Zuckerberg in testimony to a U.S. congressional committee: "We are committed to keeping people safe on our services and to protecting free expression, and we work hard to set and enforce policies that meet those goals. We will continue to invest extraordinary resources into content moderation, enforcement, and transparency." (Tag: **OF, LL**) (Mac, et al, 2021) Disclosures concerning human trafficking (Linebaugh, 2021) and hate speech give the lie to the commitment to "keeping people safe," while Facebook's protection of free expression evidently ends where the displeasure of authoritarian governments threatens its revenue flow (see **EIR** examples).

Moving Target Admissions are a staple of Facebook's defensive posture given its troubled history with hate speech. Among serial admissions involving hate speech cited in Marantz (2020) we have Guy Rosen, Facebook vice-president for integrity:

"We don't allow hate speech on Facebook. While we recognize we have more to do ... we are moving in the right direction." (OF, MT) And spokesperson Drew Pusateri:

"We've invested billions of dollars to keep hate off of our platform. ... A recent European Commission report found that Facebook assessed 95.7% of hate speech reports in less than 24 hours, faster than YouTube and Twitter. While this is progress, we're conscious that there's more work to do." (OF, MT, GI) This qualifies under GI since hate speech that goes viral (among the 95.7 % speedily identified or the 4.3% tardily caught) renders the eye-catching expenditure completely irrelevant.

From Sophie Zhang's angry farewell memo: "In the three years I've spent at Facebook, I've found multiple blatant attempts by foreign national governments to abuse our platform on vast scales to mislead their own citizenry ... We simply didn't care enough to stop them." When Zhang alerted higher-ups to a known bug exploited by the government of Honduras to create fake accounts, fake followers, and fake "likes" for promoting and disseminating pro-government propaganda, an executive told her, "I don't think Honduras is big on people's minds here." (Halpern, 2021) (IF, EIR) Similarly, when the Turkish government threatened to shut down Facebook unless it blocked posts from a group alerting Syrian Kurds of impending Turkish attacks, Sheryl Sandberg, Facebook's C.O.O., responded, "I'm fine with this."

(Halpern, 2021) (IF, EIR)

2.2 Delay, Linger, and Wait

This practice signals that employee values do not align with company priorities and has a pernicious effect on the morale of those attempting to maintain the organization's purported ethical standards.

Between December 2017 and January 2018, Facebook personnel in London called attention to Britain First, characterizing it as a hate group. On the first occasion, public policy director Neil Potts responded, "Thanks for flagging and we are monitoring this closely." However, he claimed that while it was a close call, the group did not meet Facebook's definition of hate group: "... *those that advance hatred as one of their primary objectives, or ... have leaders who have been convicted of hate-related offenses.*" (IF, DLW) When another worker noted that one of its leaders, Jayda Fransen, had been convicted of hate crimes against British Muslims, Potts replied, "Thanks for flagging. 'I'll make sure our hate org subject-matter experts are aware of this conviction.'" (IF, DLW)

A month later, absent action banning the group, the employee posted, "Happy new year! The Britain First account is still up and running, even though ... it clearly violates our community standards. Is anything being done about this?" "Thanks for circling back," Potts responded, (IF, DLW) after which the thread went dormant.

Subsequently, a white Briton, radicalized by following Britain First on social media, was convicted of the July 2017 murder by vehicular assault of one Muslim man and injury to twelve others near a London mosque. “Within six weeks, Britain First and [its prominent ally] Tommy Robinson had been banned from Facebook.

(Pusateri, the Facebook spokesperson, noted that the company has ‘banned more than 250 white supremacist organizations.’)” (OF, LL)) (Marantz, 2020)

This example bears comparison with the persistent managerial indifference to alarms raised by Roger Boisjoly and colleagues about O-ring reliability prior to the fateful launch of the Challenger space shuttle. Engineers whose judgment was overridden suffered deep demoralization and long-term psychological injury.

2.3 Effects on Workers

Individual voices reported in our references, almost without exception, describe the prevailing work atmosphere in terms saturated with disappointment, resentment, exasperation, and demoralization. This appears to be a matter of longstanding – a chronic condition rather than a phenomenon restricted to this or any other particular moment. It persists up to the present moment in regard to recurring inconsistencies and deficiencies in dealing with both disinformation and hate speech in contexts as varied as the Covid-19 pandemic, the violent insurrection of 6 January 2021 in the U.S. and the Russian invasion of Ukraine. (Mac et al, 2022; Timberg et al, 2022) Marantz, for example, reports instances of frustration among content moderators about inconsistent, *ad hoc* application of Facebook’s published standards reaching back more than eight years. These predate even the notorious 2015 decision to leave standing a post advocating “a total and complete shutdown of Muslims entering the United States” including the assertion that [all Muslims, all 1.8 billion of them] “have no sense of reason or respect for human life.” That decision precipitated an internal firestorm not entirely extinguished by the rationalization that it was an instance in which an exception was made to Facebook’s rules concerning hate speech because of the “newsworthiness” of the post.

It transpired that this sort of carve-out was subsequently formalized in an internal document, entitled Known Questions, provided to moderators as a guide for interpreting the Implementation Standards. By 2017, the guide stipulated that posting content calling for the exclusion of a group sharing a protected characteristic from entering a country or continent, something previously considered hate speech, was to be permitted. Among the illustrative examples cited was the former “exceptional” statement, “I am calling for a total and complete shutdown of Muslims entering the United States.” There is little need to guess at the demoralization of those who had criticized the original decision predicting, “Once you set a precedent of caving on something like that, how do you ever stop?” (Marantz, 2020)

One recurrent point of resentment is the immorality of holding employees to NDAs while Facebook, piously asserting its benign role in the world, repeatedly courts harm by violating its published standards.

Consider this poignant statement by an anonymous FB employee: “Nobody wants to look in the mirror and go, I make a lot of money by giving objectively dangerous people a huge megaphone.” (Marantz, 2020) What a revealing way of characterizing the degrees of freedom this individual enjoys as ethical professional.

The very latest flare-up of concerns of this sort forms a backdrop to Facebook’s aggressive push to recruit new users in Africa and the global south to counterbalance the recent hemorrhage of users Facebook has sustained in the developed world. Critics, including Facebook whistleblower Frances Haugen, express concern “over an apparent lack of safety controls” in Africa, particularly relating to Covid-19 and vaccine-related disinformation. As usual, Facebook deploys its public relations machine, extolling its “global team of 40,000 working on safety and security, including 15,000 people who review content in more than 70 languages – including Amharic, Somali, Swahili and Hausa,” (Tag: ?) and reminding critics that misinformation is a “complex and constantly evolving societal challenge for which there is no ‘silver bullet.’” (Tag: ?)

Simultaneously, critics lament the lack of direct local investment and the scale of the effort in relation to the seriousness of the problem: “There seems to be as little as possible real investment on the continent in terms of engaging people directly or hiring people with real local knowledge,” and “Currently, in comparison to the wealth of the company and its social responsibility ... it [resources committed to content moderation] is pretty minimal.” (Burke, 2022) The force of these concerns is heightened by the terms of Facebook’s campaign for new users: The offer is free internet, with the stipulation that access is filtered through Facebook. “Across Africa, Facebook *is* the internet.” (Malik, 2022) Thus, Facebook’s serious shortcomings in regard to disinformation are strongly magnified.

We leave as an exercise for the reader the application of appropriate tags to the indicated statements above but offer the following hints from the testimony of Frances Haugen: “... though only 9% of Facebook users speak English, 87% of the platform’s misinformation spending is devoted to English speakers.” And “It seems that Facebook invests more in users who [generate the most profit], even though the danger may not be evenly distributed based on profitability.” (Silberling, 2021)

2.4 The Particular Plight of Content Moderators Hired as Contract Workers

Although a small fraction of its 15,000 content-moderators are Facebook employees, Facebook relies heavily on contract labor to do the job. Ellen Silver, Facebook’s vice president of operations, said in a blog post last year that the use of contract labor allowed Facebook to “scale globally” – to have content moderators working around the clock, evaluating posts in more than 50 languages, at more than 20 sites around the world. (Silver, 2018)

The use of contract labor also has a practical benefit for Facebook: it is radically cheaper and thus helps Facebook maintain a high profit margin. The median Facebook employee earns \$240,000 annually in salary, bonuses, and stock options. (Price, 2018) Vendor-based content moderators, hired by major consulting firms like Accenture and Cognizant, are paid a small fraction of the average pay of Facebook

employees, do not receive healthcare and other benefits comparable to Facebook personnel, and work in what amount to glorified “call centers.” They are bound by strict nondisclosure agreements (NDAs) which, although ostensibly intended to prevent leaks of users’ personal information and protect moderators from retaliation by irate users whose content has been taken down, also serve the comforting purpose of insulating Facebook and its vendors from uncomfortable scrutiny about working conditions and the severe emotional toll on moderators exacted by the nature of the work. (Newton,

2019a)

Aside from complaints about inconsistent enforcement of platform rules concerning hate speech and misinformation, all content moderators also have to deal with repeated exposure to some of the most toxic material circulating on the web. This content includes text, images and video depicting child sexual abuse, violent assault, rape, murder, and suicide. It cannot come as any surprise that content moderators, in general, are susceptible to work-related psychological harm, from periodic attacks of anxiety and panic to post-traumatic and secondary stress disorder (PTSD and SSD).

(Innodata, 2021)

For contract workers, the traumatic effects of exposure are intensified by the pace imposed by supervisors who set additional stress-inducing targets for efficiency (process items rapidly, at least 200 per day) and ‘accuracy’ (interpret correctly Facebook’s mutable, often vague policies, in at least 95% of cases), as measured by periodic Facebook audits. Workers are subject to a daily schedule where breaks for lunch, bathroom, and “wellness” are calibrated to the second. Demerits that can affect their salary and retention are assessed for ‘excessive’ time away from the work screen and marginal slippage in efficiency or accuracy. (Newton, 2019a)

One supervisor of content-moderators testified, “You can ask for a meeting, present your bosses with bullet points of evidence, tell them you’ve got team members who are depressed and suicidal – doesn’t help. Pretty much the only language Facebook understands is public embarrassment.” (Marantz, 2020)

When it comes to psychological counseling, the same differential conditions apply.

“Facebook is very proud of the fact that Facebook employees get proper psychological support,” [Martha] Dark [co-founder of Foxglove, the UK non-profit advocacy group that has been assisting content moderators in their efforts to improve working conditions] says. “But outsourced moderators who do exactly the same job, and look at exactly the same content, have a [tightly scheduled] ‘wellness session’ once a week. This is not proper, meaningful, clinical long term mental health support in the way that they get at Facebook.” Naturally enough, ‘wellness sessions’ are conducted by ‘wellness coaches’ – Facebook NewSpeak for someone who is not a medical doctor, nor even a trained psychologist, but is nevertheless tasked with looking after moderators. Under their contract, ‘wellness coaches’ can’t diagnose PTSD and can’t treat it. All they can do is tell workers to try things like karaoke or deep breathing to cope. (Foxglove, 2021)

And after the job ‘eats them up’ and they are terminated, contract workers lack access to long-term psychiatric care. (Newton, 2019a)

In 2020, Facebook was forced to make a \$52 million settlement to its vendor-based content moderators in a class-action lawsuit over psychological damage. It cannot come as a surprise, then, that several consulting firms providing content moderators on contract now require new hires to sign a ‘voluntary’ statement acknowledging that their jobs may negatively affect their mental health and cause PTSD.

Sign a statement. That’ll fix it!

2.5 Rhetoric vs. Reality

In the spirit of our analysis, we should like to ask, “What are Facebook executives saying about their responsibilities toward those contract workers carrying out this essential service for the company upon whom they are indirectly visiting such misery?” We should like to ask, as well, “What are executives of the consulting services saying about their responsibilities toward these workers whose service they exploit, directly exposing them to grievous harms, paying a wage not even twice the Federal Poverty Level, reaping contracts worth almost three times what they ‘invest’ in those workers?”

Let’s look at the evidence. “We are grateful to the people who do this important work to make Facebook a safe environment for everyone,” Facebook said in a statement. “We’re committed to providing them additional support through this settlement and in the future.” (Newton, 2020) And Ellen Silver, Facebook’s vice president of operations, “We care deeply about the people who do this work. They are the unrecognised heroes who keep Facebook safe for all the rest of us. We owe it to our reviewers [moderators] to keep them safe too.” (Boran, 2020)

Silver’s anodyne rhetoric is routinely echoed by spokespersons for the consulting firms that supply content moderation cannon fodder. (Perrigo, 2021)

We have refrained from applying a taxonomic category to these statements because we lack an appropriate label for such outrageous lies. It is only necessary to juxtapose the chilling testimony of numerous contract content moderators (Bernal, 2021; Gray, 2019; Scullion, 2020; Newton, 2019a; Perrigo, 2022) concerning the appalling conditions under which they labor – under intense pressure to ‘clear’ 150 or more toxic ‘tickets’ a day, while maintaining an “accuracy” rate in the high 90’s, measuring their off-screen time to the carefully watched second – to understand the enormity of the sham.

Mark Zuckerberg’s leaked response, when questioned about this in a private meeting with Facebook employees, artfully mixed in a touch of **LL**, **MT**, and implicit **GI** to craft a finely textured tissue of lies that transcends mere taxonomic categorization:

“We work with different outside firms so, that way, we can scale up and down and work quickly and be more flexible on that. It’s one of the main reasons we do it around the world in different places, get people to work in all the different languages. But there are the challenges that you’re saying, which is we want to make sure that these folks who are affiliated with the company and very much part of our family as a company are treated well

and have the same kind of support that employees would have when dealing with difficult jobs which a lot of people here have. Some of the reports, I think, are a little overdramatic. From digging into them and understanding what's going on, it's not that most people are just looking at just terrible things all day long. But there are really bad things that people have to deal with and making sure that people get the right counseling and space and ability to take breaks and get the mental health support that they need is a really important thing. It's something we've worked on for years and are always trying to probe and understand how we can do a better job to support that." (Newton, 2019b)

In the end, perhaps, it isn't principally a matter of words. Perhaps the question most pertinent here is, "How do Facebook and consulting firm executives conceive their relation to these workers in the context of the relation of one human being to another?" Here it doesn't seem so difficult to read. Considering the mutual disclaimers: "FB doesn't apply quotas and put pressure on consulting firms," and "We [the consulting firms] have got to meet FB standards for accuracy and efficiency," the two parties are scrupulously free of any responsibility for harms suffered by content moderators. The really important thing, of course, is to 'scale globally,' to 'scale up and down and work quickly,' never mind the damage to the poor, thin-skinned devils whose reports of trying to process graphic depictions of child sexual abuse, beheadings and savage assaults are likely 'a little overdramatic.'

In fact, massive evidence of the pressure-filled robotic discipline imposed on contract content moderators suggests they are seen by Facebook and its vendors as nothing more than imperfect, but cheap and disposable, carbon-based substitutes for the mythical Artificial Intelligence Engine, the eagerly awaited perfection of which will cleanse the Internet of all the toxic filth and hate that Facebook and other social media platforms amplify and perpetuate. We set this formulation – content moderators as cheap, disposable computing units – side by side with Kant's Categorical Imperative – so 18th century, don't you know? – to help us all appreciate the profound contribution our present-day Philosopher King, Mark Zuckerberg, has made to the foundation of ethics.

3 Pushing Back Against the Megalith

Short of quiescence in the face of refractory response from their superiors, what are the options open to technical professionals raising serious criticism of company policies?

3.1 Actions by Groups and Individuals

One possibility is concerted action by a large group as in the mass walkouts by Google workers in November 2018 and the Facebook walkout of June 2020. In the case of Google, the action was the result of disclosure of the \$90,000,000 severance package awarded to Andy Rubin in spite of his forced resignation under cloud of credible sexual harassment accusations. Estimates indicated that 20,000 workers participated in the action worldwide.

The Facebook walkout was precipitated by widespread disgust among Facebook employees about Mark Zuckerberg's refusal to take down or label the former U.S. president's post including the phrase, "when the looting starts, the shooting starts," at the time of massive demonstrations in response to the police murder of George Floyd.

Both these actions received widespread attention in the press and were effective in exposing the level of internal dissatisfaction with specific company policies. However, in each case it took an event of exceptional notoriety to inspire a large body of workers to participate in a narrowly focused protest.

Frances Haugen's disclosures as a whistleblower testifying before the subcommittee on consumer protection of the U.S. Senate Commerce Committee is another means of bringing to public attention policies concealed by Facebook destructive of the public interest. Haugen appears to have been motivated by what she considered two signal failures of the platform to subordinate economic interest to the general good. The action proximate to her revelations was the company's decision to shut down its Civic Integrity Team in December 2020, an action that she and many others considered reckless in view of the possibility of unrest fueled by misinformation and incitement to violence around the January 20th, 2021, transfer of power from the administration of the defeated incumbent president to the newly elected president.

The second major failure was the fact that, despite internal research that strongly suggested that the Facebook-owned app Instagram could cause harm to teenage girls, the company failed to implement recommended changes to mitigate this possibility. In both these cases, Haugen contends Facebook prioritized user engagement over the public good to push provocative and emotionally triggering posts to the top of users' newsfeeds. Haugen's actions have succeeded in drawing aside the curtain and stimulating intense public discussion on policies and practices that Facebook would otherwise have wished to maintain veiled in secrecy. Her testimony before the U.S. Congress has led to new proposals for legislation and regulation to mitigate some of the harms resulting from unconstrained social media platform policies.

Frances Haugen's role as whistleblower stands at one extreme of the spectrum of ethical professional behavior. It is worthwhile recalling the words of Terry Winograd in his 1991 keynote address to the first National Conference on Computing and Values, a direct precursor of the ETHICOMP series:

My emphasis will be on the fundamentally social nature of ethical concerns: with looking beyond the role of the individual to the larger context of discourse and action that generates the world in which individuals make choices and act. Rather than

focusing on the isolated individual faced with an ethical dilemma, I want to direct our gaze to the larger swirl of human discourse, which is the source of the interpretations, values, and possibilities that make ethical choice meaningful.

The announcement for the NCCV conference declared a vision:

To integrate computer technology and human values in such a way that the technology advances and protects those values rather than doing damage to them.

This will require acts of individual moral courage, and it will be based on a lot more. We need to create an environment in which the consideration of human values in our technological work is not a brave act, but a professional norm. We need to produce a background of understanding in which it is simply taken for granted by all computer professionals that value considerations are foremost. We need to forge everyday practices and ways of teaching that reinforce that understanding. (Winograd, 1992)

Let us stipulate, also, that, although not at the same level of individual heroism, the mass actions of Google and Facebook employees are in their way brave acts in the sense of Winograd. Each is an act at the same end of the spectrum as Haugen's, rising to meet a particular occasion. In contrast, what we should like to ask here is whether, in the face of the massive imbalance of power between workers and management, in the face of the reflexively disingenuous rhetoric of Facebook executives, there exist resources for developing the background of understanding envisioned by Winograd so that value considerations and ethical practices are everyday professional norms?

3.2 The Integrity Institute

The Integrity Institute was founded in 2021 by two data scientists, Sahar Massachi and Jeff Allen, who bring to the undertaking critical experience working on integrity issues with Facebook between 2016 and 2019. Massachi was a member of the Civic Integrity team. Allen describes his experience as working on systemic issues in the public content ecosystems of Facebook and Instagram.

The Integrity Institute represents itself as an honest broker for reliable information, education, and evolving best practices regarding integrity issues and how to address them. In this capacity, the Institute takes its mission to be that of serving relevant communities including the public, journalists, social media platforms professionals, academics, and legislators.

One of the animating visions of the Integrity Institute is Sahar Massachi's conceptualization of the social media platform as a virtual city in which the role of the integrity worker is that of a city planner, more precisely a member of a team of city planners for the platform. Instead of focusing on the 'law enforcement' aspects of integrity work – content moderation, takedowns, and other sanctions on users – he

places the focus on incorporating incentives and friction (Massachi envisions the latter as ‘driver’s exams’ and ‘speed bumps’) that promote a more harmonious, less toxic, virtual polity into the design of the platform’s information ecosystem. He thinks of this as restoring some of the constraints of real-world physics that limit the harms that can occur in a real city. In his view, the community of integrity workers has developed a repertoire of measures to prevent or mitigate many of the harms that plague social media platforms, and this collective intelligence should be available as an open resource for platforms small and large. (Massachi, 2021)

In founding the Integrity Institute, Massachi and Allen have the ambition to see the institute evolve into a forum for the theory and practice of protecting the social internet. Building on the varied expertise and experience of a growing number of institute members representing many different backgrounds and technical roles, they aspire to build an open-source integrity team modeled on that of the Open Web Application Security Project (OWASP), an online community founded in 2001, that promotes and disseminates freely-available methodologies, documentation and tools for web application security.

If the collective efforts of the institute community succeed in the necessarily dynamic project of building – and establishing a public voice for – a theoretical framework and a library of best practices and tools for protecting the informational environment and related systems of the social internet, one can foresee a limited, but important, set of favorable consequences. It will raise the expectations of the general public in regard the quality of their experiences as users of social media platforms. Similarly, it can serve to sharpen the focus of academic studies and journalistic investigation into platform policies and performance. It can serve to inform the crafting of legislation and appropriate regulation for standards of integrity. And, perhaps most important, it can serve as a ‘backbone-stiffener’ for social media professionals in advocating for more enlightened company policy. As Massachi is quick to point out, “For too long, integrity work has been a public service trapped within private entities. We’ve been identifying ways to build better cities and fighting to get them implemented, but if our proposals cut against other company goals, they might not see the light of day.” (Massachi, 2021) The work of the institute may provide a means of arguing internally for a different balance between the integrity goals and economic prerogatives of the company.

3.3 Legislative Initiatives – PATA

The evolution of the major speech platforms – preeminently Facebook and Google – in supplanting traditional print and broadcast media as the principal source of news for large segments of the public, has amplified the scale of harms associated with their inconsistent application of standards for controlling the spread of disinformation and hateful speech. Growing public attention to these harms has resulted in efforts to draft new legal and regulatory frameworks to constrain the practices of social media.

In a sense, this has been a project seeking an achievable focus. In its early stages, discussions often revolved around the question of whether, in this capacity, social

media platforms should “be seen as neutral technology companies that have simply built an ingenious set of tools for passing along content including entertainment, artistic creation and news, or have they, in fact, evolved to play a role in the sphere of public information comparable to traditional media companies[?]” (Fleischman & Rosenbloom, 2020)

In his brilliant essay, “Is the First Amendment Obsolete?”, Tim Wu discusses the challenges and paradoxes attached to First Amendment jurisprudence in the age of “cheap speech” facilitated by social media platforms. Noting the diminished role of direct censorship in the U.S. (but not, apparently in the Russia of Vladimir Putin’s latter day imperialistic adventures), Wu catalogues the familiar techniques destructive of a healthy informational environment – troll armies, reverse censorship, flooding, and propaganda robots – then underscores the changed stakes and consequences of weaponized speech: “In the 1990s, trolls would abuse avatars, scare people off AOL chatrooms, or wreck virtual worlds. Today, we are witnessing efforts to destroy the reputations of real people for political purposes, to tip elections, and to influence foreign policy. It is hard to resist the conclusion that the law must be enlisted to fight such scourges.” (Wu, 2017)

One of the ironies attached to this exercise is the peculiar role of the First Amendment, guaranteeing every American his or her sacred right to freedom of speech. Thus, those whose posts have been taken down or who have been ‘deplatformed’ often complain bitterly about the abridgement of their First Amendment rights. But these complaints are ignorant and without constitutional merit. Since the platform is a private enterprise, it is completely within its rights to establish and enforce standards that prohibit otherwise privileged speech. (What a marvel if major speech platforms were to publish and rigorously enforce reasonable standards prohibiting violent or hateful speech. ¡Ojalá!)

On the other hand, to perfect the irony, any attempt to eliminate hateful or violent speech through legislation or regulation compelling platform action toward this goal would fall afoul of First Amendment scrutiny. “To level set, most direct regulation of harmful, but legal, online content would violate the First Amendment. With a few notable exceptions, the regulation of hate speech, disinformation and (most forms of) incitement, cannot be done through outright bans or takedown mandates.” (Persily, 2021) The lesson here seems clear. In devising legislative or regulatory remedies, don’t fall into the snare of attempting to legislate the limits of permissible speech. (Don’t try, through law or regulation, “to call the balls and strikes” of permissible speech.)

The Data Care Act of 2018 (passed in the U.S. Senate but died in the House of Representatives; re-introduced in 2021) embodies one approach to avoid this trap.

Sponsored by Sen. Brian Schatz (D-HI) and based on the work of Jack Balkin and Jonathan Zittrain, the bill seeks to assign legally binding ‘fiduciary’ duties, similar to those doctors, lawyers, and accountants owe their clients, to technology companies that handle user data. By formalizing intuitive duties owed by platforms to their users, a fiduciary framework could provide leverage for individuals and regulators looking to hold platforms accountable for failures presently beyond redress. (Bowers, 2018)

The latest, and most interesting, legislative initiative is the Platform Accountability and Transparency Act (PATA), based on the work of Nathan Persily of Stanford Law School, introduced by Sens. Chris Coons (D-Delaware), Rob Portman (R-Ohio) and Amy Klobuchar (D-Minnesota). The critical need for transparency – which in the end may prove partially achievable – was set in sharp focus by the 2016 U.S. Presidential election, the Brexit referendum and Carol Cadwalladr’s reporting on Facebook’s attempt to conceal its role and the extent of its negligent culpability in the Cambridge Analytica scandal. (Cadwalladr, 2017, 2019) Persily also acknowledges the profound debt owed Frances Haugen for providing a detailed picture of the extent of Facebook’s internal knowledge of harms attributable to Facebook products and policies, the granularity of internal evidence for those harms, and the near perfect exclusion of the public from knowledge of this information. (Persily, 2021)

“That equilibrium—where firm insiders know everything and the rest of us are left to guess—is unsustainable. Facebook and the other Silicon Valley Platforms have lost their right to secrecy. We need national transparency legislation that will allow researchers, other than those tied to the profit-maximizing mission of the firms, to get access to the data that will shed light on the most pressing questions related to the effects of social media on society.” (Persily, 2021)

Key provisions of the bill provide for

- Submission of proposals to the National Science Foundation by university-affiliated researchers. For proposals supported by NSF, social-media platforms would be required to furnish the needed data, subject to privacy protections that could include anonymization or the use of ‘clean rooms’ for review of sensitive material under platform supervision.
- Authorization for the Federal Trade Commission to require regular disclosure of specific information by platforms, including ad micro-targeting.
- Provision of a legal safe harbor for both researchers and platforms protecting them from legal liability relating to privacy concerns and preventing platforms from blocking independent research initiatives.

Among the goals of research conducted under PATA are better understanding of the role algorithms play in amplifying harmful content including the influence of recommendations and algorithmic curation; better understanding of role that advertising plays in promoting hate, disinformation, incitement and other illegal activity online; the nature, extent, and revenue realized through the sale of microtargeting on the platform.

Since the shadow of Cambridge Analytica hovers over PATA, the act provides for vetting of proposals by a trusted governmental agency, the NSF, with stiff penalties to the university and individual researchers for data access malfeasance. To prevent leaks compromising user privacy, “the data reside at the firm, which is responsible for maintaining security of the research environment and monitoring all research conducted therein. Researchers need to be monitored whenever they are in touch with the data. Every keystroke must be recorded as the data analysis is conducted. Researchers may not take any data out of the research environment without a privacy review being conducted.” And “[a]lthough the government will be heavily involved in

enforcing the program of researcher access, the **datasets themselves should never be placed in government hands. It is absolutely critical that there be no risk of government surveillance or privacy intrusions as a result of this program.**"

[emphasis added] (Persily, 2021)

To further protect user privacy, all data must be anonymized or pseudonymized, with use of additional measures including differential privacy and construction of synthetic data sets where appropriate. Under the act, platforms will be protected against tort law claims for violations of user privacy: "If the platform follows all applicable regulations concerning protecting privacy in the research environment, then it will be immune from suit for the fact that it made such data available under this program. To be clear, this does not immunize them from harms identified by the researchers. If the platform is discovered to be acting fraudulently or contributing to offline harm, then that information might later end up in a lawsuit or even a criminal prosecution." (Persily, 2021)

Notwithstanding apparent meticulous care in drafting, serious questions have been raised concerning PATA by researchers otherwise favorably disposed to the purpose of the act. Daphne Keller of the Stanford Center for Internet and Society expresses reservations as to whether the act fully ensures that "researchers don't inadvertently create new privacy risks – like by storing data insecurely, or inadvertently publishing data that can be used to identify people? Is after-the-fact liability enough?"

Given that much of what we know about the role of social media in past abuses, such as the Cambridge Analytica scandal, is the work of dedicated, resourceful, and responsible journalists like Carol Cadwalladr (and despite the queasy matter of propagandists masquerading as "journalists"), Keller questions the act's reticence on a defined role for journalists and civil society organizations. (Keller, 2022)

Elsewhere, Keller makes the deeply discouraging point that, although the NSF and the Federal Trade Commission are agencies that normally command general respect for probity and scrupulousness, the current atmosphere of political polarization and the last Presidential administration's ruinous successes in politicizing and corrupting major governmental departments and agencies has resulted in a reflexive reaction of distrust that may prove an obstacle to passage of PATA. (Keller, 2020)

The really big problem here is how to find a planet and a country in which it's actually possible to pass reasonable and meaningful legislation.

3.4 The Unequal Contest

Mark Zuckerberg is prepared "to go to the mat" in case of hostile action by Elizabeth Warren or Lina Kahn, the newly appointed activist chairperson of the U.S. Federal Trade Commission. (Pengelly, 2019) Neither the U.S. Congress nor the Parliament of Great Britain seems capable of taking the measure of Zuckerberg's evasions and deceptions. (Cadwalladr, 2019) What hope is there, then, for a promising outcome in a contest between conscientious Facebook workers and the likes of Zuck, Nick Clegg, and Sheryl Sandberg? How does the conscientious individual stand up against the

tools they so easily deploy – a new flavor yogurt in the lunchroom fridge, solemnly intoned deceptions and evasions, chilly unresponsiveness, or simply being sacked?

How does one maintain ethical autonomy and agency in this space?

4 The Global Disinformational Environment

In prescient 1994 essay, “A Revolution of Values,” feminist and educational scholar bell hooks warned of the mounting crisis fueled by “a lack of meaningful access to truth.” The crisis, hooks explains, is not simply the presence of lies, but also the telling of lies “in a manner that enables most effective communication.” The violent, destabilizing, and rapidly spread disinformation on social media platforms clearly falls into this category, but the disinformation communicated by managers and company representatives to employees and the public is perhaps even more effective and dangerous in the long term. The kinds of lies presented in our taxonomy and throughout this work illustrate what hooks calls our “collective cultural consumption of and attachment to misinformation” – a culture in which people (sometimes knowingly, sometimes unknowingly) use lies to reproduce systems of domination that severely undermine the values they purport to uphold. (hooks, 1994) Indeed, the greatest “stumbling block” in the way of our efforts to dismantle the rampant culture of disinformation online may not be the blatantly violent and villainous lies of terrorists and dictators, but the paternalistic lies that are “more devoted to order than to justice” – narratives that aggressively maintain the status quo and silence dissent – all the while insisting to agree with proposals for change. (King, 1963)

Going deeper, hooks notes a symbiotic relationship between systemic oppression and disinformation: “a culture of domination necessarily promotes addiction to lying and denial.” (hooks, 1994) This is because systems of domination force the overwhelming majority of people (for instance the non-billionaire or non-Englishspeaking populations) into positions that are unsafe and uncomfortable. Rather than admit that our everyday actions and habits of being are directly connected to systems that dominate us, hooks notes that people have a tendency to use familiar lies as “stabilizing traditions” – for example, the misinformed but widespread idea that corporations have the consumers’ best interests at heart (otherwise they would lose their business!), despite overwhelming evidence to the contrary.

A global informational environment dominated by social media – such that open competition or challenge from inside or outside the company is systematically rooted out and extinguished – necessarily promotes an accompanying cultural apparatus of lies and denial. And conversely, cultural attachments that permit, welcome, and perpetuate lies and disinformation serve to empower systems of domination and oppression. Thankfully, there is a large body of extant work on how to dismantle the systems of oppression that thrive on and create the dominating culture of lies and denial at Facebook and in the tech industry in general. This work can serve as a foundation for people like Carol Cadwalladr, Sophie Zhang, and Frances Haugen (and all the conscientious Facebook and Google employees whose names we do not know

because of NDAs) who are working to change the ways people engage with disinformation in the digital age.

Following Paulo Freire and hooks' conception of education as the practice of freedom (Freire 1974; hooks, 1994), intersectional feminist and educational scholar Bettina L. Love outlines a concrete framework for abolishing systems of oppression through the vehicle of liberatory education, called the abolitionist teaching methodology (Love, 2019). The framework synthesizes many of hooks' suggestions and provides concrete examples of each in the context of education. We believe this framework has immense value not only in the struggle to dismantle the culture of disinformation in the digital age, but also to work toward building a culture truly grounded in the values of freedom and justice for all. While we regrettably do not have the space to fully explore the abolitionist teaching, we briefly recap core tenets of the practice and encourage the reader to peruse Love's book, *We Want to do More than Survive: Abolitionist Teaching and the Pursuit of Educational Freedom* (2019).

First, abolitionist teaching practice is grounded in the power of community involvement. It asks practitioners to identify and engage communities who might be harmed by the issues that practitioners are trying to address. For activists in the digital disinformation space, this might involve seeking out and hearing the testimony of people who have been persecuted or harmed as a result of disinformation, whether members of the non-English-speaking majority communities who have been harmed by the platform itself, or content moderators and technology professionals who have been harmed by tech companies' inward-facing policies. Second, abolitionist teaching requires the context of history, civics, and resistance. The disinformation tactics employed by dictators and Facebook representatives alike are not new; there are countless examples throughout history – as well as examples of how these tactics can be illuminated and overcome – that can inform our struggle to dismantle the global disinformational environment. The third tenet is the acknowledgement of intersectionality – the fact that systems of oppression (and disinformation) are inextricably intertwined, and that facing one of them means facing the rest. Similarly, it is necessary for us to confront whiteness: the system of racial and caste domination that implicitly favors the comfort and wellbeing of white, upper-class, Englishspeaking people over the lives of non-white, poor, and non-English-speaking people.

It is worth interjecting here that the abolitionist teaching practice espoused by Bettina Love and bell hooks stand in direct opposition to the oblivious, truth-averse proposals being advanced by certain white parents and officials, to remove from public school curricula and libraries all mention of disturbing history relating to slavery, discrimination, and the selective denial of the full rights of citizenship to people of color, and those of non-traditional sexual orientation and gender identification. In the current unhappy U.S. political climate, the ideas of abolitionist teaching practice are strong and salutary medicine for a sick nation.

The final two core tenets of abolitionist teaching are very personal, and strike at the heart of the dismissive and dehumanizing culture of Big Tech: the pursuit of joy and wellbeing, and freedom dreaming. This is the idea that we, as a society, must prioritize the inner, social, emotional, and intellectual wellbeing of each individual

human being on earth, starting with ourselves. Only then will we be able to “dream” of solutions that truly free us from our collective cultural addiction to lies and denial. Grounded in introspection and self-care, the abolitionist teaching framework harmonizes with Winograd’s suggestion that the integration of technology and society must be done in a way that protects the values that unite us as human beings. Abolitionist teaching, along with all of the movements of people that resist technological oppression and the tyranny of the global disinformation environment, necessarily place human values over technical ones. We believe the following prophetic appeal, which Martin Luther King, Jr. made in his 1967 book, *Where Do We Go From Here? Chaos or Community*, aptly summarizes the state of our dilemma in the digital age: “We must rapidly begin the shift from a ‘thing’-oriented society to a ‘person’-oriented society. When machines and computers, profit motives and property rights are considered more important than people, the giant triplets of racism, materialism, and militarism are incapable of being conquered.”

In the spirit of conquering the giant triplets and the endless forms of deceit that fuel them, we advocate for a swift and total abandonment of empty techno-solutionism, with replacement by human-centered solutionism: in the words of King and hooks, a revolution of values.

References

- Bernal, N. (2021, June 11). Facebook's content moderators are fighting back. *WIRED UK*. Retrieved from URL <https://www.wired.co.uk/article/facebook-content-moderators-ireland>
- Boran, M. (2020, February 27). Life as a Facebook moderator: 'people are awful. This is what my job has taught me'. *The Irish Times*. Retrieved from URL <https://www.irishtimes.com/business/technology/life-as-a-facebook-moderator-people-are-awful-this-is-what-my-job-has-taught-me-1.4184711>
- Bowers, J. (2018, December 27). 2018 was a trying year for social media platforms— and their users: Three pathways forward. *Just Security*. Retrieved from URL <https://www.justsecurity.org/62041/2018-year-social-media-platforms-and-users-pathways/>
- Burke, J. (2022, April 18). Facebook ‘Lacks Willpower’ to Tackle Misinformation in Africa. *The Guardian*. Retrieved from URL <https://www.theguardian.com/world/2022/apr/18/facebook-accused-of-lacking-willpower-to-tackle-misinformation-in-africa>
- Cadwalladr, C. (2017, May 7). The great British Brexit robbery: How our democracy was hijacked. *The Guardian*. Retrieved from URL <https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy>
- Cadwalladr, C. (2019, April). Facebook's role in Brexit -- and the threat to democracy. *Carole Cadwalladr: Facebook's role in Brexit -- and the threat to democracy / TED Talk*. Retrieved from URL

- https://www.ted.com/talks/carole_cadwalladr_facebook_s_role_in_brexit_and_the_threat_to_democracy?utm_campaign=social&utm_medium=referral&utm_source=t.co&utm_content=talk&utm_term=technology
- Foxglove. (2021, July 9). What's a blue-chip consulting firm like Accenture got to do with racist memes targeting footballers or self-harm on Instagram? *Foxglove*. Retrieved from URL <https://www.foxglove.org.uk/2021/09/07/accenture-facebook-instagram/>
- Freire, P. (1974). *Education: The Practice of Freedom*.
- Gray, C. (2019, December 9). As a facebook moderator I saw the worst of humanity. we need to be valued | Chris Gray. *The Guardian*. Retrieved from URL <https://www.theguardian.com/commentisfree/2019/dec/09/facebook-moderator-worst-of-humanity-valued>
- Halpern, S. (2021, May 2). Facebook and the Normalization of Deviance. *The New Yorker*. Retrieved from URL <https://www.newyorker.com/news/daily-comment/facebook-and-the-normalization-of-deviance>
- Hooks, b. (1994). *Teaching to Transgress: Education as the Practice of Freedom*. New York, NY: Routledge.
- Innodata (2021) *The Ethics of Content Moderation: Who Protects the Protectors?* Retrieved from URL <https://innodata.com/the-ethics-of-content-moderation/>
- Keller, D. (2020, July 16). CDA 230 Reform Grows up: The Pact Act has problems, but it's talking about the right things. *Center for Internet and Society*. Retrieved from URL <http://cyberlaw.stanford.edu/blog/2020/07/cda-230-reform-grows-pact-act-has-problems-it%E2%80%99s-talking-about-right-things>
- Keller, D. (2022, April 6). User privacy vs. platform transparency: The conflicts are real and we need to talk about them. *Center for Internet and Society*. Retrieved from URL <http://cyberlaw.stanford.edu/blog/2022/04/user-privacy-vs-platform-transparency-conflicts-are-real-and-we-need-talk-about-them-0>
- King, M. L. (1967). *Chaos or community?* Hodder and Stoughton.
- King, M. L. (1963). *Letter from Birmingham City Jail*. American Friends Service Committee.
- Linebaugh, K. & Scheck, J. (hosts). (2021, September 18). *The Facebook Files, Episode 3: 'This Shouldn't Happen on Facebook'* [audio podcast transcript]. Retrieved from URL <https://www.wsj.com/podcasts/the-journal/the-facebook-files-part-3-thisshouldnthappen-on-facebook/0ec75bcc-5290-4ca5-8b7c-84bdce7eb11f>
- Love, B. (2020). *We want to do more than survive: Abolitionist teaching and the pursuit of educational freedom*. Beacon.
- Mac, R., Silverman, C. & Lytvynenko, J. (2021, April 26). Facebook Stopped Employees from Reading an Internal Report about Its Role in the Insurrection. You Can Read It Here. *BuzzFeedNews*. Retrieved from URL <https://www.buzzfeednews.com/article/ryanmac/full-facebook-stopthe-steal-internal-report>

- Malik, N. (2022, January 20). How facebook took over the internet in Africa – and changed everything. *The Guardian*. Retrieved from URL <https://www.theguardian.com/technology/2022/jan/20/facebook-second-life-the-unstoppable-rise-of-the-tech-company-in-africa>
- Marantz, A. (2020, October 12). Why Facebook Can't Fix Itself. *The New Yorker*. Retrieved from URL <https://www.newyorker.com/magazine/2020/10/19/why-facebook-cant-fix-itself>
- Massachi, S. (2021, December 20). How to save our social media by treating it like a city. *MIT Technology Review*. Retrieved from URL <https://www.technologyreview.com/2021/12/20/1042709/how-to-save-social-media-treat-it-like-a-city/>
- Newton, C. (2019a, February 25). The trauma floor. *The Verge*. Retrieved from URL <https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona>
- Newton, C. (2019b, June 19). Bodies in seats. *The Verge*. Retrieved from URL <https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-tampa>
- Newton, C. (2020, May 12). Facebook will pay \$52 million in settlement with moderators who developed PTSD on the job. *The Verge*. Retrieved from URL <https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health>
- Pengelly, M. (2019, October 1). Zuckerberg: I'll 'go to the mat and fight' Warren over plan to break up Facebook. *The Guardian*. Retrieved from URL <https://www.theguardian.com/technology/2019/oct/01/mark-zuckerberg-facebook-elizabeth-warren-big-tech>
- Perrigo, B. (2021, September 23). 'I Sold My Soul.' WhatsApp Content Moderators Review the Worst Material on the Internet. Now They're Alleging Pay Discrimination *Time*. Retrieved from URL <https://time.com/6080450/facebook-whatsapp-content-moderators/>
- Perrigo, B. (2022, February 17). Inside Facebook's African sweatshop. *Time*. Retrieved from URL <https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/>
- Persily, N. (2021, October 26). *Testimony of professor Nathaniel Persily James B. McClatchy professor ...* Retrieved from URL <https://www-cdn.law.stanford.edu/wp-content/uploads/2021/10/2021-10-26/Testimony-Before-the-Senate-Homeland-Security-and-Governmental-Affairs-Committee-including-Proposed-Platform-Transparency-and-Accountability-Act.pdf>
- Price, R. (2018, April 17). The median pay at Facebook is more than \$240,000 a year. *Business Insider*. Retrieved from URL <https://www.businessinsider.com/facebook-median-pay-240000-2017-2018-4>
- Scullion, G. (2020, May 22). Chris Gray 'dealing with PTSD as a result of working as a social media content moderator' on Apple Podcasts. *Bringing design closer with Gerry Scullion*. Retrieved from URL

<https://podcasts.apple.com/us/podcast/chris-gray-dealing-with-ptsd-as-a-result-of/id1450117035?i=1000475475110>

- Silberling, A. (2021, October 5). Facebook whistleblower Frances Haugen testifies before the Senate. *TechCrunch*. Retrieved from URL <https://techcrunch.com/2021/10/05/facebook-whistleblower-frances-haugen-testifies-before-the-senate/>
- Silver, E. (2018, July 26). Hard questions: Who reviews objectionable content on Facebook – and is the company doing enough to support them? *Meta*. Retrieved from URL <https://about.fb.com/news/2018/07/hard-questions-content-reviewers/>
- Timberg, C., Dwoskin, E., & Albergotti, R. (2021, October 29). Inside facebook, Jan. 6 violence fueled anger, regret over missed warning signs. *The Washington Post*. Retrieved from URL <https://www.washingtonpost.com/technology/2021/10/22/jan-6-capitol-riotfacebook/>
- Winograd, T. (1992). Computer, Ethics, and Social Responsibility. *Computing and Human Values: Proceedings of the 1991 Conference, New Haven: Research Center on Computing and Society, 1992.*, 1–18.
- Wu, T. (2018). Is the First Amendment obsolete? *Michigan Law Review*, (117.3), 547. <https://doi.org/10.36644/mlr.117.3.first>

IT ‘Experts’ Considered Harmful: A Concerning Case of False Expertise in the Legal System

Reuben Kirkham¹[0000-0002-1902-549X]

¹ Monash University, Victoria, Australia
reuben.kirkham@monash.edu

Abstract. Over the last few years, this author has engaged in cases before administrative tribunals that relate to matters of Computer Science. The author has encountered what might be termed notions, which are, to put it mildly, strange. For example, he was told that ‘Microsoft Excel’ could only be used effectively by ‘academic computer scientists’, found the tribunal was closed due to an inability of a regulator (of information rights) to organise PDF’s, and (whilst abroad) had a Tribunal try to connect him to a video hearing via ISDN. What is most concerning is that these notions are being supported by self-proclaimed ‘IT experts’, be it those employed by the Courts and Tribunals, regulators, or specialist judges. The issue is the conflation of ‘IT experts’ with computer scientists, and the lack of regulation of ‘IT experts’, who can become recognised as such after completing a few multiple-choice quizzes. I argue that the answer is more effective regulation of people presenting themselves as ‘experts’ and the need for professional bodies to develop appropriate qualifications and assessments to ensure that the wrong people cannot present themselves as experts.

Keywords: expert witnesses; legal systems; IT experts.

1 Some examples that I have encountered...

Administrative tribunals (hereon “Tribunals”) make decisions that impact the lives of a large proportion of the population. For example, each year, UK Tribunals alone hear several hundred thousand social security cases, employment claims, and immigration cases, totaling over 300k cases disposed of each year.¹ These decisions can often be life-changing for those involved (Ellis, 2013).

Increasingly, there is a need for Tribunals to deal with matters of computer science. This can include where these matters relate to a question of fact that a Tribunal is expected to decide, or whether someone can access proceedings remotely using video-conferencing software like Zoom (e.g. due to a disability). My background is that of a computer scientist working on human-computer interaction, rather than being a trained lawyer. Nevertheless, I have gained some experience of how Tribunal’s deal with these

¹ <https://researchbriefings.files.parliament.uk/documents/CBP-8372/CBP-8372.pdf>

matters, having engaged with the Information Rights jurisdiction over the last seven years. In some ways, it has been a surreal experience.

This short article is presented from the viewpoint of ‘showing’, rather than ‘telling’. As such, most of this paper is provided in Appendices, which include copies of Tribunal transcripts and orders. I summarise some examples below, with a selection of further evidence being provided later in this document for the reader to peruse for themselves:

- **Example 1: Closure of the Information Tribunal due to difficulties in organising PDF’s.** The (Information) Tribunal jurisdiction was closed down because the UK Information Rights regulator found it too difficult to organise PDFs into bundles. The then President of the relevant Tribunal accepted this assertion. This was despite a user-friendly guide having been prepared for this exercise.² When this matter was raised in the press, the Judicial Press Office wrote to the journalist threatening to sue the journalist ‘on behalf of the judiciary’. An excerpt from the journalists original article is reported as **Appendix A**.
- **Example 2: Using Microsoft Excel requires an ‘academic’ computer scientist.** A senior Judge responsible for Information Rights asserted that using relatively simple procedures in Microsoft Excel was something that required the “*technical competence of a computer science academic, and not one that may reasonably be expected of a public authority*”. The Judge then attempted to prevent me from obtaining a transcript of the hearing before her. I successfully appealed that decision to the Upper Tribunal with the result of confirming in law the right of litigants to obtain transcripts of hearings (with the case being reported as *Kirkham v Information Commissioner (GIA)* (Tribunal procedure and practice - record of proceedings) [2019] UKUT 381 (AAC)). An account of the worrying situation before the First-tier Tribunal as covered by a member of the press can be found here (<https://www.computerweekly.com/news/252509483/Government-bodies-refuse-FOI-requests-on-basis-of-misleading-database-search-times-says-academic>)
- **Example 3: Trying to do remote hearings via ISDN.** Before the pandemic, a Tribunal attempted to force me to use ISDN at the cost of around \$1000 an hour to connect to hearings. In another case, a different Tribunal struck it out for my moving abroad, on the basis it was too challenging to do remote hearings. This strike out was overturned by an exasperated judge of the Upper Tribunal, with his observations on this being provided as **Appendix B**. The issue of access to online justice using an appropriate system was finally resolved by Upper Tribunal Judge Thomas Church, who decided to make directions for the courts to be configured in line with my instructions. His order is provided as **Appendix C**.

²<https://www.youtube.com/watch?v=Ij2Q-2sq5iI&list=PLPs6vztSaE0gtpFB62vq6tGT6flcRdRof&index=4> (this was linked from the Judiciary.gov.uk website)

- **Example 4: Requiring in person hearings instead of using Zoom.** In a hearing, it was asserted that claimants should not do hearings from abroad but wait until they returned. This was said to be a practice from “*War Pensions*” cases (where many applicants to the tribunal lived abroad and have significant disabilities). In effect, this group would have had to wait an extended time of their pensions due to the Tribunals’ refusal to do remote hearings. The relevant part of the transcript, which former Tribunal ‘President’ McKenna spent a year improperly attempting to suppress, is provided in **Appendix D**.
- **Example 5: Inexpert Judicial ‘Data Protection’.** This example is particularly worrying. The Judge responsible for tribunals data protection (i.e. the “*judicial data protection network lead*”³ responsible for “*supervis[ing] the data processing activities of Tribunals*”) expressed the belief that “*you do not build your systems around the rogue, do you?*”. Of course, computer security is fundamentally about protecting from rogue actors. The Tribunal that this Judge was President of also failed to protect the personal information of judicial office holders (e.g. home addresses) – an example of this is provided as **Appendix E**. A further excerpt from the transcript is provided as **Appendix F**.

These problems have now had some press attention since I began engaging with IT journalists, who have been equally bemused and concerned.⁴ This work has also led to an academic article, in respect of the limited technological understanding of the information rights jurisdiction (Kirkham, 2018). Nevertheless, the judiciary has not taken much in the way of proactive steps to solve these types of problems.

2 The IT ‘Experts’

The underlying issue is really one of false expertise, where Tribunals have been listening to people who say they have the relevant knowledge to determine these matters. All of those examples involved one or more self-proclaimed ‘experts’, including:

- A “*Head of Digital and IT Architecture*” at the Information Commissioner’s Office. He was said to be an expert but admitted he did “*not hold a degree in computing or IT*” but had various certificates, all assessed by “*multiple choice*” quizzes after a few days of attendance. He claimed to be qualified in SQL but said it was beyond his expertise to query an SQL database. This official gave evidence regarding Example 2 but was possibly involved in Example 1. An excerpt of the transcript is provided as **Appendix G**.

³ https://www.judiciary.uk/wp-content/uploads/2019/01/Supplementary-SPT-report-Dec-2018_final.pdf

⁴ For example, Gareth Corfield’s article: https://www.theregister.com/2020/04/08/ico_tribunal_cases_halted_bundle_bewilderment/ The original version resulted in an improper legal threat sent on behalf of the then Chamber President, Alison McKenna, to the media outlet in question.

4

- An (unknown) technical advisor (or group of advisors) working for the Tribunal Service (HMCTS). This advisor (or group of advisors) was implicated in Example 3, but possibly Examples 4 and 5.
- A senior Judge (former Tribunal President) responsible for Information Rights appeals in the UK and being the UK's "*judicial data protection network lead*" for Tribunals. This Judge was involved in **all** the numbered Examples, at least in part.
- Lawyers in the Information Commissioner's Office and certain Barrister's chambers claiming to be competent to argue 'information law' cases, whilst also saying that matters of computer science are not within their expertise. These lawyers were involved in Examples 1 and 2.

The total involves dozens of asserted experts. Of course, what they have supported does not belong in the real world. But the involvement of a wide range of so-called 'IT experts' illustrates the troubling issue – namely the issue of false expertise.

3 Resolving the Ethical Problem – Some Initial Steps

The examples I have given have all undermined the fair administration of justice, especially when they are repeated outside of my cases. It means, for example, that cases are wrongly decided or that people (e.g. with disabilities) are unable to access proceedings. The most extreme example in the UK was the Post-Office case, where IT experts presented a similar problem (concerning a specific computer system) with the end result of many dozens of people being falsely convicted of criminal offences (Marshall et al., 2020; Wallis, 2021). The difference is that my examples are systematic and took place within a specialist branch of the legal system, which is (supposedly) designed to be able to deal with matters of information technology or computer science.

The underlying issue is this: there is no mechanism to sort the wheat from the chaff. In the eyes of the legal system, it appears that someone can become an 'IT expert' after having passed some multiple-choice quizzes, attended a few workshops, and having spent time employed in the area. Yet, plainly this is insufficient. Nor does it suggest that the principles of relevant ethical codes are always being followed – for example, the ACM Code of Ethics expects that practitioners "*Maintain high standards of professional competence, conduct, and ethical practice*" (at 2.2.).⁵ This is far removed from what is happening on the ground in Tribunals.

What can the academic computer science community do about this? The issue is a lacuna: there are no clear standards that protect from self-proclaimed 'IT experts'. Nor is there a recognition of the particularly high standard of expertise and ability needed in our field, which is relatively a very fast-moving one, where even genuine experts cannot be entirely sure.

⁵ There are other ethical principles at play, such as avoiding harm (in the ACM Code at 1.2) and not discriminating (at 1.4). However, this paper is of a limited length, so this wider analysis is for another day.

This means that to begin with, there is a need to be clear about the limitations of ‘IT’ qualifications – at the least, national organizations such as the BCS should be explicit that their short courses and multiple-choice quiz assessed qualifications are not sufficient to make someone an expert (if only to prevent Tribunals misconstruing them as being such). Ideally, the starting point should be a PhD in a relevant field, with a track record of publications in well-recognized venues. In time, we should be looking at developing bespoke qualifications for testing the ability of experts who do not have PhDs.

There are other opportunities to improve the system beyond regulating ‘IT experts’. For instance, there is presently little judicial training or guidance available on matters of computer science (asides from some references to online courts and the implications for vulnerable applicants in the updated Equal Treatment Bench Book⁶): this is different to other fields, where the Royal Society has provided guidance, e.g. (Daeid et al., 2020). Providing such guidance could be a great step forward. Academic computer scientists should also seek to set policy within the justice system, when such opportunities arise. Finally, one can do what I have done, which is to approach specialist journalists and ensure press coverage of these issues.

4 Conclusion

This paper has outlined and evidenced a problem of serious concern, namely how self-appointed ‘IT experts’ have been able to present certain problematic notions, and have them accepted within the justice system. This is a state of affairs that needs to be addressed. In highlighting this issue and providing direct evidence, it is hoped that academics are in a better position to identify and address this problem as it arises.

Appendix A – Excerpt from Journalists original article

The following text appeared in the original version of the article on the Register, before the publication received an (improper) legal threat from the Judicial Press Office on behalf of former Tribunal President McKenna⁷:

“It is quite unusual for legal cases to be paused just because one side can’t put a stack of PDFs into order. The underlying cause here appears to be, if the ICO’s version of events is to be believed, the First-Tier Tribunal’s apparent refusal to work from electronic bundles combined with the ICO’s inability to organise PDF files for judges’ convenience.

Moreover, the speed with which the tribunal imposed the delay on ICO cases at the ICO’s request, seemingly without consulting with the other parties first, may cause

⁶ <https://www.judiciary.uk/publications/december-interim-revision-of-the-equal-treatment-bench-book-issued/>

⁷ The full original article can be found here: <https://perma.cc/VE89-CTFJ>

some to wonder whether the government department's favoured status with the tribunal extends to judgments as well as delays."

Appendix B – Observations of Upper Tribunal Judge Edward Jacobs on Remote Hearings

The below excerpt is from a full decision of the Upper Tribunal, who allowed my appeal against my case being struck out, partly on the alleged basis that a remote hearing between the UK and Australia was impossible to arrange.

“Ground 1 – impossible to list for hearing

12. The Chamber President’s first ground related to the difficulties in making practical arrangements for the hearing of Dr Kirkham’s appeal. In summary, this is what the President said. Dr Kirkham had asked for expedition, but made frequent interlocutory applications that delayed listing. He also took up a post at an Australian University. Coming to the practicalities, Dr Kirkham had proposed a video hearing split over at least four mornings. He did not address the Registrar’s concerns about the increase in the costs of the other parties and of the tribunal itself that this would involve. The Chamber President decided that his proposal would not be fair, because:

- it would increase the other parties’ costs unreasonably;
- it would be unreasonably disruptive of the tribunal’s usual business;
- Dr Kirkham was vague about when he would be back in this country;
- it would not be fair to delay listing for his return.

She concluded: ‘I share the Registrar’s concerns that the Appellant has left the jurisdiction and thus put himself beyond the reach of the usual sanction for unreasonable behaviour.’ There are a number of flaws in this reasoning.

13. I begin with the alleged lack of co-operation. Arguing for a particular listing arrangement is not of itself a failure to co-operate with the tribunal. Dr Kirkham put his proposal. The Chamber President’s grounds suggest that she either had to take it or reject it. That is not correct. The tribunal is responsible for deciding on the listing of an appeal. It will take account of the wishes of the parties and the convenience of their representatives, but the final decision is a matter for the tribunal. The tribunal should have made a decision on listing and proceeded accordingly, without allowing delay for interlocutory issues.

14. Coming to the ability to deal with the case fairly and justly, this includes ‘seeking flexibility in the proceedings’ and ‘ensuring, so far as practicable, that the parties are able to participate fully in the proceedings’ (rule 2(2)(b) and (c)). The Chamber President’s grounds contain no reference to those factors. Dr Kirkham cannot be blamed for taking up employment abroad. There is no suggestion that this was a tactic to delay the proceedings on the appeal. Courts and tribunals conduct hearings by telephone or video link and take account of time differences. Just off the top of my head, the Upper Tribunal has conducted hearings with parties in Switzerland, Malta and Canada. It also held a short permission hearing by telephone with Dr Kirkham in Australia just before Christmas 2019. So the technology is not a problem and Dr

7

Kirkham is comfortable with it. The important feature of this case is the length of the hearing suggested by Dr Kirkham. As I have said, this is a matter for the tribunal to decide, not for him to dictate. That aside, the video time could be reduced by a combination of written arguments and pre-reading, leaving a shorter video hearing for clarification and questions. It may also be possible that some adjustment to the normal sitting hours could reduce the number of days involved. Courts and tribunals are no longer tied to the normal business hours for hearings, so a degree of compromise is possible. On that basis, it ought to be possible to list this case in a way that is, in the Chamber President's own words, 'consistent with the achievement of the overriding objective.'

*(The full case reference is Kirkham v Information Commissioner and UK Research and Innovation [2020] UKUT 93 (AAC). After that decision, the Tribunal doubled down, and proposed to ban **all** information requesters from abroad in a case that involved Julian Assange: <https://www.computerweekly.com/news/252495653/Victory-for-free-speech-and-openness-after-tribunal-confirms-no-territorial-restrictions-to-FOIA>)*

Appendix C – Observations of Upper Tribunal Judge Thomas Church



THE UPPER TRIBUNAL (ADMINISTRATIVE APPEALS CHAMBER)

UPPER TRIBUNAL CASE NO: GIA/2320/2019

TRIBUNAL PROCEDURE (UPPER TRIBUNAL) RULES 2008

Appellant:	Dr Reuben Kirkham
Respondent:	The Information Commissioner
Tribunal:	First-tier Tribunal (General Regulatory Chamber) (Information Rights)
First-tier Tribunal no:	EA/2018/0036
Date of decision:	02 October 2019

OBSERVATIONS

1. Dr Kirkham wishes to exercise his right to renew his application for permission to appeal to the Upper Tribunal at an oral hearing. He lives in Australia. Dr Kirkham has requested a video link with screen sharing.
2. On 05 December 2019 I made Directions asking the Appellant to explain why he considered that a telephone hearing would be inadequate for the purposes of explaining his grounds of appeal, and that the use of "live" visual aids in the form of a powerpoint presentation, and a video link with screen sharing were necessary.
3. On 18 February 2020 the Appellant responded to my Directions by email. I am satisfied by the explanation given by the Appellant and consider it to be proportionate and in the interests of justice to direct a hearing by video link with a facility for screen sharing.

CASE MANAGEMENT DIRECTIONS

In exercise of the powers conferred on me by rule 5 of the Tribunal Procedure (Upper Tribunal) Rules 2008, I direct that:

1. this application for permission to appeal to the Upper Tribunal is listed for an oral hearing with the judge in London and the Appellant attending by live video link from Melbourne, Australia with a facility for screen sharing, being a laptop with an unrestricted internet connection that can access a broadcast by clicking on a link accessible by the judge and the Respondent or her representative (should they opt to participate in the hearing);
2. the Appellant shall, within 14 days of the issue of these Directions notify HMCTS of any dates to avoid for the listing of his application;



GIA/2320/2019

3. the listing clerk shall then list the matter for hearing in accordance with Direction 1 above for the first available date taking into account any limitations on the Appellant's availability notified in accordance with Direction 2 above;
4. in setting the time for the hearing the listing clerk shall have regard to the time difference between London and Melbourne;
5. the maximum time estimate for the hearing of the application is 1 hour, and with this in mind the Appellant shall restrict his powerpoint presentation to a maximum of 15 slides;
6. the Appellant should attend (by video link) and have his set of the Upper Tribunal papers to hand in case he or the judge wishes to refer to them;
7. the Respondent is to be notified of the date, time and place of the oral hearing but need not attend nor be represented at the hearing, although she may if she wishes. Whether she attends or not she may file a written submission but, if it is to be considered, it must be received by the Upper Tribunal and the Appellant at least five working days before the hearing. If permission is given the Respondent will have a further opportunity to put her case later.

(Signed on the original)

Thomas Church
Judge of the Upper Tribunal

Dated 05 March 2020

Appendix D – Former President McKenna's explanation of how she required war pensions claimants to travel for hearings

JUDGE MCKENNA: ... If so, the starting point would be that we would have the hearing at a time when you were in the country. It is very common for us to do that in jurisdictions where, for example, I used to be President of the War Pensions and Armed Forces Compensation Chamber and appellants were often abroad for long periods of time and we waited until they came back."

...

JUDGE MCKENNA: I have just been over there [to Australia]. Yes, I do not know if we have the technology. My starting point would be that we would wait until you were able to attend in person. If the technology becomes available to me, I could reconsider that, but there is a full project that is rolling out over several years, so it is not ready at the moment.

DR KIRKHAM: As I say, I think... If the court has a mind, there are two different types of links. You can do it by ISDN now, but –

JUDGE MCKENNA: No, we would have to do it by own system. We would not do it on a system that was –

10

DR KIRKHAM: No, I am talking about the two systems, if I can be allowed to say? There are two protocols, ISDN is basically the –

JUDGE MCKENNA: You would not know about the pilot protocol; it has not been published. It is ongoing testing.

DR KIRKHAM: I am aware of the two communication protocols you can use separately. I am aware that most of the courts are using ISDN and –

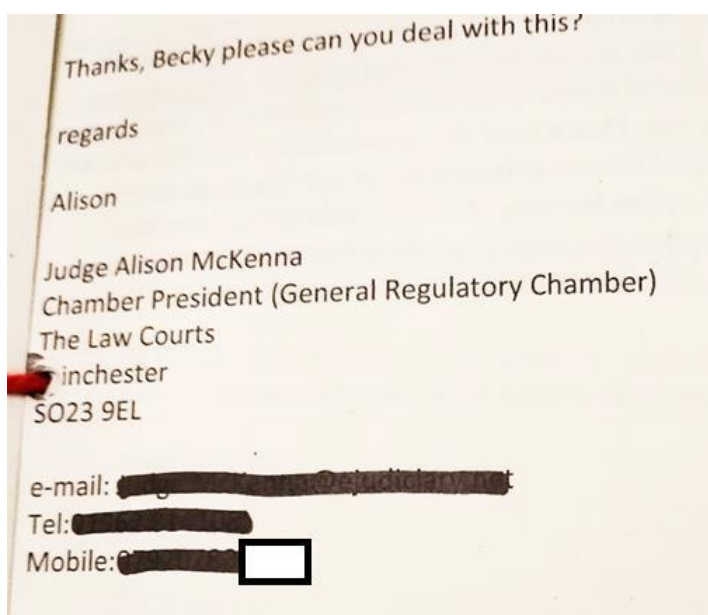
JUDGE MCKENNA: Dr Kirkham, it is your case, if you leave the country, we will try and accommodate you if there is any need for a further hearing in any case, but you are not able to dictate the terms of that today.

DR KIRKHAM: I am not trying to, but I am trying to explain that differences –

JUDGE MCKENNA: I understand you are leaving the country, but there are other people who are not and who want their cases heard, and we have to devote appropriate resources to you when you started a case, but leave the country before it is finished. I cannot turn everything else on its head in order to accommodate you.

Appendix E – Illustration of how Judge McKenna’s Tribunal attempted to ‘redact’ personal information

As can be seen, the redaction using a marker pen fails when the document is held up to the light.



Appendix F – Relevant Excerpts from the Transcript where Judge McKenna explains her approach towards data protection.

This is a small excerpt of the exchanges between the then Tribunal President and myself. The context is whether or not it was appropriate for the Commissioner to rely on asserted security practices as a reason to refuse to do appropriate (i.e. automated) searches in response to a Freedom of Information Act (2000) request. The Commissioner's claim was it was against their existing policy, so should not be done.

JUDGE MCKENNA: Are you suggesting that as part of considering what is a reasonable estimate, we should consider what a rogue employee might do in breach of their terms and conditions of employment?

DR KIRKHAM: Yes, because the concern is security. ... You don't design it on the basis that all your employees can be trusted. It's designed on the basis that you accidentally employ one bad apple, right?

JUDGE MCKENNA: Mr Smithes, is your evidence that a member of staff would be in breach of their terms and conditions of employment for running that sort of system at home?

A. Yes.

JUDGE MCKENNA: Okay. I think that is all we need to know.

...

DR KIRKHAM: The automation doesn't harm data security in any way.

JUDGE MCKENNA: Yes. I understand that but in terms of the time, you are consistently saying, well if you do not trust your staff, then you know, whatever, and I am saying it is reasonable for an employer to have rules in place for their staff and you are saying that the way they should assess the situation is assuming that the employees are not going to comply with the rules -

DR KIRKHAM: And you will see that 99% of them will, but 1% probably won't.

JUDGE MCKENNA: But you do not build your system around the rogue, do you?

Appendix G – Discussion of Level of IT Expertise of the Information Commissioners Office

Most of what follows is a cross examination of the person responsible for information technology in the Information Commissioner's Office.

Q. Obviously, I have downloaded your disaster recovery plan, so have you and your team practised restoring your system in the event of a disaster?

A. In the case of this system, I- it hasn't been done during my time at ICO.

Q. So, these backups you're taking on tapes, how do you know they actually contain the data? How do you know if you're not taking these backups and checking them to see if you can restore the system?

A. There is notification within the backup agent to assure that, but we have not yet found a fault in disaster recovery during my time at the ICO.

12

Q. When was it last done?

A. I do not know. I didn't know I'd be questioned on it, so I haven't checked.

Q. But it's not within the- roughly, is it done every five years? 10 years? Well, when do you plan to do the next one? That might give that away, you know.

A. We're in the process of changing the way we backup, so we are looking to do it just after April, when the next financial year rolls around.

...

Q. So, one person, you trust one person at your [inaudible] provider and if that one person isn't trustworthy, the ICO is screwed? In terms of they could have a massive data loss. One person?

A. Yes.

JUDGE MCKENNA: It depends what the person did, does it not?

DR KIRKHAM: Yes, but one person could walk out the door. There's one person in a company could walk out of the door with all the ICO's data.

A. I would doubt it with that system.

Q. But that could be the entire case management system, for example?

A. Yes.

...

A. I'm not an Excel expert. I have not had great deal of success with Excel, so I wouldn't say that it was simple.

...

A. I don't use Excel on a day-to-day basis. It is not my tool of choice.

Q. Okay.

A. It is not something that I would feel comfortable doing, but that's not to say that it isn't possible.

...

DR KIRKHAM: So, let's say it wasn't Excel, but your choice of programming language, right? Whatever you're expert on. I assume joining one table together with a primary key and comparing it to another table with a primary key is not a difficult exercise within the scheme of things?

A. I mean, no.

Q. You agree? It's smoother and all manner of IT staff would understand and appreciate it?

A. A member of IT staff with database or business analysis experience, yes. But I don't know whether any member of IT staff have done it.

Q. Okay, but there'd be a member, it's highly likely that you'd employ someone who'd be able to do this?

A. With capability, yes. That specific role, no.

Q. Okay. In terms of training someone to do basic operations on Excel, this wouldn't be an expensive exercise? This might be a couple of days' training. Would that be a fair ball park?

A. Yes.

References

- Daeid, N. N., Biedermann, A., Champod, C., Hutton, J., Jackson, G., Neocleous, T., Spiegelhalter, D., Willis, S., Kitchin, D., & Wilson, A. (2020). *The use of Statistics in legal proceedings: A primer for the courts*. The Royal Society.
- Ellis, R. (2013). *Unjust by design: Canada's administrative justice system*. UBC Press.
- Kirkham, R. (2018). How long is a piece of string? The appropriateness of search time as a measure of 'burden' in Access to Information regimes. *Government Information Quarterly*, 35(4), 657–668.
- Marshall, P., Christie, J., Ladkin, B., Littlewood, B., Mason, S., Newby, M., Rogers, J., Thimbleby, H., & Thomas, M. (2020). Recommendations for the probity of computer evidence. *Digital Evidence and Electronic Signature Law Review*, 18.
- Wallis, N. (2021). *The Great Post Office Scandal: The Fight to Expose A Multimillion Pound Scandal Which Put Innocent People in Jail*. Bath Publishing Limited.

From the Page to Practice: Support for Computing Professionals Using a Code of Ethics

Don Gotterbarn¹[0000-0002-7267-4616], Michael S. Kirkpatrick²[0000-0002-7200-4102], and Marty J. Wolf³[0000-0003-0617-942X]

¹ East Tennessee State University, Johnson City, Tennessee, USA

² James Madison University, Harrisonburg, Virginia, USA

³ Bemidji State University, Bemidji, Minnesota, USA
don@gotterbarn.com

Abstract. We describe the Proactive CARE framework for integrating ethical decision making into the daily decision-making process and computing workflow of computing professionals. It is a framework that can help an organization move to more deeply embedding ethical practices into computing. Proactive CARE is based on the ACM Code of Ethics and Professional Conduct (the Code). We identify features of the Code that make it well-suited for a framework such as Proactive CARE. We layout the rationale for a proactive approach and demonstrate the use of Proactive CARE in a case study. We show how Proactive CARE can be used in an educational setting, and we suggest ways to incorporate Proactive CARE into professional settings.

Keywords: ethical decisions, codes of ethics, professional computing, professional practice

1 Introduction

For many people, the “right thing to do” seems obvious. It is based on past experiences, things learned from parents, religious texts, and the constant social cues around us. However, issues in computing ethics are often new and require thought beyond our familiar simple ethical responses. Often people believe “Evil is done by evil people; I am not an evil person and therefore I cannot do evil.” This is a fatal premise and often goes unacknowledged by those engaged in ethical decision making. When professionals take this view, even unknowingly, they approach their work underestimating the real ethical challenges of their work and may simply complete a system without even suspecting that there may be ethical problems with it. They will not have even considered taking the time to do the ethical analysis and consider the impact of the system on a broad range of stakeholders.

The harm that can be caused by software made by (presumably) good people has been well documented. Google’s search auto-completion has made suggestions based on racial tropes, thereby oppressing Black women and other marginalized groups (Noble, 2018). Facial recognition algorithms have been shown to perform

significantly worse on dark-skinned faces, facilitating discriminatory applications (Buolamwini & Gebru, 2018). Ostensibly neutral classification algorithms have led to discriminatory effects in a variety of fields, including insurance and law enforcement (O’Neill, 2016).

Psychological studies have found that the fatal premise is common and found that our immediate ethical reactions, which involve the intuitive processing of information, frequently precede ethical decision-making that is slower, more conscious, and more logical (Tenbrunsel, 2004; Basserman, 2012). Unfortunately, the fatal premise is reinforced when an organization primarily focuses its staff development on technical skills acquisition and ignores practicing how to consider the impacts of technical decisions on society.

Even a code of ethics may help reinforce the fatal premise when it is mere window dressing. On the other hand, when a code of ethics is a document that reflects a deeply ingrained set of attitudes and practices that guide the day-to-day decisions of individuals and the organization, the fatal premise is regularly challenged. While few, if any, organizations have achieved this ideal, its pursuit is worthwhile in that it aligns the success of the organization with the good of society. Making ethical decisions in practice requires interpreting, choosing, and assessing our activities as professionals (Miller, et al., 2017).

In this paper, we describe the Proactive CARE framework for integrating ethical decision making into the daily decision-making process and computing workflow. In Section 2, we identify features of the ACM Code of Ethics and Professional Conduct (the Code) that help start the process of change. Section 3 lays out the rationale for such a move. Section 4 introduces Proactive CARE. Section 5 demonstrates a Proactive CARE case study. Section 6 demonstrates how Proactive CARE can be used in an educational setting. In Section 7 we suggest ways to incorporate Proactive CARE into professional settings.

2 The ACM Code of Ethics and Professional Conduct

The ACM Code of Ethics and Professional Conduct starts with the premise that computing professionals, which includes anyone who “uses computing technology in an impactful way,” have a responsibility to act in support of the public good. It is structured as 25 principles that come with guidance indicating what a professional should aspire to. These principles help guide professionals to contribute to society in ways that both minimize unintentional ethical mistakes and maximize positive impacts. The Code supports professionals in interpreting, choosing, and assessing their activities. As an ethical framework, it is a guide to proactive action that inspires professionals to do good by identifying and then addressing opportunities to engage in more ethical computing practices (Brinkman et al., 2016; Gotterbarn et al., 2018).

Rather than emphasizing a retributive framework focused on violations, the Code calls on computing professionals to proactively identify opportunities for doing good.

Each principle starts with “a computing professional should ...”. This perspective, and the Code itself, implies that it is essential that computing professionals develop

the ability to recognize and navigate ethical challenges both before and as they arise. Professionals engaged in this practice know that the Code must be treated as a complex holistic set of principles. Doing so helps surface constraints that ought to be placed on ICT systems. More importantly, the Code calls on computing professionals to proactively identify opportunities for doing good.

The Code's first principle, "Contribute to society and human well-being acknowledging that all people are stakeholders in computing," goes beyond simple advocacy of seeking opportunities to do good in projects one is working on. The guidance says that computing professionals "are encouraged to actively contribute to society by engaging in pro bono or volunteer work." Pro bono work is more than merely charitable volunteerism—it means using one's special professional skills for the public good in a broader context.

The reference to pro bono work is simply a reminder to good people. The Code has an underlying assumption that computing professionals want to do the right thing, but the complexity and pressure of their work can make it difficult for them to recognize how to achieve this goal. This challenge is acknowledged in the Code and partially addressed by each principle's guidance that indicates what a professional might aspire to. Determining the best approach in a particular situation can be challenging when principles conflict, and making the best choice requires the ability to weigh multiple perspectives and values with nuance. Principle 2.2 specifically recognizes the need to develop ethical skills and makes doing so obligatory for computing professionals.

Another strength of the Code is the way it was developed. It was developed in a public way that identified principles that represent the conscience of the global computing profession rather than the potentially narrow aims of a single organization. That way, even though people may bring different experiences of "ethics" to their work as computing professionals, being computing professionals is one thing they have in common. The ACM Code of Ethics and Professional Conduct captures the shared values of the profession, but there is more to do.

3 A Case for a Proactive Approach

The Code advocates for careful ethical analysis using its principles. It suggests parameters for such deliberation, and it identifies the public good as paramount in making decisions where there is tension among multiple principles. The choice that upholds the public good most should be the one prioritized. When faced with multiple choices, sound decisions are made by careful evaluation of the relevant principles. Putting these values into practice requires tools beyond those present in the Code.

Support is also needed to prevent a practice whereby ethical decision making guided by the Code frames out relevant ethical considerations. The principles in the Code are meant to be interpreted in a holistic and cohesive manner, rather than as distinct policy guides. Focusing on a single principle can lead to an analysis that misses important considerations. One might be tempted to focus on Principle 1.6, Respect Privacy when developing a Contact Tracing app, but other principles, such as Principle 1.2, Avoid Harm, apply in this situation as well.

Further, when a code of ethics is seen merely as a compliance tool, ancillary materials intended to facilitate the use and understanding of the code may not move an organization toward the ideal mentioned above. For example, the case studies included in a booklet (The Code, 2018) published with the Code outlined a methodology to analyze case studies. This methodology, called Consider, Analyze, Review, Evaluate (CARE), invites adopters “to apply the Code as a framework for analyzing ethical dilemmas.” CARE “provides an outline for judging whether possible actions ... [are] consistent with both the letter and the spirit of the Code.” This initial methodology illustrated how one might apply the Code in a retrospective case analysis to develop a deeper understanding of how the principles of the Code interacted with one another. The questions provided by CARE are effective for guiding a user’s exploration of ethical issues and the Code and for evaluating whether certain actions have violated the Code. This approach focuses on one function of a Code, namely using it to apply and understand past experiences of failures to follow the Code. Those who are restricted to this narrow focus are limiting their use of the Code and missing out on opportunities to use it to guide positive decision-making and to incorporate its values into their daily practice.

After-the-fact case studies are educational, but when those analyses stop at fingerpointing, an opportunity is missed. There is value in taking the next step and using the analysis in conjunction with the Code to identify opportunities for changing conversations around project development to prevent similar lapses in the future and to identify opportunities for doing good. Once those opportunities are identified, integrating the Code, which was designed to guide and inspire, into the practice of computing for in-the-moment analysis, can lead to both products and processes that better integrate ethical values. This aspirational approach moves the computing professional away from using the Code to impose a narrow moral vision on others toward developing an understanding of the broad impact of the project.

Even when computing professionals are skilled in case analysis, there is more to do to achieve the goal of ethical practice. Ethical practice requires the ability to identify features of both software and society that may have meaningful ethical impacts. It requires the ability to have potentially challenging conversations with colleagues about issues that weave together both technical and ethical issues. It requires the ability to know the limits of one’s professional abilities along multiple axes. It requires the ability to seek outside expertise: knowing when, how, and whom to ask. Integrating ethical decisions into professional work requires developing skills to address challenges as they arise, which takes practice beyond just after-the-fact analysis.

Without consistent practice in considering the impact of technology on people prior to and during the development cycle, significant issues are more likely to be missed. Narrowing our focus can sometimes frame ethical questions out of consideration. Ethical decision-making processes take metacognitive attention to consider aspects beyond the way situations are framed for us. Sometimes computing systems fundamentally change the nature of society in ways that we did not consider. It takes proactive development processes to Consider relevant ethical elements, Analyze the impact of those elements, Review pertinent responsibilities, authority, and alternate

approaches, and Evaluate those alternatives. A structured framework that can help us to recognize the impacts of a system on a broad range of stakeholders is required to make a sound ethical decision from amongst all possible choices.

4 Proactive CARE

Proactive CARE for Computing Professionals (Gotterbarn et al., 2021) supports incorporating ethical analysis into the software development process using the Code’s principles. The framework is intentional about identifying opportunities for more ethical computing practices. Proactive CARE is general enough to be integrated into most typical software development processes. It augments the Code by giving computing professionals a basis for conversation about the ethical and social impact of their work. Its questions guide computing professionals through a process to consider and evaluate the ethical dimensions and opportunities present in a project at various points and prior to making decisions. It helps recognize and address the ethical components of daily technical practice and engage with colleagues about alternative designs and paths forward. Proactive CARE helps people identify ethical obligations and opportunities. It moves the guidance of the Code away from moral imperialism toward an understanding of the broad impact of projects, the values that are central to what “you should do,” and what you can do to enhance the value the project brings.

Proactive CARE is an iterative and adaptive process that invites its practitioners to regularly and continually engage with elements from the CARE case analysis tool. We summarize the approach here, explaining the intent of the four elements. The full framework, provided in the Appendix for reference, consists of a set of questions derived from the principles of the Code.

Consider: Code that is worth developing is going to affect other people. Consider whose behavior, circumstances, experiences, and work process will be affected. These considerations lead to identifying stakeholders whose situations ought to be addressed. Understanding a broad range of stakeholders allows for the consideration of alternative approaches to constructing the system.

Analyze: Analysis should involve the Code. How do its principles come into play with the identified considerations? Analyze the technical facts of the system and the consequences of each of the alternative approaches for each stakeholder. Analysis should surface conflicting values: among stakeholders, between technical concerns and ethical concerns, and among principles in the Code.

Review: Analysis leads to the identification of trade-offs and conflicting values and to opportunities to change paths to make a product that does a better job of advancing the principles in the Code. Review those alternatives, values, and opportunities. Identify creative alternatives that bridge those gaps. Loopback to “Consider” and “Analyze” before proceeding.

Evaluate: Review leads to a clearer understanding of the alternatives and a deeper understanding of the Code. Evaluation identifies the best approaches and articulates their strengths and trade-offs. There is likely tension surrounding the potential actions. There may not be an optimal choice, merely ones with different trade-offs. Evaluate the knock-on effects of each of the alternatives, knowing full well that you may not have them right. Resolving these tensions requires some stamina. Furthermore, evaluation requires continued monitoring of decisions to later determine if it was a good one and learn from that experience.

Proactive CARE users identify multiple alternatives to the situation they are facing. As part of the process, they analyze the ethical aspects of alternatives under consideration, determining likely outcomes and impacts. Proactive CARE also incorporates rights, relationships, and values that might be overlooked in systems and by teams that focus primarily on quantified data. This guided decision support framework leads to informed decisions that include evaluation of the ethical impact of the alternative solutions under consideration. When those decisions are supported by the principles in the Code, the resulting system is undergirded by an international set of professional values.

5 Proactive CARE: A Case Study

As Proactive CARE is intended to be used in active, dynamic software development situations, case studies that demonstrate its application of the Code do so at key decision-making moments. Here we present such a case study that is based on a significant real-world data breach¹. Throughout the narrative, we demonstrate how Proactive CARE could have been used at key moments and contrast this approach with the response taken by Equifax.

Equifax, one of the three largest credit reporting agencies in the USA, collects and aggregates credit and demographic information on over 800 million individual consumers and more than 88 million businesses worldwide. To support its online dispute portal web application, Equifax used Apache Struts Toolkit, an older application built on a popular programming framework for building Java web applications (Foster, 2018).

Imagine that you are a software engineer who specializes in cybersecurity at Equifax. In March 2017, you have reached your fifth anniversary with the company, giving you sufficient experience to understand both general security concerns and threats specific to Equifax. Due to the sensitivity of the data collected by Equifax, the company is obligated to ensure these records are accurate and secure. Equifax's code of ethics has a section on "Protecting our information and assets" that states you are "responsible for protecting them and using them with care." Recent high-profile data

¹ This analysis is based on a case study and data provided by M. de Roche and D. Gotterbarn. The original study was developed by some members of the IFIP and ACM ethics committees. June 2021. Special acknowledgements to D. Gotterbarn, M. de Roche, M. Havey, M. Kirkpatrick, D. Kreps, and K. Miller.

breaches at Experian and Target have made company leadership anxious about the possibility of a similar data breach.

Consider: The stakeholders affected by a potential breach include the consumers and businesses whose data has been collected. The breaches at other companies could act as an opportunity to prepare, and Equifax could use these events to justify an updated risk assessment.

Analyze: Credit reports are used for many purposes, so a breach could affect these stakeholders' financial management, employment, qualification for government assistance, and many other important areas of their lives. Principle 2.1 advises computing professionals to be "cognizant of any serious negative consequences," which would include the impacts of a data breach. Stakeholders are right to expect Equifax to take steps to prevent breaches.

Review: As a software engineer, you could be assigned to a variety of tasks. One possibility would be to perform a security audit of existing applications; alternatively, you could build new applications that provide new services to users. The latter could improve the customer experience or improve marketing. At the same time, security audits could identify potential threats to a large number of stakeholders; however, it is not always clear whether such audits will be successful, and funding for such tasks is limited.

Evaluate: Given your background, it would be prudent to assign you to a security audit while assigning more junior developers to build new applications. The concern about budgets is legitimate and it is impossible to predict when security breaches will happen. However, the recent high-profile breaches, particularly at Equifax's competitor Experian, have provided a warning that should be heeded.

6-8 March 2017

On Monday, 6 March, Apache discovered a critical vulnerability in their software. On Tuesday, one day after the vulnerability was discovered, Apache, acting responsibly, issued a patch to address the flaw and warned its customers of the risk and the need to immediately implement the patch. On Wednesday the US Department of Homeland Security's Computer Emergency Response Team (CERT) redistributed the patch, emphasizing the importance of its immediate installation (Franceschi-Bicchierai, 2017). As a cybersecurity professional, you are subscribed to a number of mailing lists and other news sources where this vulnerability and patch are discussed. You read about the patch on Tuesday and receive an alert about CERT's advisory on Wednesday.

Consider: Since this critical vulnerability potentially affected all Apache Struts applications, Equifax's complaints portal was put at risk. Your experience makes you well positioned to recognize the scope of the problem and how to

mitigate it. Assigning you to install the patch will cause delays in your other work but leaving the application vulnerable places consumers at risk.

Analyze: Principle 2.9 clearly articulates that breaches cause harm and computing professionals should “take action to secure resources,” so patching this vulnerability should be a priority. Equifax’s security department agrees and has a general policy that requires critical vulnerabilities to be patched within 48 hours. Failing to address this issue in a timely manner violates public trust in credit agencies and Equifax’s own commitment to not harm customers.

Review: Software modifications, including applying patches, introduces the risk of accidentally breaking something and requires testing that can take time. While the patch is being evaluated, you could install a web access filter to detect and stop outside attacks on known vulnerabilities. However, this installation could distract you from getting the patch installed.

Evaluate: While security breaches are unpredictable and there is no guarantee that this vulnerability would be exploited, both Apache and CERT have emphasized its importance. The potential for significant impact on stakeholders should make this patch your top priority and the best approach would be to submit a change request to apply the software patch throughout the system as soon as possible. This is also consistent with the 48-hour patch installation policy. You resolve to monitor the situation.

17 March 2017

Ten days after the warning, Equifax installed the patch and ran a scan to determine if the patch was installed correctly. However, there were some installations where the patch was not applied correctly and some servers using the Apache Struts Toolkit were not scanned. Consequently, several servers were left vulnerable to attack. Because you are still concerned about the severity of the vulnerability, you have followed the progress of the patch and become frustrated with the delays and worried about the completion of the patch.

Consider: Equifax’s delays, mistakes in installation, and incomplete scanning have continued to put stakeholders at risk, even though they are unaware of this fact. After the installation and scan have been completed, you could consider the vulnerability to be fixed or you could advocate for a more complete audit.

Analyze: Principle 2.5 describes the importance of “comprehensive and thorough evaluations of computer systems and their impacts,” which supports the argument for a thorough audit and risk assessment after the patch. However, vocally insisting on such a response could risk alienating colleagues if their professionalism and competence are accidentally questioned. You weigh the impact on your relationship with colleagues versus the potential risk to all of Equifax’s clients.

Review: An initial discussion with your manager could help you to raise your concerns while avoiding interpersonal mistakes. If that approach is unsuccessful or you have concerns about retribution, Equifax has authorized employees to report concerns anonymously using their “Integrity Line” (Equifax, 2017)².

Evaluate: Given your experience and specialization in cybersecurity, your manager and colleagues should respect your judgment when making such a recommendation. Raising your concerns with your manager or the Integrity Line would be appropriate and could potentially uncover other problems that are currently unknown.

29 July 2017

Over four and a half months after the CERT notification about the Apache Struts vulnerability, a security researcher began scanning the company’s public-facing infrastructure. This scan uncovered a site that “looked like a portal made only for employees but was completely exposed to anyone on the internet” (FranceschiBicchierai, 2017). The site required no authentication but displayed several search fields, allowing anyone to “force the site to display the personal data of Equifax’s customers.”

The Equifax security department discovered “suspicious network traffic” associated with its online dispute portal and took the portal offline on 30 July (EPIC, 2019). The ensuing investigation revealed that, from 13 May to 30 July, hackers utilized simple commands to determine the credentials of network accounts at Equifax to access and exfiltrate sensitive personal information. The attackers moved from the web portal to other servers because the systems weren’t adequately segmented from one another, and they found usernames and passwords stored in plain text that then allowed them to access further systems. Because Equifax failed to renew expired security certificates, attackers were able to bypass login protocols undetected and pull data out of the network for several months (Keller & Canfield, 2018). Equifax eventually patched this vulnerability.

After hearing about this security investigation, you try to gather information to review and evaluate what has been done. When you ask about this newly discovered vulnerability you are told that it is being managed by the other team. A few more days go by and there is no public announcement about the breach and the theft of sensitive data. You are told that the company needs to organize several things to protect their interests, including the purchase of an identity theft company whose services they could sell to those whose data had not been protected by Equifax. You are concerned about the delay and consider it unethical to put the company’s requirements before the

² Equifax maintains an “Integrity Line” to report ethics concerns which they encourage employees to use any time they have a concern that can be submitted anonymously. “Equifax has an opportunity to improve every time you ask a question or raise a concern.”

safety of the customers and others whose sensitive data they hold. Not only has Equifax failed to protect its clients' data, but now it is depriving consumers of an opportunity to take immediate precautionary measures to protect themselves from identity theft and fraud. You consider leaking information about the breach to a media contact.

Consider: As before, Equifax's actions (including delaying a public announcement) put consumers and businesses at risk. However, now Equifax's reputation is also at risk with such a disclosure. If the announcement is made before the full scope of the breach is understood and Equifax is prepared to remedy the harm, the company's standing with the public may be further and unnecessarily damaged. In considering leaking the information, you are placing a high value on institutional transparency, though at the cost of confidentiality.

Analyze: Principle 2.9 guides that affected parties should be "notified in a timely and clear manner, providing appropriate guidance and remediation." It is unavoidable that the breach, once it becomes public, will damage Equifax's relationship with the public. In choosing to acquire an identity theft company, rather than contract for such services, Equifax has potentially introduced unnecessary complexity and delay. This choice illustrates a mismatch between the rights and values of the affected parties and those of Equifax, as the latter may view the delay as worthwhile to secure identity theft services at a lower cost. Principle

1.2 indicates that it "may be necessary to 'blow the whistle' to reduce potential harm," but "capricious or misguided reporting of risks can itself be harmful." As a software engineer rather than a business analyst, you may not be in a position to assess all of the relevant information.

Review: Equifax is both morally and legally obligated to disclose the breach and to mitigate the harm caused by it. This disclosure will inevitably hurt Equifax's reputation. At the same time, Equifax has a fiduciary duty to its shareholders, which includes managing operating costs and reputational harm. As an employee and computing professional, you are also expected to honor confidentiality (Principle 1.7). Leaking information about the breach is, inherently, a violation of confidentiality and you should "consider thoughtfully whether such disclosures are consistent with the Code." Your role in the company makes it unlikely that you could influence the acquisition decision, but you could note your objections to your manager or the Integrity Line.

Evaluate: Although there are good financial reasons for cutting costs by acquiring an identity theft company, the affected parties have been at risk for five months, with known data exfiltration for three months. Mitigating this harm should have priority over reducing operating costs, and Equifax should not wait to acquire the identity theft company before disclosing and addressing the breach. At the same time, it would be inadvisable for you to leak the information to a media contact at this point. In doing so, you could potentially be introducing even more delay as the leadership of Equifax would shift focus to the public relations

impact. In addition to harm to the company and those affected by the breach, you could be fired or face legal consequences.

7 September 2017

After the discovery on 29 July, the company did not inform the public of the breach. Weeks were spent hiring cybersecurity experts, informing select groups of the breach, and purchasing an identity protection company so they could sell its services to consumers who had their data stolen. The company did not publicize the breach until 7 September on Twitter, more than a month after they discovered it had happened. Their announcement only indicated unauthorized “access: occurred from mid-May first infiltration through July” not making clear that extensive personal identifying information including names, home addresses, phone numbers, date of birth, social security numbers, driver’s license numbers, and credit card numbers had not only been accessed but also stolen (i.e., exfiltrated from its systems) (Keller and Canfield, 2018).

6 Bootstrapping Proactive CARE

One of the challenges we face is getting computing practitioners to adopt approaches such as Proactive CARE into their development process. Addressing that difficulty head on is beyond the scope of this paper; however, there is a way to begin chipping away at it. Wolf and Greer (2021) have developed a simple wrapper that helps instructors integrate Proactive CARE into just about any computing project. While their approach may be incorporated into a single assignment here or there, it is most impactful when it is incorporated into multiple assignments in multiple courses throughout the computer science curriculum. Students who learn software development using processes that integrate the consideration of social and ethical concerns into the development of code will come to understand “that it’s just the way it’s done,” and, increasingly, they will expect to do so as they move to their professional careers. At the very least they will be experienced in having conversations that surface social and ethical concerns.

In addition to the programming assignment template, Wolf and Greer provide rubrics that can be used to evaluate student work. One rubric is designed to evaluate students who are not experienced in the practice of ethical reflection, and the second rubric is helpful in evaluating student work when they are more experienced. The more frequently that students use Proactive CARE (or any reflective practice), the better they should become at broad considerations of social and ethical concerns surrounding projects.

7 Deploying Proactive CARE

In many ICT organizations the primary focus is on meeting functional requirements within budget and schedule. Computing professionals working in a rapidly changing discipline adopt a divide and conquer methodology to narrowly focus on the complex technical task of meeting specifications and avoiding mistakes along the way. When commercial incentives reward piece work production, even by computing professionals, those incentives frame out responding to critical ethical issues even if they might be recognized. Addressing a broader professional obligation seems like a distraction rather than an essential part of “meeting specifications.”

Appeals to a lack of time, budget, or skill create a culture where discussing professional ethics is not valued and ethical issues are ignored. When organizations have moved a step beyond by designating an internal ethics authority who presents the annual half-day staff ethics training that strongly emphasizes legal compliance, complex ethical judgment is reduced to merely box checking. Psychological studies have shown that the compliance model of professional ethics is viewed by practitioners as binary checklists and does not seem to involve ethical judgment (Basserman, 2012). The presence of the checklist reduces their felt need for careful judgment.

In either environment the commercial emphasis on technical skills and carefully engineered product development are entrenched, and Proactive CARE can be used to move beyond mere compliance box-checking. Typically, code development is monitored and improved by group technical code reviews. Responsible software development identifies and responds to the needs of those impacted by the system. Once those stakeholders are identified the design can be modified so they are positively, not negatively impacted. One technique to promote and assist in this identification is to add questions from Proactive CARE to technical code reviews. These questions address how the system might be modified to improve the stakeholders’ situations. Making these questions part of the technical review improves the ethical culture and provides an opportunity for the developers to identify possible positive contributions.

A common element to various stages of a development life cycle is a task list, project management plan, design checklist, test plan, etc. Individual principles from the Code can be placed at the top of assignments with the suggestion that employees take the principle into account when doing their tasks. The Code asserts that a computing professional “...should reflect upon the wider impacts of their work, consistently supporting the public good.” The Code provides guidance for addressing ethical risks and guidance identifying ethical opportunities to support the public good. The Code makes clear that the concern to “do it right” includes caring for those who are “affected directly or indirectly by their work.”

Simple questions at the end of every task list like “How can we better contribute to society?” can initiate individual reflection. Questions like those in Proactive Care that reflect principles of the Code rather than the principles of faster and cheaper development will lead to improvements in both the process and product of work. Companies that positively promote an ethical culture better attract employees, are

more profitable with fewer recalled products, and have lower training costs (Midwest 2019, Obaidy 2019). Additionally, an ethical culture can help attract top talent and retain employees. People like working in an ethical environment where their contributions can make a positive difference. Proactive CARE supports the view that if you want more kindness in the world (computing's impact to be more positive), you need to put it there. The results of this change in development should be rewarded.

Individual reflections on Proactive CARE questions then serve as a basis of conversation during all stages of development. Integrating Proactive CARE questions into the development process encourages developers to reflect on and converse about them at every stage of the development cycle. These questions help establish a mindset. They are not a checklist of yes or no responses but they require elaboration and creative thinking. When each individual CONSIDERs the questions, the team has a basis for conversation about the ethical and social impact of the project under development.

Establishing an approach like Proactive CARE in an organization redistributes power, ownership, and pride in product quality to all who take part in the practice of Proactive CARE. When Proactive CARE becomes part of technical code review and employees are rewarded for recognizing how to more responsibly practice their profession, people will speak up and become more involved in their work. Risk analysis will extend beyond removing potential problems with the development and include potential improvements to the product so that it will have broader positive impacts.

The entire computing profession benefits when questions related to ethical issues can best be answered by thoughtful proactive consideration of the fundamental ethical principles consistent with the Code that the public good is the paramount consideration.

8 Conclusions

Because Proactive CARE is based on the widely adopted ACM Code of Ethics and Professional Conduct, it gives everyone a common starting point when computing professionals from different countries/companies/educational backgrounds come together to work on a project. With this common starting point there is a better chance of positive ethical action and outcomes.

Using Proactive CARE moves people beyond a common perception that codes of ethics are merely useful for determining whether organizations or individuals ought to bear some responsibility for failed projects or unforeseen harm caused by a "successful" project.

Proactive CARE does not make ethical decision making any easier. It merely increases the chances of that decision being better. It helps bring to the fore the unquantifiable nuances faced by those incorporating the Code into the technical decision-making process. The framework illustrates one way to approach ethical concerns rationally and systematically so that the harmful impacts of the system being developed are reduced.

Practicing Proactive CARE supports Principle 1.1 “Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing.”

References

- Association for Computing Machinery. (2018). ACM Code of Ethics and Professional Conduct. Retrieved from <https://www.acm.org/code-of-ethics>
- Association for Computing Machinery. (2018). *The Code*. Retrieved from <https://www.acm.org/binaries/content/assets/about/acm-code-of-ethics-booklet.pdf>
- Bazerman, M. & Tenbrunsel, A. (2012). *Blind Spots: Why We Fail to Do What's Right and What to Do about It*. Princeton University Press.
- Brinkman, B., Gotterbarn, D., Miller, K., & Wolf, M. (2016). Making a positive impact: updating the ACM code of ethics. *Communications of the ACM* 59(12): 7—13. <https://doi.org/10.1145/3015149>
- Buolamwini, J. & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, in *Proceedings of Machine Learning Research* 81: 77—91. Retrieved from <https://proceedings.mlr.press/v81/buolamwini18a.html>
- Electronic Privacy Information Center (EPIC). (2019). *Equifax data breach*. <https://epic.org/privacy/data-breach/equifax/>
- Equifax. (2017). *Code of Ethics and Business Conduct*. https://www.equifax.com/assets/corp/code_of_ethics.pdf
- Foster, K. (2018). *What are the ethical implications of the Equifax data breach?* Business Ethics Advisors. <https://johnkevinfoster.com/equifax-data-breach/>
- Franceschi-Bicchierai, L. (2017). *Equifax was warned*. https://www.vice.com/en_us/article/ne3bv7/equifax-breach-social-security-numbers-researcher-warning
- Gotterbarn, D., Bruckman, A., Flick C., Miller, K., & Wolf, M. (2018). ACM code of ethics: a guide for positive action. *Communications of the ACM* 61(1): 121—128. <https://doi.org/10.1145/3173016>
- Gotterbarn, D., Kirkpatrick, M., Miller, K., Tihi, J., & Wolf, M. (2021). *Proactive CARE for Computing Professionals*. <https://ethics.acm.org/wp-content/uploads/2021/03/Proactive-CARE-for-Computing-Professionals.pdf>
- Keller, A. & Canfield, K. (2018). CONSOLIDATED CONSUMER CLASS ACTION COMPLAINT, In re: Equifax, Inc. Customer Data Security Breach Litigation. Case 1:17-md-02800-TWT Document 374 Filed 05/14/18. https://www.equifaxbreachsettlement.com/admin/services/connectedapps.cms.extensions/1.0.0.0/ed93e6d9-c6b0-4829-994c-a7687661917f_1033_ConsolidatedConsumer-Class-Action-Complaint-20180514.pdf

- Minnwest Bank. (2019). *Small business - How ethics can help your bottom line*.
<https://www.minnwestbank.com/blog/small-business-how-ethics-can-help-yourbottom-line>
- Miller, K. W., Wolf, M. J., & Grodzinsky, F. S. (2017). This ethical trap is for roboticists, not robots: On the issue of artificial agent ethical decision making. *Science and Engineering Ethics*, 23(2), 389—401. <https://doi.org/10.1007/s11948-016-9785-y>
- Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press. New York.
- O'Neill, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Books.
- Obaidy, H. (July 26, 2019). *How an ethical workplace impacts your bottom line*.
<https://emtrain.com/blog/workplace-culture/ethics-impacts-bottom-line/>
- Rawls, John. (2009). *A Theory of Justice*. Harvard University Press.
- Tenbrunsel, A. & Messick, D. (2004). Ethical fading, the role of self-deception in unethical behavior.” *Social Justice Research*, 17(2), 223—236. <https://doi.org/10.1023/B:SORE.0000027411.35832.53>
- Wolf, M. J. & Greer, C. (2021). *Adding responsible CS to a programming assignment*.
<https://www.bemidjistate.edu/academics/departments/mathematics-computerscience/rcs/teaching-modules/adding-responsible-cs-to-a-programming-assignment/>

Appendix

Proactive CARE for Computing Professionals

ACM Committee on Professional Ethics, January 2021

When faced with complex situations, ethical issues are certain to arise, and it is helpful to use a guided process that leads to informed decisions among alternative solutions to the situation. Proactive CARE (Consider, Analyze, Review, and Evaluate) is a process that uses a set of questions to help computing professionals think through such issues. Proactive CARE uses the ACM Code of Ethics and Professional Conduct as a framework to identify and address opportunities to engage in more ethical computing practices.

CONSIDER who might be affected and how: Whose behavior, circumstances, job, or experiences might be affected positively or negatively by the proposed solution? What values might they bring? Who has the ability to act in the current situation? Whose expertise should be part of the project, especially considering expertise in domains other than computing? What possible alternative solutions address different stakeholder needs and impacts? What are the most important consequences (e.g. shortterm, long-term, local, global, and environmental) for each of the alternative solutions for each of these stakeholders? How will conflicting stakeholder interests be addressed? Are there historical factors or injustices that are not readily apparent and require heightened consideration? Are there details you should clarify before you decide?

ANALYZE the situation's details: For each plausible alternative solution, what stakeholder rights and relationships between stakeholders are likely to be affected? What obligations do you have that are relevant to each of the stakeholders? What stakeholder values are in conflict? What aspects of the Code's guidance are most relevant? (Different principles may be most relevant to different alternatives, both now and in the future.) What technical facts are most relevant to your decision and your system? How might personal, institutional, or legal values influence the choice of possible alternatives? How might the system's design or implementation be improved based on observations so far?

REVIEW other obligations and limitations: What responsibilities, authority, practices, or policies seem to be most important to the alternatives in your analysis? Which alternative seems to be the most compelling? If negative consequences are unavoidable, what mitigations need to be enacted to limit this impact? Are there predictable reactions from others that will need to be addressed? Are there other creative alternatives to the ones you've considered so far? (If so, loop back to

“Consider” and “Analyze” those alternatives before proceeding.) What potential actions might you take to make a positive difference?

EVALUATE the best course of action: Which of the alternatives considered seems to be the fairest and just? Will it aid or hinder violating Code Principles? Are there other Principles in the Code that are relevant to your deliberations about this action? What are the trade-offs and why do they exist? What are the secondary and unintentional effects of the choices you have made? After you have made your decision and after you implement the chosen alternative, what will you monitor in the future to help determine if this was a good decision? How will you (and others) judge the quality of this decision when you look back at some time in the future?

“It Belongs in a Museum!”: A Practical Take on the Question of Artificial Intelligence and Moral Patiency

Lauri Tuovinen^[0000-0002-7916-0255]

University of Oulu, Oulu, Finland

lauri.tuovinen@oulu.fi

Abstract. One of the more philosophical issues in artificial intelligence (AI) ethics is the debate on whether an AI-based entity such as a robot can ever be considered a moral patient. The debate revolves around the question of whether such an entity could conceivably have interests that would warrant the conclusion that it should have rights akin to human or animal rights. This paper explores an alternative point of view, where the appropriate analogy is not a sentient being but a cultural artifact such as a great work of art. Such an artifact, while not strictly speaking a moral patient, is arguably due some level of moral consideration and moral agents may have an obligation to protect it. By considering AI-based entities as cultural artifacts, an argument can be made that 1) such entities do, in some cases, have substantial intrinsic value based on their cultural significance, 2) such entities have certain special qualities that distinguish them as a category from other, closely related technological artifacts, and 3) the above considerations are already applicable to entities that exist today, not just to hypothetical future sentient or conscious machines.

Keywords: artificial intelligence, moral patiency, intrinsic value, cultural heritage, robot rights

1 Introduction

Autonomous systems based on artificial intelligence (AI) is one of the central themes in AI ethics, because the decisions made or actions taken by such systems may have substantial ethical implications. AI-based systems may, for example, evaluate the eligibility of loan or job applicants, in which case the fairness and explainability of the system’s decisions are important concerns. Others, such as autonomous vehicles and surveillance applications, may jeopardize human safety or liberty, so it is crucial to ensure their robustness as well as address the issue of legal and moral responsibility for adverse consequences.

Among the more philosophical issues being debated is the status of AI-based autonomous systems with regard to moral agency and patiency. Simply put, a moral agent is an entity with moral obligations, whereas a moral patient is an entity with moral rights. Traditionally, artificial entities have been excluded from both categories, but as ideas such as artificial life and artificial consciousness have gained traction, so

has the idea that artificial entities may yet be created that deserve to be acknowledged as moral agents and/or patients.

It seems safe to say that no currently existing AI system satisfies any meaningful set of criteria for either moral agency or moral patiency, so it is tempting to declare that these concepts are, at least for the time being, purely philosophical and have little or no relevance to practical AI ethics today. Arguably, however, the situation is not quite that simple. For example, while an AI system that makes decisions with morally relevant consequences may not be a moral agent per se, its decision-making processes may be so complex and opaque that it is not at all clear who should be held accountable for the decisions. A survey on the current status of research in the area of artificial moral agents has been carried out by Cervantes et al. (2020).

Discussions of the possible moral patiency of AI-based entities are generally based on the premise that for such an entity to be considered a moral patient, it must have the capacity to suffer, like a human being or an animal does. Basl (2014), for example, invokes the concept of psychological interest and concludes that until artificial consciousness is achieved, the welfare of machines is a non-issue in moral deliberations. Floridi (2002), on the other hand, has argued for the intrinsic (moral) value of what he refers to as information objects, a broad category of which AI systems form a relatively small and seemingly unremarkable subset. In any case, in practice we do often concern ourselves with, for want of a better word, the welfare of entities that cannot reasonably be argued to have psychological interests. This is illustrated by, for example, the international moral outrage triggered by the destruction of the Buddhas of Bamiyan in 2001 (Centlivres 2008) and the mourning of the partial destruction of Notre-Dame de Paris in a fire in 2019.

In both of the above examples, it can be reasonably assumed that for many of the people emotionally affected by the incidents, the artifacts in question did not hold any particular personal or religious significance. Instead, much of the sadness and anger came from a sense that something of immense cultural significance to humanity at large had been irrevocably lost. This seems to imply that while it may be a stretch to call inanimate objects moral patients, they may nevertheless embody values that moral agents have, within some parameters, an obligation to preserve. For the sake of brevity, and to distinguish this quality of manmade artifacts from moral patiency in the strict sense, let us refer to it as the *moral worth* of such objects.

If we accept that some artifacts have moral worth as defined above, the logical next question to ask is to which artifacts this concept of moral worth should be extended. In particular, are there some circumstances under which we would have to ascribe some degree of moral worth to AI-based entities? Furthermore, do AI-based entities represent a special case in this context, or can they be subsumed under some more general category of artifact? This paper examines these questions by exploring parallels and distinctions between AI-based entities and other types of cultural and technological artifacts; based on these, it is argued that the question of inherent moral worth of AI artifacts is relevant to practical AI ethics in the here and now.

The remainder of this paper is organized as follows: In Section 2, we examine more closely the notion of cultural artifacts having moral worth, starting with physical artifacts and expanding the discussion from there to digital artifacts. In Section 3, we

then apply the findings of this examination to AI-based entities in order to better understand if and when such entities could be considered to have intrinsic value as artifacts of cultural significance. Finally, in Section 4, we turn to the question of how AI-based entities deemed valuable should be preserved for posterity. The conclusions are presented in Section 5.

2 Moral worth of artifacts

In order to have a more formal account of the concept of moral worth to work with, let us first establish some basic criteria for artifacts that can be said to possess this quality. Above, we identified the cultural significance of an artifact as the source or foundation of its moral worth, but what kind of significance are we talking about more specifically? Intuitively, at least the following criteria must be satisfied:

1. The significance of the artifact must not be (wholly) instrumental in nature; it must be perceived to possess some value simply by virtue of existing.
2. The significance must be based on collective interest; ownership of the value of the artifact must be perceived as shared among humanity at large or a substantial subset thereof.

Both criteria are necessary. If the first one is not satisfied, the value of the artifact is not intrinsic but depends on its potential to be used to achieve some purpose that could also be achieved by some other means. If, on the other hand, the second one is not satisfied, the value depends on a specific individual or group of individuals having a claim to it, and damaging the artifact is a moral wrong against these individuals rather than the artifact itself.

If, however, both criteria are satisfied, there is some justification for saying that the artifact has genuine intrinsic value, independent of its utility or its relationships with individual stakeholders, and that preserving that value is a collective moral duty. With this definition, we can now proceed to examine the properties of an artifact that contribute to its cultural significance and, by extension, to its moral worth.

2.1 Physical artifacts

As an archetypal example of a highly significant cultural artifact, let us consider the Mona Lisa. Should the painting be accidentally or deliberately destroyed, this would no doubt be considered a tragic loss by many people around the world, so the criterion of collective interest is satisfied. Moreover, while for some people the significance of the painting is at least partially instrumental in nature – for example, they have an opportunity to go view it in the Louvre and get pleasure for doing so – this is not generally the case.

Why is this piece considered so important and valuable? There are several obvious factors that contribute to its perceived significance:

- The painting has great *aesthetic value*, being universally regarded as a masterpiece.
- It is *associated* with a notable historic figure, Leonardo da Vinci.

- It is considered highly *representative* of a particular time period and cultural context, namely the Italian Renaissance.
- It is relatively *ancient*, having survived in remarkably good condition for five centuries.

Crucially, the painting is also *rare*, in fact *unique*: there is, and can be, only one original piece in existence. Even in the hypothetical case that it were possible to perfectly replicate (rather than just imitate) it, it seems unlikely that a replica would be considered in every way equivalent to the original, since the original would still be a distinct individual perceived to possess non-material properties that the replica does not have. Destruction of the original would therefore constitute irrevocable loss of a major piece of humankind's collective cultural heritage, and protecting it against such an occurrence is arguably a moral imperative.

Notably, Seddon (2011) explicitly argues that cultural heritage can be considered a moral patient, and while he may be using the term in a less strict sense than, for example, Basl, the point is that cultural heritage is inherently worthy of moral consideration, as opposed to its moral worth depending on its contribution to human welfare. Cultural heritage is a broad concept, covering much more than just physical artifacts, so there is no reason to exclude digital artifacts from consideration. However, digital artifacts have certain special characteristics that make it necessary to discuss them as a distinct category.

2.2 Digital artifacts

Like the physical world, the digital world provides a medium for diverse forms of creative expression. It is worth noting that these are not limited to digital counterparts of traditional forms: the programmability of digital objects enables them to evolve over time in response to internal or external stimuli, deterministically or nondeterministically. In other words, the digital medium enables artifacts to have *behavior*, which until the Information Age was only achievable in a limited fashion through highly specialized engineering.

Whether there exists a digitally created work of art that is, or one day will be, comparable in significance to the Mona Lisa is debatable. There may be works that are considered by experts in the field of digital art to be masterpieces, but the field is so young that we simply do not have the kind of historical perspective on it that we do on the Renaissance. However, gaining such a perspective is literally a matter of time, and there is no evident reason why a digital artifact could not, if preserved, eventually become something universally regarded as an important piece of cultural heritage.

The question of rarity or uniqueness in the context of digital artifacts is somewhat complicated and merits a separate discussion, since it is another aspect that distinguishes them from physical artifacts. An artifact composed entirely of digital data can be replicated endlessly, with each replica being perfectly identical to the original. This being the case, it does not seem to make sense to identify different instances of the artifact as separate individuals, although the rising interest in buying and selling digital items as non-fungible tokens (Kugler 2021) is perhaps generating some confusion about this.

The fact that the concept of individuality generally does not make sense in the context of digital artifacts has implications for their preservation. If an artifact is purely digital, then the form of its physical manifestation is wholly irrelevant, so even if there is a specific copy of the artifact that can with some justification be called the original, long-term preservation of the artifact is likely to require discarding it in favor of retaining the essence of the artifact. Not only may it be necessary to make copies of the artifact to avoid losing it due to the original storage medium being unreliable: it may also need to be converted into a different representation format to account for the original format being rendered obsolete. Furthermore, if the artifact has behavior, it may even be necessary to rewrite some of its program code to ensure that the essence of the artifact is preserved even after the original implementation platform can no longer be kept functional.

3 AI entities as cultural artifacts

Above, when discussing the cultural significance of artifacts, we have focused on works of art, mainly because great works of art serve well to illustrate the idea of cultural significance translating into moral worth. While there are some artifacts incorporating AI that are explicitly conceived as art (see e.g. Tidemann & Brandtsegg 2015), generally AI-based entities are not works of art but of engineering. This does not exclude such entities from the realm of cultural artifacts, but may affect how we assess their cultural significance. In this section we will examine the nature of AI-based entities as cultural artifacts in more detail in order to better understand what – if anything – about them makes them interesting as a special category.

3.1 Cultural significance of AI entities

Above, in Subsection 2.1, we tentatively observed that the perceived cultural significance of a great work of art is not explained by any single factor but rather by an *intersection* of multiple factors. By the same reasoning, if we want to identify AI-based entities of high cultural significance, we can find particularly interesting candidates among those where several contributing factors intersect. In fact, we can identify several parallels here between AI-based entities and works of art:

- Association: a culturally significant AI entity could be, for example, one associated with a milestone achievement or notable figure in the history of AI research.
- Representativeness: a culturally significant AI entity could be one that is considered a classic example of a particular approach to AI and/or a particular period in the history of AI.
- Aesthetic value: conceivably, e.g. a highly novel and elegant neural network architecture could be considered to have substantial cultural value regardless of the instrumental value of its practical applications; traditional software engineering is already known to have its own aesthetics where code may be

judged “beautiful” or “ugly” independently of its technical correctness (see e.g. Kozbelt et al. 2012).

Applying the concept of gaining significance through age is, for reasons discussed above, somewhat problematic here. For one thing, computer science is a young discipline and AI even younger, so even the oldest existing AI entities can hardly be characterized as ancient, except in highly relative terms. Furthermore, attempting to determine the (physical) age of a digital artifact may not even make sense, since there is no meaningful difference between any two copies of the artifact. If the entity includes both digital and physical components, the situation is more complicated; we shall return to this theme in Section 4.

The youth of AI, combined with the current boom in AI (and especially machine learning) research, means that there are many firsts being achieved – potential milestones of substantial cultural significance. Human intelligence being a natural yardstick for measuring progress in AI, many of these milestones entail an AI system defeating a human champion at a game regarded as a test of intelligence. A classic example of such a system is Deep Blue (Campbell et al. 2002), the IBM chess-playing computer that beat Garry Kasparov in a six-game match in 1997, making it the first AI chess player to defeat a reigning world champion. More recently, similar milestones have been achieved at games that are considerably harder for a computer to master, such as the classical board game Go (Chen 2016) and the television quiz game *Jeopardy!* (Ferrucci 2012); from a purely technical standpoint, these are more significant achievements, but on the other hand, the great cultural significance of the game of chess itself arguably amplifies the significance of Deep Blue versus Kasparov.

Another type of game commonly used as a test for the capabilities of AI systems is the imitation game, where both human and AI participants produce some kind of output, while human evaluators try to distinguish between human-generated and AI-generated outputs. The most famous example of such a game is undoubtedly the one proposed by Alan Turing, known as the Turing Test (Pinar Saygin et al. 2000), which tests the ability of an AI system to converse like a human. Various experiments have also been conducted to test whether people are able to distinguish between humangenerated and AI-generated art (Daniele et al. 2021). These latter ones are particularly interesting in the context of our discussion of AI-based entities as cultural artifacts, especially since the question of the authorship of AI-generated works of art can be viewed as another instance of the more general question of whether AI-based entities can have some form of agency. However, discussing this theme any further is outside the scope of this paper.

3.2 Distinguishing characteristics of AI entities

The AI in an AI-based entity is a software artifact, but whereas the behavior of traditional software is pre-determined by the programmer, AI software, specifically AI based on machine learning (ML), is a special case. An ML system is, in a sense, programmed by itself: at the heart of it is a computational model whose parameters are not set by the developers but learned by the system by applying a learning

algorithm to a set of input data. In other words, what distinguishes AI-based entities from other software-based entities is that in addition to behavior, they have the *ability to learn new behavior*, i.e. to alter their own behavior in a highly intricate manner in response to new information.

Particularly when an AI system continues to learn from new data after initial training and deployment, and when the new inputs are uncontrollable, the system may develop a certain degree of individual personality, which is something that never happens with software programmed in the traditional way. There has recently been a lot of academic and even mainstream interest in language models such as GPT-3 (Floridi & Chiriatti 2020), and these exemplify how deploying an AI system in an uncontrolled environment and leaving it to learn from stimuli coming from that environment may lead to unanticipated outcomes. One notable case from recent years is that of Tay, Microsoft's experimental Twitter bot, which began to generate inflammatory content as a result of its interactions with human Twitter users (Hunt 2016).

In general, the opaque and unpredictable nature of some AI systems constitutes a major part of what makes them potentially ethically problematic when they are allowed to make decisions without human supervision. However, arguably this ability to develop a personality of sorts could also be considered another contributing factor when assessing the moral worth of an AI-based entity. The implication here is not that such an entity should be regarded as a person in any sense, but that there is some intrinsic value to be found in emergent AI behavior that is novel and interesting. Developing meaningful standards for this is, of course, a different matter.

4 Preserving AI entities for posterity

When we think about preserving artifacts as cultural heritage, we tend to think about museums. Interestingly, literature searches combining the term “artificial intelligence” with terms such as “museum” or “cultural heritage” mainly turn up a large number of papers on applications of AI in studying and preserving cultural heritage and creating museum experiences. Few, if any, of the results are about AI *as* cultural heritage, so academic interest in this area appears to be primarily directed at AI systems as tools rather than as artifacts interesting in themselves.

This does not mean, however, that AI is absent from museums except as a supporting technology. The Computer History Museum in Mountain View, California, for example, has had exhibitions with a substantial AI aspect, such as *Mastering The Game: A History of Computer Chess* and *Where To? A History of Autonomous Vehicles* (Computer History Museum 2005, 2014). The former is particularly interesting, not only because it featured the previously mentioned Deep Blue and its famous match against Garry Kasparov, but also because digital games in general commonly include some form of AI and have recently come to be regarded as cultural heritage to be preserved. We shall therefore briefly discuss the preservation of video games before moving on to the question of what it means to preserve an AI-based entity.

4.1 Games as cultural heritage

Video games as cultural heritage have been studied by a number of scholars across a range of disciplines. Guttenbrunner et al. (2010), Barbier (2014), Maier (2015), Eklund et al. (2019) and Guay-Bélanger (2021) have all examined the principles, practices and challenges of the preservation of video games from various perspectives. Among these, Eklund et al. and Guay-Bélanger discuss the preservation of games as cultural heritage on a conceptual level, while Guttenbrunner et al. and Maier examine the technical and legal challenges involved. Barbier discusses preservation efforts taking place outside traditional cultural institutions.

A common theme running through all of these discussions is the idea that if preservation efforts focus on exhibiting physical artifacts such as game boxes and gaming hardware, something essential about the games as cultural heritage will be lost. The interactive nature of games distinguishes them from other works of (popular) culture, and preserving this nature is obviously not achieved unless the game can be played. Eklund et al. and Guay-Bélanger go beyond this and argue for a broader view of how game culture should be understood and preserved, but there appears to be a consensus that preserving games as playable experiences is in any case a necessary element of preserving the cultural heritage they represent.

As discussed in Subsection 2.2, the operational lifetime of software artifacts is limited, and games are no exception: the disks, cartridges and other media on which games are stored decay over time, and perhaps even more importantly, so does the hardware of the consoles and computers on which the games are played. Long-term preservation of the interactive experience therefore generally requires opting for emulating the games on modern platforms rather than investing in an (ultimately futile) effort to keep the original hardware and software functional. This approach even makes it possible for games to remain commercially relevant long after the original release platform has become obsolete, as exemplified by the video game vendor GOG.com, which offers a selection of classic games made playable on modern platforms using emulators such as DOSBox.

4.2 Preserving the essence of AI entities

In the preservation of AI-based entities we encounter a similar situation where preserving the behavior of the system is not, in the long run, compatible with preserving the entity in its original configuration. Specific physical components may, of course, be worth preserving in themselves; to continue with our running example, the Deep Blue hardware used in the seminal Kasparov match would be an obvious candidate. However, from an AI perspective, the essence of Deep Blue was its chessplaying ability, and should we wish to preserve this, it would have to be done in such a way that the preserved form does not depend on the original implementation platform. As with other digital games, this could be accomplished by emulating the Deep Blue software on another platform.

We can extend the notion of preserving the interactive experience to cover other types of interactive AI-based entities, such as conversational agents. However,

interactive applications are but a subset of AI systems, and if the system does not involve an interactive experience, then it may not be necessary to preserve it as a running system. The behavior of an ML-based AI system emerges from the process by which the system is trained, and in theory, the behavior could be preserved in a completely platform-independent way by meticulously documenting the training process in such detail that the behavior can be reproduced as long as there is any platform capable of performing the required computations available.

In practice, this will not be always technically feasible. Furthermore, in those cases where it is feasible, the result will inevitably be a snapshot of the system frozen in time, which may be what is desired, but not necessarily. Above, in Subsection 3.2, we observed that an essential distinguishing feature of some AI-based entities is their ability to learn continuously from new inputs, potentially resulting in them developing their own individual personalities. In such cases it could be argued that to fully capture and preserve the essence of the entity in question, it should be maintained in an operational state *in a dynamic environment* where it continues to receive new data and the resulting changes in its behavior can be observed.

5 Conclusion

In this paper we examined the question of AI and moral patiency from an angle that has not been previously adopted in the literature. While it seems clear that no AI-based entity created so far is a moral patient in the sense of having moral rights, it was argued that some of them may have some degree of moral worth, derived from their significance as cultural artifacts. By comparing and contrasting AI-based entities with other types of artifacts, we arrived at partial answers to three questions: 1) when can an AI-based entity be said to have cultural significance, 2) what constitutes the essence of an AI-based entity, and 3) how should the essence of a culturally significant AI-based entity be preserved.

The translation of cultural significance into moral worth is hardly a straightforward matter. Moreover, AI is a relatively young discipline, and it is difficult if not impossible to see, from today's perspective, which parts of its history so far might one day be considered so monumentally important that preserving them would be viewed as a collective moral duty. On the other hand, there seems to be no reason to assume an upper limit on the cultural significance of AI-based entities that would prevent them from ever achieving such a status. Given that practical AI ethics has been all about acknowledging the importance of intrinsic values in guiding the development and evaluating the impact of AI systems, it would be a natural development to start looking into the intrinsic value of the systems themselves.

References

- Barbier, B. (2014) Video Games and Heritage: Amateur Preservation? *Hybrid*, (1).
<https://doi.org/10.4000/hybrid.1107>

- Basl, J. (2014) Machines as Moral Patients We Shouldn't Care About (Yet): The Interests and Welfare of Current Machines. *Philosophy & Technology*, 27(1), 79–96. <https://doi.org/10.1007/s13347-013-0122-y>
- Campbell, M., Hoane, A. J. & Hsu, F.-h. (2002) Deep Blue. *Artificial Intelligence*, 134(1–2), 57–83. [https://doi.org/10.1016/S0004-3702\(01\)00129-1](https://doi.org/10.1016/S0004-3702(01)00129-1)
- Centlivres, P. (2008) The Controversy over the Buddhas of Bamiyan. *South Asia Multidisciplinary Academic Journal*, (2). <https://doi.org/10.4000/samaj.992>
- Cervantes, J.-A., López, S., Rodríguez, L.-F., Cervantes, S., Cervantes, F. & Ramos, F. (2020) Artificial Moral Agents: A Survey of the Current Status. *Science and Engineering Ethics*, 26(2), 501–532. <https://doi.org/10.1007/s11948-019-00151-x>
- Chen, J. X. (2016) The Evolution of Computing: AlphaGo. *Computing in Science and Engineering*, 18(4), 4–7. <https://doi.org/10.1109/MCSE.2016.74>
- Computer History Museum (2005, August 11) Computer History Museum Debuts New Exhibit, Mastering the Game: A History of Computer Chess. Retrieved from <https://computerhistory.org/press-releases/new-exhibit-mastering-the-game-a-history-of-computer-chess/>
- Computer History Museum (2014, May 08) Computer History Museum Launches New Exhibit on the History of Autonomous Vehicles. Retrieved from <https://computerhistory.org/press-releases/where-to-exhibit/>
- Daniele, A., Di Bernardi Luft, C. & Bryan-Kinns, N. (2021). “What Is Human?” A Turing Test for Artistic Creativity. In: *Artificial Intelligence in Music, Sound, Art and Design. EvoMUSART 2021*. Lecture Notes in Computer Science, vol. 12693. Springer, Cham. https://doi.org/10.1007/978-3-030-72914-1_26
- Eklund, L., Sjöblom, B. & Prax, P. (2019) Lost in Translation: Video Games Becoming Cultural Heritage?. *Cultural Sociology*, 13(4), 444–460. <https://doi.org/10.1177%2F1749975519852501>
- Ferrucci, D. A. (2012) Introduction to “This is Watson”. *IBM Journal of Research and Development*, 56(3.4), 1:1–1:15. <https://doi.org/10.1147/JRD.2012.2184356>
- Floridi, L. (2002) On the intrinsic value of information objects and the infosphere. *Ethics and Information Technology*, 4(4), 287–304. <https://doi.org/10.1023/A:1021342422699>
- Floridi, L. & Chiriatti, M. (2020) GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, 30(4), 681–694. <https://doi.org/10.1007/s11023-020-09548-1>
- Guay-Bélanger, D. (2021) Assembling Auras: Towards a Methodology for the Preservation and Study of Video Games as Cultural Heritage Artefacts. *Games and Culture*. <https://doi.org/10.1177%2F15554120211020381>
- Guttenbrunner, M., Becker, C. & Rauber, A. (2010) Keeping the Game Alive: Evaluating Strategies for the Preservation of Console Video Games. *International Journal of Digital Curation*, 5(1), 64–90. <https://doi.org/10.2218/ijdc.v5i1.144>
- Hunt, E. (2016, March 24) Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. Retrieved from <https://www.theguardian.com/technology/2016/mar/24/taymicrosofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter>

- Kozbelt, A., Dexter, S., Dolese, M. & Seidel, A. (2012) The aesthetics of software code: A quantitative exploration. *Psychology of Aesthetics, Creativity, and the Arts*, 6(1), 57–65. <https://psycnet.apa.org/doi/10.1037/a0025426>
- Kugler, L. (2021) Non-Fungible Tokens and the Future of Art. *Communications of the ACM*, 64(9), 19–20. <https://doi.org/10.1145/3474355>
- Maier, H. (2015) Games as Cultural Heritage: Copyright Challenges for Preserving (Orphan) Video Games in the EU. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*, 6(2), 120–131.
- Pinar Saygin, A., Cicekli, I. & Akman, V. (2000) Turing Test: 50 Years Later. *Minds and Machines*, 10(4), 463–518. <https://doi.org/10.1023/A:1011288000451>
- Seddon, R. F. J. (2011) *The Ethical Patiency of Cultural Heritage*. Dissertation, Durham University. Retrieved from <https://philpapers.org/rec/SEDTEP>
- Tidemann, A. & Brandtsegg, Ø. (2015) [self.]: an Interactive Art Installation that Embodies Artificial Intelligence and Creativity. In: *C&C '15: Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*. Association for Computing Machinery, New York. <https://doi.org/10.1145/2757226.2764549>

Working with Affective Computing: Exploring UK Public Perceptions of AI enabled Workplace Surveillance.

Lachlan Urquhart [0000-0001-5144-5024] Alexander Laffer 2[0000-0003-2463-9135] and Diana
Miranda 3[0000-0002-8605-5031]

¹ University of Edinburgh, Edinburgh, Scotland

² Bangor University, Bangor, Wales

³ University of Stirling, Stirling, Scotland

lachlan.urquhart@ed.ac.uk

Abstract. This paper explores public perceptions around the role of affective computing in the workplace. It uses a series of design fictions with 46 UK based participants, unpacking their perspectives on the advantages and disadvantages of tracking the emotional state of workers. The scenario focuses on mundane uses of biometric sensing in a sales environment, and how this could shape management approaches with workers. The paper structure is as follows: section 1 provides a brief introduction; section 2 provides an overview of the innovative design fiction methodology; section 3 explores wider shifts around IT in the workplace; section 4 provides some legal analysis exploring emergence of AI in the workplace; and section 5 presents themes from the study data. The latter section includes discussion on concerns around functionality and accuracy of affective computing systems, and their impacts on surveillance, human agency, and worker/management interactions.

Keywords: affective computing; surveillance studies; workplace monitoring; design fiction; human agency; emotions.

1 Introduction: Affective Computing in the Workplace

Affective computing (AC) systems are emerging in the workplace changing the nature of workplace cultures and enabling technologically mediated professional relationships. Better understanding of employees' emotions may have benefits like protecting worker wellbeing, but significant risks are likely to emerge such as forms

of tracking that benefit employers at the expense of employee interests. In this paper, we explore public perceptions of mundane uses of AC and biometric surveillance systems at work. Through a series of ten online workshops with 46 members of the UK public, we used a novel design fiction led methodology to elicit insights. We observed concerns around AC functionality and accuracy, negative impacts on human agency and interactions, and general anxieties around expanding surveillance infrastructures. To contextualise these concerns beyond our empirical findings, we present analysis of the legal dimensions of AI use at work and the wider histories of information technologies being integrated into the workplace.

2 Methodology: Design Fictions

To explore a range of Affective Computing (AC) use-cases in a situated manner, we developed an innovative narrative approach (Laffer 2022). This draws on **Design Fiction** (Bleecker 2009; Jensen & Vistisen 2017; Coulton et al, 2017) – notably the use of diegetic prototypes; technology that exists within a fictional world – and aspects of **Contravision** (Mancini et al. 2010), where positive and negative outcomes for each use-case are created. A storyline narrative, created using Twine (an interactive fiction writing tool), was presented in online workshops to engage our participants (n=46) with mundane examples of AC. This supported a narrower focus on people, social practices, and technological impacts. This process of ‘domestication’ (Auger 2013) helped to quickly familiarise participants with socio-technical aspects of emergent systems and go beyond just the technological aspects. The full Twine can be viewed here: <https://eaitwine.neocities.org/> and the opening image of the AffecTECH narrative is provided below (Figure 1).



Figure 1: Advert for AffecTech, a fictional technology company, presented to participants during online workshops as part of a Twine narrative.

Participants were separated into three older groups (65+) (n=13); three younger groups (18-34) (n=12), two groups of disabled participants (n=10) and two groups of people belonging to UK ethnic minorities (n=11). Workshop discussions were recorded, transcribed, and uploaded to qualitative analysis software Nvivo for thematic analysis.

In our workplace scenario, a new emotion sensing system (AffecTECH) has been introduced into a call center so employees making sales calls are monitored and evaluated through voice and other biometric data. The system provides immediate performance feedback and maintains these records for use in staff evaluation by managers. The technology is introduced via a diegetic email (figure 2) from the protagonists' manager. The contravision element is composed of positive and negative feedback about the system by two of the protagonists' colleagues who had piloted the technology.

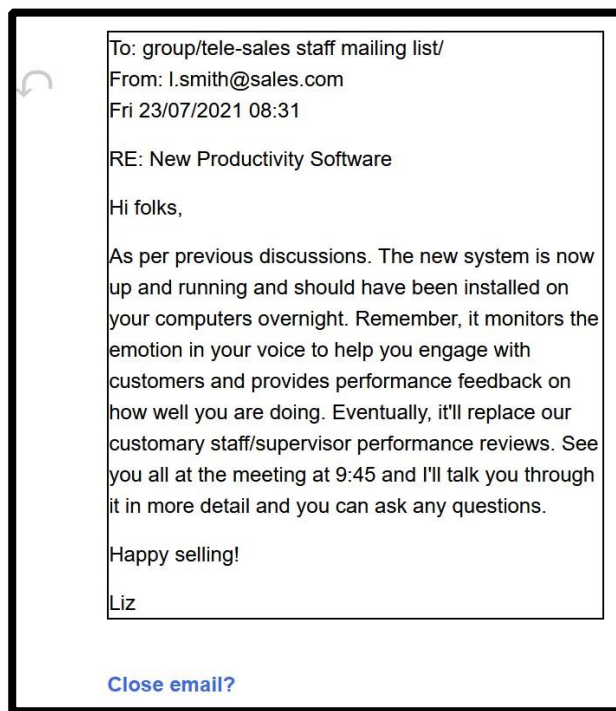


Figure 2: Email introducing the workplace AC use-case presented to workshop participants as part of a Twine narrative.

3 Technologically Mediated Workplace Surveillance

Before presenting results, it is useful to reflect on wider literature and concerns around incorporation of new IT systems into the workplace, as this has long raised legal and design issues. For example, in Scandinavia, researchers and trade unions in the 1970's and 80's tried to shape the early emergence of IT at work to address risks like deskilling of workers tasks, and IT enabling poor management practices (Bjerknes and Bratteig, 1995). Their goal was to instill a form of **workplace democracy** where workers participated in the design of IT to ensure they would continue to flourish at work and new IT systems would not negatively impact existing working practices (Floyd, 1989). Moving forward to present day, current workplace IT incorporates data driven machine learning tools for overt or covert profiling and management of the workforce (Edwards, Martin and Henderson, 2019). As Jarrahi et al (2021) highlight, we are moving to an age of 'algorithmic management' in workplaces. Yet lessons from 40 years ago have not been learned, and many ethical and legal concerns remain. Four brief examples show this. **Gig economy platforms** such as Deliveroo or Uber track and monitor movements and performance of workers remotely, yet workers have limited agency to change the platform (Prassl, 2019). **Automated hiring systems** utilise machine learning to determine if the candidates match arbitrary psychometrics, which can in turn lead to discrimination (Monedero, Dencik, and Edwards, 2020). Some systems even claim to detect emotional states, reiterating the turn to metricising emotions (HireVue – Chen and Hao, 2020). Employer **wellness programmes** utilise employee wearables to track vital statistics (e.g., heart rate, activity) but can leave privacy concerns around data handling unaddressed (Iliadis and Pederson, 2018). Exacerbated by the shift to remote working prompted by the global Covid-19 pandemic, employers have been increasingly monitoring employees through **covert surveillance using webcams, facial recognition, and bespoke software**. This challenges trust relationships between employer and employee, raises privacy concerns, and surfaces differentiated impacts of surveillance, particularly across genders (Stark, Stanhaus and Anthony 2020). Yet, we see commercial vendors such as Uniphore offer to monitor customer sentiment to provide opportunities for sellers to add value to their calls, and build a trusting relationship with a customer¹. Similarly, video conferencing suite Zoom, is looking to deploy emotional AI in video analytics.² Given the concerns around the ability of emotional AI to even detect emotions in the first place, and the resulting criticisms of it being a form of 'pseudoscience' (Crawford, 2021), the commercial plans to roll out such systems at scale in workplaces are concerning.

These shifts appear to be here to stay. Ball (2021) has highlighted that “thoughts, feelings, and physiology” are a key target in **workplace surveillance** through use of biometrics, wearables, and emotion tracking e.g., in call centres (Bromuri, Henkel, Iren and Irovi, 2020). Ball's earlier work on call centres explored negative impacts of extensive workplace surveillance (Ball and Margulis, 2011), and as monitoring of

¹ <https://www.uniphore.com/>

² <https://www.protocol.com/enterprise/emotion-ai-sales-virtual-zoom>

voice tone is delegated to machines in the future (McStay, 2018), our scenario draws together these trends.

4 Surely the Law can help?

To tackle ethical and legal challenges posed by AC, the **proposed EU AI Act**³ may benefit workers by preventing certain AI applications from ever becoming available on the EU market in the first place. The legislation seeks to **regulate AI** in a product safety driven manner, utilising different standards depending on the levels of risk a system poses, from prohibited, to high to low risk. Hiring systems, for example, are deemed to be **high-risk AI systems** (HRAIS) in the AI Act. These are systems where AI is used for ‘recruitment or selection of natural persons, including ‘advertising vacancies, screening or filtering applications, evaluating candidates in the course of interviews or tests’ (Annex III, AIA). Other workplace HRAIS include AI used for making ‘decisions on promotion or termination of work-related contractual relationships, for task allocation, and for monitoring performance and behaviour of persons in such relationships’ (Annex III, AIA). A system being deemed a HRAIS means it faces a wide range of design and development requirements for the providers and users/deployers (McStay and Urquhart, 2022). For example, there are requirements to comply with strict data governance requirements around training datasets; providing human oversight for operators to intervene when systems go wrong; and use of conformity assessment processes to ensure systems adhere to legal requirements before being available for sale (Urquhart, Crabtree and McGarry, 2022). Yet, not all emotion sensing technologies are treated as strictly. In other contexts, they are only set to be regulated as **low risk systems**, requiring operators merely to be transparent with persons that they are interacting with an AI in the first place (if it is not already obvious from the context of use) (Art 52, AIA).

EU data protection authorities are concerned that the provisions in the AIA are not strict enough. They are calling for all emotional AI systems to be treated as **prohibited** forms of AI, unless for research or health purposes (EDPS/EDPR, 2021). This would see them treated in the same way as other prohibited systems such as social crediting or automated, live, public space facial recognition by the police (see Urquhart and Miranda, 2021). Further, unlike with data protection laws such as the EU General Data Protection Regulation 2016, which provides a range of data subject rights over personal data, the proposed AI Act does not provide direct rights for individuals who may be subject to AI systems (Edwards, 2022). This is concerning, given the gaps within the existing data protection and privacy law landscape workers data interests, and the AI Act will not provide direct recourse either.

A recent report from Allen and Masters (2021) examines the challenges of AI in workplace and the issues workers face in asserting control over how their data is used by employers. For example, they highlight that under European Convention on Human Rights case law there can still be a reasonable expectation to private and

³ <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>

family life during employment (as per *Barbulescu v Romania*).⁴ The state needs to provide appropriate legal frameworks to guard against unnecessary, disproportionate intrusions by employers, and the facts of any specific case will be key to determining if an intrusion is justifiable or not. Under data protection law even in employment relationships, a **lawful basis** is needed for processing data (Art 6, GDPR), which guards against employers acting in uninhibited ways. The (former) EU Article 29 Working Party (WP)⁵ highlights that **consent** can rarely be a legal ground for processing in an employment setting, due to it not being freely given (A29 WP Report, 2017). Other grounds, such as **legitimate interests** of the employer (such as to monitor for fraudulent employee activity on networks) are possible but need to be balanced against interests of workers whilst also being proportionate and necessary. The most common legal ground for employers processing employee personal data is **necessity** ‘for the performance of the employment of contract the data subject is party to... or in order to enter into a contract’ (Art 6(1)(b) GDPR). Whilst this is broad and can capture many uses of personal data in the workplace, it is not limitless. The Art 29 WP stresses that data minimisation in any workplace system is important to protect workers, by not storing data longer than needed, and ensuring that any monitoring is transparent and covered by clear workplace policies. One safeguard for workers when employers are conducting high risk data processing activities (which Emotional AI systems like AffectTECH would likely be) is for them to conduct a **data protection impact assessment** (DPIA) (Art 35, GDPR). A DPIA is a valuable tool for demonstrating compliance with GDPR more broadly, as per Art 5(2) of the legislation (see Urquhart, Lodge and Crabtree, 2019). There is some discretion in how impact assessments can be carried out, and some uncertainty persists around when high risk processing occurs (this is the trigger to conduct a DPIA). The UK Information Commissioner Office has usefully helped demystify this for the employment context by providing high risk examples namely: using employee data at scale in profiling e.g. in recruiting and accessing programmes; use of biometrics for access to workplaces or identity verification including fingerprint/voice/facial recognition; tracking location of employees and CCTV monitoring; remote working monitoring; web and cross device tracking; and large scale profiling via wearable based health monitoring.⁶ It is clear AC technologies could integrate with a large variety of these systems, and their intrusiveness suggests they would similarly be deemed high risk and require a DPIA.

Another legal development that may help stem some of the issues around AI and AC use in the workplace comes from proposed **EU Directive for Improving Working Conditions in Platform Work** (IWCPL)⁷. This targets specific types of work, namely gig economy platforms such as Uber and Deliveroo, as opposed to

⁴ <https://hudoc.echr.coe.int/eng?i=001-159906>

⁵ Now the EU Data Protection Board. This older report was still written to reflect GDPR changes, so remains relevant guidance.

⁶ <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-dataprotection-regulation-gdpr/data-protection-impact-assessments-dpias/examples-of-processinglikely-to-result-in-high-risk/>

⁷ <https://ec.europa.eu/social/main.jsp?langId=en&catId=89&newsId=10120&furtherNews=yes>

workplaces more generally. Given the exploitation these workers often face though, this is a valuable step in protecting them. The proposed law seeks to bring in new mechanisms to help workers deal with **algorithmic tracking**. For example, Article 6 provides for transparency requirements on use of automated monitoring/decision making systems. Here platforms need to tell workers about ‘**automated monitoring systems**’ that are electronically supervising and evaluating their performance. They also need to disclose information about ‘**automated decision-making systems**’ which take decisions ‘significantly affecting’ the workers conditions. This includes information around the ‘categories of actions monitored, supervised and evaluated’ not just of the platform itself but also by clients and what ‘main parameters’ factor into automated decision making processes. It also limits platforms use of personal data of workers where it has to be data ‘intrinsically connected to and strictly necessary for the performance of their contract’ and cannot include their ‘emotional state’, their health, psychology, private conversations, and activities during ‘off time’ (Article 6, IWCPL). This is coupled with provisions for human monitoring of automated systems (Article 7, IWCPL) and **review of significant automated system decisions** when they impact working conditions for workers. This entitles workers to discussion and clarification of the circumstances, reasons and facts around a decision with a ‘human contact person at the digital labour platform’ (Art 8, IWCPL). This law promises some positive changes and highlights the type of thinking that would benefit other work contexts too (e.g., call centres). We now

turn to our data to unpack public concerns and the sorts of issues future legal frameworks may need to address.

5 Data and Analysis: Public Perceptions of AC & Design Fictions

5.1 Functionality and accuracy of EAI systems

Most participants were **ambivalent or negative towards our affective computing (AC) use case**, being particularly concerned with the **punitive potential** of the technology and the opportunity for employers to use them as a means to fire staff without engaging with them. Some participants even argued that the systems should only be implemented if they were used to benefit the employee, and that there should be a ‘*no detriment policy*’ (Tom Y2) to their use.

Despite participants’ concerns, a range of **benefits** were highlighted, chiefly around the potential to **improve performance**. Some participants felt it could be used as a training tool, helping individuals to ‘*monitor their performance*’ and ‘*know where to improve*’ and the company can then ‘*help the person to develop the sort of spots where they are lacking*’ (Elias Y3). Similarly, it was argued that the system could usefully counter perceived individual failings: ‘*...if the person that was being evaluated was surly and bad tempered with the customers all the time, and it spots a trend, fine. That person, for want of a better phrase, might need reeducating*’ (Cliff D1). Other participants felt it could help with difficult situations and customers,

potentially even supporting neurodiverse customers. The older demographic were most positive in this regard, drawing on their own experiences as customers or those having to deal with difficult customers.

Another suggested benefit of the system was that it might contribute to **improvements in workplace culture**, for example by providing a degree of objectivity in evaluation, *'because you're going to cut out any nepotism or any favoritism, because everyone's going to be judged on the same caliber'* (Paul Y2). Impersonal feedback from an affective computer system was also seen as a benefit by one participant in its capacity to reduce interpersonal threat between staff and managers, heightened due to asymmetries in power: *'I think, I would like it better than the supervisors having to check and that and give you that kind of a view. Because this feels like, at least you can kind of question it without feeling like you're undermining their authority or their influence'* (Yasmine E2). Additionally, it might add **competition in a sales environment** or enhance **management opportunities**. These latter elements were dependent on employee characteristics and need to be *'specified to different sectors'*, as Alice (Y3) states, it is *'...dependent on what you're ringing up for...in terms of the model employee, you're going to have different personalities for different sectors and different positions.'* She continues to discuss how **context** is key to the value of the system for both employee and customer, noting *'...[if] I'm booking a holiday, that's fine. I don't mind someone being happy and chirpy. If I'm ringing up someone to discuss life insurance, you don't want someone on the end of the phone like, "Well, how long do you think you're going to live for?'*

Despite some discussion of benefits, most participants were **negative, even hostile**, to the use of emotion sensing tools in the workplace, highlighting their potential to add more **pressure on employees**, be **distracting** and be used **punitively**. Linda (O3) articulates these three points in her criticism of the system: *'...this could possibly cause someone to lose their job. And also, the pressure put on someone, knowing they're being constantly monitored and having little alerts popping up, must be horrendous. I'd hate it...'* Of even more concern is that it could have **mental health implications** for participants: *'I mean, under the circumstances, you're already stressed with the work. So I think this kind of technology will not help to your mental health.'* (Louis Y3). This anxiety over mental health implications was most pronounced in the groups consisting of disabled participants, both in terms of impact but also how it might contribute to discrimination: *'there's people here that, they've got anxiety, depression, panic attacks, whatever, it doesn't mean they can't sit and do the job but it would put them under even more stress. So that wouldn't be fair even though they could do the job perfectly well.'* (Penny D2). This reinforces the importance of attending to the needs of diverse groups to guard against discrimination of workers with protected characteristics. This is true in terms of research, designing an inclusive approach for collecting data on citizen attitudes, and more broadly when we consider the implementation of new forms of technology.

Participants went on to raise a number of important issues about the implementation of emotion sensing, concerned that it demonstrates a **lack of trust** in employees, with Andy (O1) arguing that employers should, *'[t]rust your employee or get another employee'*. This in turn contributes to **bad morale in the workforce**, who

feel ‘tense’ because ‘*You’ve got the feeling like the bosses can’t trust you, so they’re putting in a system*’ (Samuel Y1). The system itself is criticized as being **highly invasive**, potentially causing you to ‘lose people because they would feel that it’s too intrusive on their job’ (Joanne O2); and potentially open to **manipulation**, which in some cases will be inevitable, demonstrating a lack of trust in existing labour relations: ‘*It depends who is using the system because then they can manipulate it to do whatever and it will be manipulated in certain situations*’ (Carol Y3). Participants were skeptical of the **effectiveness** of such a system and that it may in fact be detrimental to customers. These negative factors are connected to notions of **accuracy** and participants’ view of the **over-simplistic** assumption that the system would be able to interpret the employees’, or indeed customers’, emotions correctly, failing to account for the prevalence of **individual variation**, that people are different and changeable. As Emily (D2) argues:

‘Everybody’s way of going about talking is different. How is it going to be that reliable? Say if you make a joke that’s slightly dry humor, sarcastic, the person on the end of the phone, that could be how you should interact with them because that could be them, but then the automatic bot that’s recording the calls might not pick up on that. It might pick up on that as being rude. I just think it’s not ideal’?

This key issue underpins a range of concerns around **how the data is interpreted**, the potential for **flaws in the system** and the involvement of different actors leading to **further inaccuracy**. This is illustrated by Cliff’s (O1) negative appraisal of the system and AI more generally. He argues that, ‘*AI is not perfect. It’s flawed, as we know*’ because ‘*[s]ystems are built by humans. Humans built in errors and things like that*’, which he exemplifies with reference to the 2019 Boeing airplane crashes where, he claims, ‘the system didn’t react as it was supposed to’.

In its most egregious form, there were anxieties around the potential for **coded bias and discrimination** embedded in these surveillance systems, understood by participants as when the system, ‘*picks out a certain subgroup that it favours in favour of another subgroup*’ (Patrick D2). This discrimination may take the form of unfairly limiting who is employed or negatively impacting an existing employee, leading to employees who ‘*don’t have any personality, they don’t have empathy. They don’t have anything else apart from just fitting what this machine wants them to.*’ (Patrick D2). Ultimately, for many participants, it was seen that the **system needs human intervention** to ensure it worked effectively and as intended and argued that EAI was a poor substitute for an effective manager.

Despite not separating the groups by gender, the data potentially supports Stark et al’s (2020) findings on gendered perceptions of workplace surveillance. We found that where there does seem to be divergence in group consensus, it is female participants (Linda O3; Rosie Y2; Brenda D1;) who seem more critical of the technology and/or male participants who present a more positive perspective (Elias Y3; Harry O3; Cliff D1) captured in this interchange from group D1:

Brenda: *I wouldn't be happy. I would feel, I think, totally invaded by it. And [partly causing] paranoia, I think.*

...

Brenda: *Feels all negative to me.*

Cliff: *I wouldn't say it's all negative.*

A caveat needs to be added that other demographic (and personal) characteristics might be contributing factors. For example, Harry and Elias could be distinguished from other group members based on politics or ethnicity.

5.2 Surveillance, worker agency and changing interactions

For some participants, this new emotion sensing system was seen as an **extension of surveillance mechanisms** already in place in workplace settings: *Systems like this, but nowhere near as sophisticated as this, already exist... these days there's software that looks up what you are doing as a worker. I don't know to what extent it measures performance, but it knows exactly what you do at any given time. This is just one stage further on, that's all.* (Tim O1). Signaling the existence of current surveillance technology contributes to a sense of legitimacy or at least resignation towards its implementation (Ball, 2021) in the group discussion. However, if for some participants this 'normalises' the system, for others this makes it even less acceptable as being more **invasive**, continuously present and **lacking human input**. At the most extreme end, it led to participants feeling '*totally invaded*', again contributing to negative mental health implications, such as '*paranoia*' (Brenda D1). This level of surveillance, combined with limitations on an employee's agency and opportunities for interaction, it leads to a workplace environment where employees become '*robots*' (Lauren D2) and dehumanised: '*they don't have any personality, they don't have any empathy. (...) just fitting what this machine wants them to.*' (Patrick D2).

Exacerbating these issues is the **power asymmetries and lack of control** employees have over the implementation, use and functionality of these systems. This leads to fears around how they would be used in automating management structures, and employee evaluation. Phillip (O1) articulates this well stating it may '*replace reviews and appraisals and that sort of thing. You could imagine, the option is there to be sacked automatically by email, without even speaking to anybody. And that would be a bit of a concern, I think. It's about, yeah, where it ends up.*' Similarly (Penny D2) states employees only option could be quitting if they do not like these systems.

6 Conclusions

We have explored the wider context of legal concerns and historical issues of incorporating new IT, like AC, into the workplace. By using our innovative design fiction approach to collecting citizen attitudes, this research has revealed a number of important concerns around implementation of AC in the workplace. While most participants were **ambivalent or negative towards** the use case, a number of **benefits**

were suggested. These focused on the technologies use for **training to improve performance** and potential to **enhance workplace culture**. However, these were both seen to be contingent on **context**, in terms of the individual and sector.

The pre-existence of a range of **surveillance mechanisms in the workplace legitimised** the implementation of AC systems for some participants, while others found themselves **resigned** to their use. Most participants were **strongly negative** towards the use of AC systems in the workplace, pointing to the **pressure** they put on employees and their potential to be **distracting** and used **punitively**, with **mental health implications**. This latter was most strongly seen in groups self-identifying as disabled, highlighting the importance of this research that collected citizen attitudes from groups who have traditionally been ignored in the development of new technology, the implementation of which may reinforce existing inequalities.

Participants argued that these surveillance systems contribute to **bad morale in the workforce**, highlighting a **lack of trust** in employees. The systems are seen as both **highly invasive** and **open to manipulation** while at the same time **in-effective** and **inaccurate** relying on **over-simplistic** interpretations that is unable to account for **individual variation**. Participants raised legitimate concerns about **systemic issues**, including **coded bias and discrimination** as well as the dangerous potential for **human error**. As with many of the other use-cases discussed in the workshop, participants were keen for some element of **human intervention** to support or balance out decisions made by AI systems.

A key concern emerging from this research is the reinforcement of awareness that these workplace uses of AC are being introduced in contexts with existing **power asymmetries** and they have great potential to magnify inequalities between employers and employees; this is something citizens understand and are anxious about, seeing the potential for the negative impact on labour relations and individual wellbeing. These concerns must be attended to in the development, implementation and legislation surrounding AC deployment in the workplace.

Acknowledgements

Funding: We would like to thank our funders for supporting this research under Economic and Social Research Council grant ES/T00696X/1 [All] and UKRI Engineering and Physical Science Research Council grants EP/T022493/1 and EP/V026607/1 [Urquhart]

References

- Allen, and Masters, D. (2021) *Technology Managing People – The Legal Implications*. Trade Union Congress.
- Article 29 Working Party, (2017) *Opinion 2/2017 on Data Processing at Work*. European Commission.

- Auger, J. (2013) Speculative Design: Crafting the Speculation. *Digital Creativity* 24(1) 11–35. <https://doi.org/10.1080/14626268.2013.767276>.
- Ball, K., (2021) Electronic Monitoring and Surveillance in the Workplace. *European Commission Joint Research Centre* DOI:10.2760/5137, JRC125716.
- Ball, K and Margulis S.T (2011) Electronic Monitoring and Surveillance in Call Centres: a Framework for Investigation. *New Technology, Work and Employment* 26(2) 113–126.
- Bjerknes, G. and Bratteteig, T. (1995) User Participation and Democracy: A Discussion of Scandinavian Research on System Development. *Scandinavian Journal of Information Systems*, 7(1):73–98
- Bleecker, J. (2009). *Design Fiction: A short essay on design, science, fact and fiction*. Near Future Laboratory. <http://blog.nearfuturelaboratory.com/2009/03/17/design-fiction-a-short-essay-on-design-science-fact-and-fiction/>
- Bromuri, S., Henkel, A. P., Iren, D., & Urovi, V. (2021). Using AI to predict service agent stress from emotion patterns in service interactions. *Journal of Service Management*, 32(4), 581–611. <https://doi.org/10.1108/josm-06-2019-0163>
- Chen, A. and Hao, K. (2020) Emotion AI researchers say overblown claims give their work a bad name. *MIT Technology Review*. <https://www.technologyreview.com/2020/02/14/844765/ai-emotion-recognitionaffective-computing-hirevue-regulation-ethics/>
- Coulton, P. Lindley, J.G. and Sturdee, M. and Stead, M. (2017) Design Fiction as World Building. In: *Proceedings of Research through Design Conference 2017*
- Crawford, K. (2021) *Atlas of AI*. Yale University Press
- Edwards, E (2022) *Regulating AI In Europe: Four Problems and Four Solutions* Ada Lovelace Institute.
- European Data Protection Board and European Data Protection Supervisor (2021) EDPB-EDPS Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).
- Floyd, C. Mehl, WM, Resin, F.M Schmidt, G. Wolf. G (1989) Out of Scandinavia: Alternative Approaches to Software Design and System Development. *Human Computer Interaction* 4, 253–350
- Iliadis, A; Pedersen, I (2018) The fabric of digital life: Uncovering sociotechnical tradeoffs in embodied computing through metadata *Journal of Information Communication & Ethics in Society*16 (3) 311- 327 10.1108/JICES-03-2018-0022.
- Jarrahi, M. H., Newlands, G., Lee, M. K., Wolf, C. T., Kinder, E., & Sutherland, W. (2021). Algorithmic management in a work context. *Big Data & Society*. <https://doi.org/10.1177/20539517211020332>
- Jensen, T., & Vistisen, P. (2017). Ethical Design Fiction: Between storytelling and world building. *The Orbit Journal*, 1(2) 1–14. <https://doi.org/10.29297/orbit.v1i2.56>

- Laffer, A (2022) Using an online narrative approach to explore diverse participants' understanding of emerging technology: Citizen's perspectives on living with emotional AI in *SAGE Research Methods: Doing Research Online*. London: Sage.
- Mancini, C., Rogers, Y., Bandara, A., Coe, T., Jedrzejczyk, L., Joinson, A., Price, B., Thomas, K., & Nuseibeh, B. (2010). ContraVision: Exploring Users' Reactions to Futuristic Technology. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 153–162. <https://doi.org/10.1145/1753326.1753350>
- McStay, A. (2018) *Emotional AI: The Rise of Empathic Media*. SAGE
- McStay, A and Urquhart, L. (2021) In Cars (are we really safest of all?): Interior Sensing and Emotional Opacity, *International Review of Law, Computers & Technology*, DOI: 10.1080/13600869.2021.2009181
- Prassl, J (2019) What If Your Boss Was an Algorithm? The Rise of Artificial Intelligence at Work. *Comparative Labor Law & Policy Journal* 41(1) 123.
- Sánchez-Monedero, J, Dencik, L, Edwards, L (2020) What does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. *ACM/FAT*20: proceedings of the 2020 conference on fairness, accountability, and transparency*. Available at: <https://dl.acm.org/doi/10.1145/3351095.3372849>
- Stark, L, Stanhaus, A. and Anthony D.L (2020) "I Don't Want Someone to Watch Me While I'm Working": Gendered Views of Facial Recognition Technology in Workplace Surveillance, *Journal of the Association for Information Science and Technology* 71(9) 1074-1088 <https://doi.org/10.1002/asi.24342>
- Urquhart, L, Lodge, T. and Crabtree, A. (2019) Demonstrably doing accountability in the Internet of Things, *International Journal of Law and Information Technology*, Volume 27, Issue 1, Spring 2019, Pages 1–27, <https://doi.org/10.1093/ijlit/eay015>
- Urquhart, L, McGarry, G. and Crabtree, A. (2022) Legal Provocations for HCI in the Design and Development of Trustworthy Autonomous Systems (Forthcoming).
- Urquhart, L, and Miranda, D. (2022) Policing faces: the Present and Future of Intelligent Facial Surveillance, *Information & Communications Technology Law*, 31:2, 194-219, DOI: 10.1080/13600834.2021.1994220

Evaluation of the impact of health technologies on life expectancy and Avoidable mortality in Asian and Europe countries: case study of Japan, South Korea, Lithuania and Luxembourg

Younes Karrouk¹, Driss Ezouine², Mohammed Hassani Zerrouk³ and Mario Arias Oliva⁴

¹Global Research Review, Tangier, Morocco

²Global Research Review, Rabat, Morocco

³Abdelmalek Essaadi university, Tangier, Morocco

⁴Complutense University, Madrid, Spain

younes.karrouk@gmail.com

Abstract. Health is undergoing a major revolution propelled by the increasingly personalized needs of patients. In this sense, technology for health will play a key role in the development of health, however, we certainly do not know the impact of technology on health indicators. This study aims to analyze how the Health technology could increase life expectancy and reduce avoidable mortality in Asiatic and European countries. Indicators related to avoidable mortality and life expectancy can provide a general “starting point” for evaluating the effectiveness of health technology in those countries: Japan, South corea, Lituatia and Luxembourg.

Keywords: Avoidable mortality, life expectancy, Health technology, Japan, South Korea, Lithuania, Luxembourg.

1 Introduction

This The evolution in health since the last century has made a spectacular link. The increase in lifespan is a good indicator of this. If it did not exceed 46.6 years on average in the world in the mid-1950s, it was 71.4 years in 2015 and could even reach 90 years for women in certain countries in 2030 (OCDE, 2017).

Hamilton, the most important biologist of the second half of the 20th century, concludes that age-related deterioration is inevitable because of all the specifics and the only radical genetic changes that extend lifespan: “After a few hundred years of draconian eugenic measures (...) the duration of human life could be lengthened very slightly (...) let's say it would go from 70 years to 75 years. » (Hamilton, 1996).

However, and contrary to what Hamilton suggests, the evolution of health technologies partly explains longevity. Health technologies therefore play an essential role in the functioning of a health system. Medical devices, in particular, are

indispensable for the prevention, diagnosis and treatment of diseases and the rehabilitation of patients. Therefore, health technologies have an essential role in the functioning of a health system. Medical devices, in particular, are indispensable for the prevention, diagnosis and treatment of diseases and the rehabilitation of patients. Therefore, the importance of health technology assessment has become an essential supporting tool for the core functions of an effective global health system (World Health Organization, 2011).

Two indicators of health status reflect the essential components of the length of life and its quality. Life expectancy is a fundamental indicator of the overall health status of a population; avoidable mortality draws attention to premature deaths that could have been avoided or whose causes could have been treated. In this context, it is necessary to know the correlation between the use of technology and health indicators such as: life expectancy and avoidable mortality.

2 Impact of medical technologies on health

Technology has influenced the way in which care is provided in many ways: by expanding the range of treatable pathologies and categories of patients; by the replacement of certain interventions, or better targeting; by intensifying the treatment of certain conditions; and by the modifications induced in the care procedures.

The diffusion of health technologies – along with other factors related to income level, reimbursement systems, medical culture and demographic evolution – has been one of the main drivers of the spectacular growth of health expenditure observed in OECD countries since the middle of the 20th century. Depending on the method used, it is estimated that between one-fifth and 70% of this increase is directly attributable to health technologies (Chernew & Newhouse, 2012).

According to a study of a subset of diseases with high prevalence in the United States, such as cardiovascular disease, cancer and infectious diseases, the additional cost induced by technologies in the past has been offset by results worth acceptable and is thereby legitimized. Overall, the resources employed in the development and adoption of health technologies have brought satisfactory gains in terms of human longevity and survival.

Considering health technologies through the prism of their “value”, it is possible to classify them into three groups (Chandra and Skinner, 2008, 2012). The first group are those that are effective and bring high benefit. Many of these are cheap, unsophisticated techniques that are widely used. Examples include aseptic techniques, vaccines, the combined use of beta-blockers and aspirin, cataract surgery and antiretroviral treatments used to combat HIV (Chandra & Skinner, 2012).

The second group includes technologies that are effective in certain indications but whose application tends to extend to populations and cases in which their clinical utility is less. Diminishing marginal benefit dilutes the value of these technologies. Included in this group are many diagnostic technologies (such as radiology and endoscopy) (OCDE, 2017).

The last category covers technologies, for which the evidence of a therapeutic interest is weak, or even non-existent, and which are equivalent, from a clinical point of view, to "observation" or to other less complex conservative interventions. . Many of these interventions are costly and carry the risk of iatrogenic injury. This category includes certain spinal surgery techniques, various diagnostic tools, such as liver function tests, and various devices, such as devices used to measure pulmonary arterial pressure (OCDE, 2017).

We then see that health technology has a visible impact on health, however, the studies carried out before do not give an idea of the impact of health technology according to the geographical aspect and the level of socio-economic development of countries.

As a result, we will proceed to the assessment of health technologies (belonging to the second category) in two geographical areas: Europe and Asia, by choosing two countries per continent (a developed and in developing countries).

3 Methods

The research method used in this research is a case study method with a qualitative approach, by measuring the impact of technology on health and more particularly on life expectancy and avoidable mortalities through prevention, we have taken all the materials making up the technological section, namely CT scanners, total, Gamma Cameras, total, Mammographs, total and Radiotherapy equipment total over a period of 20 years from 2000 to 2020.

We are going to make a comparison between two samples of two countries of Europe and two countries of Asia and we will see the evolution of these variables over time also we will look for a correlation and we will conclude on the impact technology on the other variables of the two samples.

4 Data

The database is taken from the official OECD website covering the period from 2000 to 2020: <https://stats.oecd.org/>.

Given the different interpretations to which the expressions below may lend themselves, they are defined as follows for the purposes of this article:

Medical technology (WHO): The definition of Medical technology is the application of organized knowledge and skills in the form of medicines, medical devices, vaccines, procedures and systems developed to solve a health problem and improve quality of life.

Avoidable mortality (OECD iLibrary, 2019): Avoidable mortality is defined as the causes of death that effective primary prevention and public health interventions (i.e. before the onset of diseases/ injuries, to reduce the incidence) would essentially prevent.

Life expectancy (Manton, 2007): Life expectancy is the average number of years a person in a population could expect to live after age x. It is the life table parameter most commonly used to compare the survival experience of populations.

We then present the data corresponding to the evolution of the three indicators: Medical technology (Table 1), Avoidable mortality (Table 2), Life expectancy (Table 3) during the period 2000-2020 for the four countries corresponding to our case of study: Japan, South Korea, Lithuania and Luxembourg.

Table 1. Evolution of Medical Technology between 2000-2020.

Year	Medical technology*					
	Japan	S.Korea	Asia	Lithuania	Luxemburg	Europe
2000	176,93	55,44	116,185	19	66,46	42,73
2001	176,93	55,47	116,2	20,23	65,67	42,95
2002	176,93	65,76	121,345	21,79	69,47	45,63
2003	188,1	73,28	130,69	22,2	79,7	50,95
2004	188,1	77,74	132,92	24,72	82,95	53,835
2005	188,1	81,49	134,795	25,89	83,85	54,87
2006	192,41	90,38	141,395	30,28	84,65	57,465
2007	192,41	106,11	149,26	25,37	85,42	55,395
2008	192,41	112,34	152,375	32,53	88	60,265
2009	202,14	117,46	159,8	36,98	86,39	61,685
2010	202,14	117,12	159,63	43,26	84,82	64,04
2011	202,14	123,17	162,655	49,85	79,09	64,47
2012	214,74	129,6	172,17	56,22	75,32	65,77
2013	214,74	132,69	173,715	58,5	71,76	65,13
2014	214,74	131,79	173,265	58,66	71,9	65,28
2015	223,95	133,35	178,65	57,84	66,72	62,28
2016	223,95	139,25	181,6	61,71	63,42	62,565
2017	223,95	142,89	183,42	62,57	62,05	62,31
2018	223,95	146,22	185,085	64,25	60,84	62,545
2019	223,95	151,02	187,485	68,36	62,9	65,63
2020	223,95	151,02	187,485	74,8	73,14	73,97

Source: <https://stats.oecd.org/>

* Calculated by summing the different variables that make up the variable Medical technology (Health Care Resources) SCANNER CT per million inhabitant + Radiation therapy equipment, total + Positron Emission Tomography (PET) scanners, total + Mammographs, total + Gamma cameras, total + Magnetic Resonance Imaging units, total.

Table 2. Evolution of Avoidable mortality /100000 inhabitant between 2000-2020.

Year	Avoidable mortality /100000 inhabitant					
	Japan	S.Korea	Asia	Lithuania	Luxemburg	Europe
2000	137	240	188,5	343	169	256
2001	134	230	182	365	161	263
2002	130	228	179	358	167	262,5
2003	128	222	175	357	162	259,5
2004	125	211	168	356	133	244,5
2005	123	199	161	380	135	257,5
2006	118	185	151,5	391	145	268
2007	116	176	146	400	140	270
2008	113	166	139,5	366	113	239,5
2009	110	161	135,5	326	125	225,5
2010	108	155	131,5	321	114	217,5
2011	113	146	129,5	307	112	209,5
2012	100	138	119	301	111	206
2013	97	131	114	295	108	201,5
2014	94	124	109	278	108	193
2015	90	117	103,5	276	98	187
2016	87	111	99	268	102	185
2017	85	104	94,5	245	95	170
2018	83	100	91,5	232	41	136,5
2019	84	97	90,5	226	41	133,5
2020	84	97	90,5	226	42	134

Source: <https://stats.oecd.org/>

Table 3. Evolution of life expectancy at age 65 (men and women) between 2000-2020

Year	Life expectancy at age 65 (men and women)					
	Japan	S.Korea	Asia	Lithuania	Luxemburg	Europe
2000	39,9	32,5	36,2	31,4	35,6	33,5
2001	40,5	32,9	36,7	31,1	35,7	33,4
2002	41	33,2	37,1	31	35,9	33,45
2003	41	34	37,5	31,3	34,2	32,75
2004	41,5	34,5	38	31,6	37	34,3
2005	41,3	35,1	38,2	31,1	37,1	34,1
2006	41,9	35,6	38,75	31,2	37,3	34,25
2007	42,2	36,2	39,2	31,3	36,7	34
2008	42,2	37	39,6	32	38,4	35,2
2009	42,9	37,8	40,35	32,4	39	35,7
2010	42,5	38	40,25	32,6	38,9	35,75
2011	42,4	38,6	40,5	33,2	39,4	36,3
2012	42,7	38,7	40,7	33,3	39,8	36,55
2013	43,1	39,5	41,3	33,3	41	37,15
2014	43,5	40,2	41,85	33,8	41,1	37,45
2015	43,6	40,6	42,1	33,3	40,7	37
2016	44	41	42,5	33,6	41,6	37,6
2017	44	41,3	42,65	33,8	40,3	37,05
2018	44,2	41,5	42,85	34,2	40,9	37,55
2019	44,4	42,5	43,45	34,8	41,6	38,2
2020	44,4	42,5	43,45	32,7	40,4	36,55

Source: <https://stats.oecd.org/>

5 Results and discussion

The study results revealed that, in Japan , a growth of +26.57% devoted to the adoption of health technology has generated a significant drop in avoidable mortality in the order of -38.68% and has increased life expectancy of the order of + 11.27% during the study period 2000-2020.

On the other hand, in South Korea, a growth of +172.4% devoted to the adoption of health technology has generated a significant drop in avoidable mortality of the order of -59.58% and has increased life expectancy of the order of + 30.76% during the study period 2000-2020.

Regarding Luxemburg, a growth of +10.05 % devoted to the adoption of health technology has generated a significant drop in avoidable mortality in the order of 75.14% and has increased life expectancy of the order of + 13.48% during the study period 2000-2020.

Finally, in Lithuania, a growth of +293.68% devoted to the adoption of health technology has generated a significant drop in avoidable mortality in the order of 34.11% and has increased life expectancy of the order of + 4.14% during the study period 2000-2020.

By comparing the consolidated indicators of Europe and Asia, it can be seen that an increase in the adoption of health technology from +73.11 % in Europe, has generated a significant drop in avoidable mortality in the order of -47.65% and has increased life expectancy of the order of + 9.10%.

On the other hand, in Asia, we can see that with +61.36 % increasing health technology, it has generated a significant drop in avoidable mortality in the order of - 51.98% and has increased life expectancy of the order of + 20.02%. That is mean, with a less engagement in health technology, Asiatic countries is more performing than the European countries regarding life expectancy and avoidable mortality indicators.

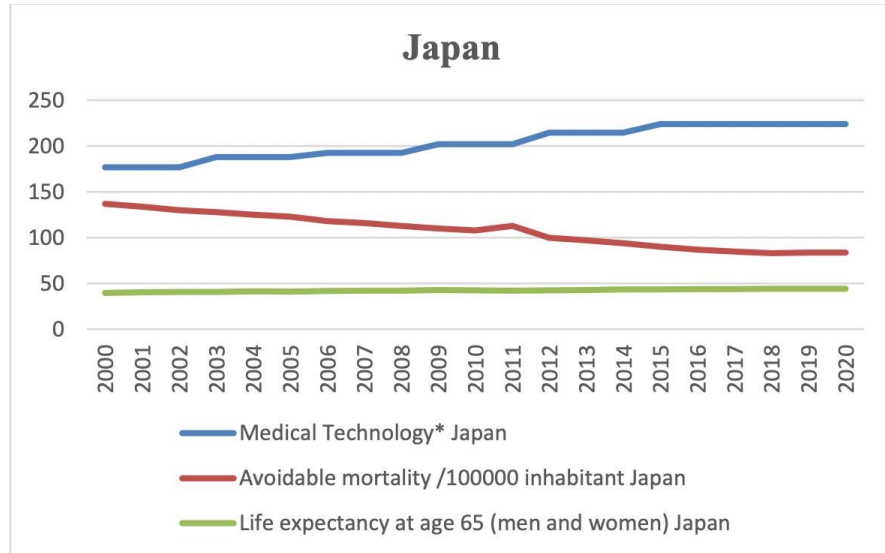


Fig. 1. Evolution of life expectancy and avoidable mortality compared to medical technology in Japan.

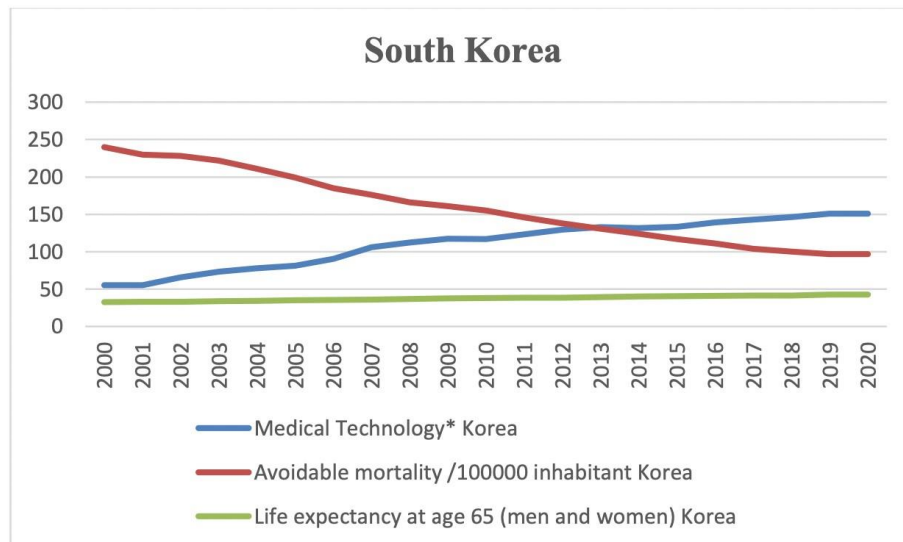


Fig. 2. Evolution of life expectancy and avoidable mortality compared to medical technology in South Korea

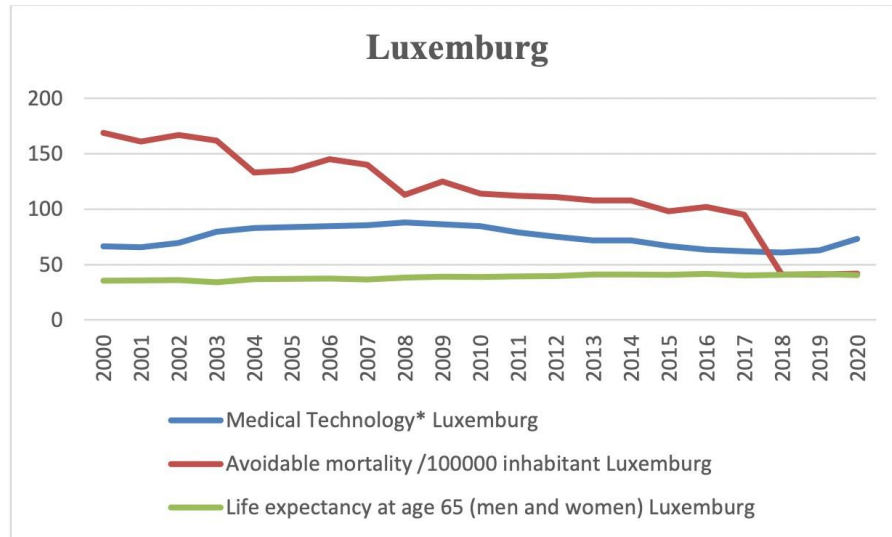


Fig. 3. Evolution of life expectancy and avoidable mortality compared to medical technology in Luxemburg

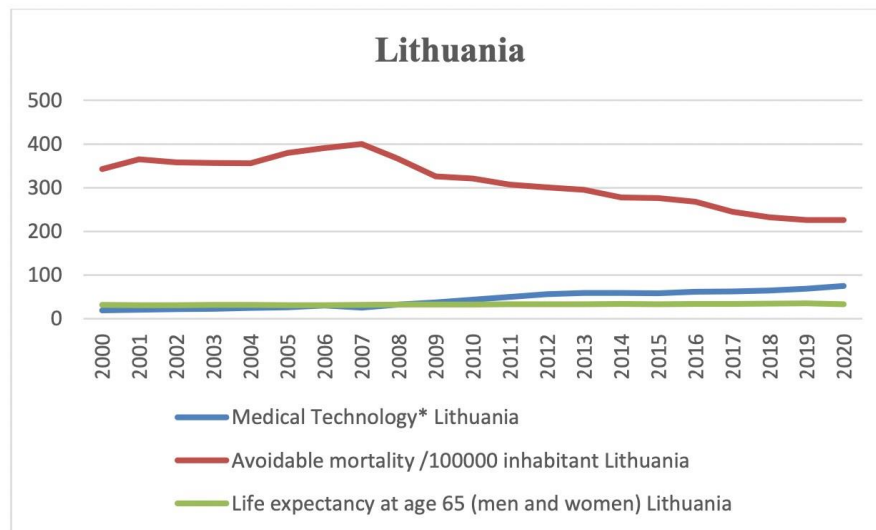


Fig. 4. Evolution of life expectancy and avoidable mortality compared to medical technology in Lithuania

6 Conclusions

In conclusion, the present study revealed that health technology has a positively significant impact on health. However, this impact varies from one country to another, knowing that with a minor deployment of technology on the part of Luxembourg compared to the other countries covered by our study, it had the best results concerning avoidable mortality and life expectancy. The same observation was observed when comparing Asia and Europe (Asia is more efficient than Europe).

These results push us to conduct other studies necessary to reveal other parameters related to the use of health technology and which could explain the considerable differences in results between the countries and continents covered by our study.

Acknowledgments

This research was supported by Telefonica and its Telefonica Chair on Smart Cities of the Universitat Rovira i Virgili and the Universitat de Barcelona (project number 42.DB.00.18.00)

References

- Chandra, A. and J. Skinner (2012). “Technology Growth and Expenditure Growth in Health Care”, *Journal of Economic Literature*, Vol. 50, No. 3, pp. 645–680.
- Chernew, M. and Newhouse, J. (2012). “Health Care Spending and Growth”, *Handbook of Health Economics* Vol. 2, Elsevier.
- Hamilton, W. D. (1996). *Narrow Roads of Gene Land*, Oxford University Press, vol. 1.
- Manton, K.G. (2007). in *Encyclopedia of Gerontology* (Second Edition)
- OCDE. (2017). “New Health technologies: managing access, value and sustainability “, <https://www.oecd.org/health/health-systems/Les-nouvelles-technologies-desant%C3%A9-S>
- OCDE. (2017). *New Health technologies: Managing access, value and sustainability*, pp 10-11.
- OECD iLibrary (2019). *Avoidable mortality (preventable and treatable)*. <https://www.oecdilibrary.org/sites/3b4fdbf2en/index.html?itemId=/content/component/3b4fdbf2en#:~:text=Based%20on%20the%202019%20OECD,injuries%2C%20to%20reduce%20incidence>.
- WHO, <https://www.euro.who.int/en/health-topics/Health-systems/health-technologies-and-medicines/policy-areas/health-technology-assessment>
- World Health Organization. (2011). *First WHO global forum on medical devices: context, outcomes, and future actions*. World Health Organization. http://www.who.int/medical_devices/gfmd_report_final.pdf.

Release the robot dogs:

Teaching with professional codes of ethics

Stacy A. Doore and Azalea Yunus

Department of Computer Science, Colby College, Waterville, ME, USA

sadoore@colby.edu azalea.yunus@colby.edu

Abstract. There are many barriers to embedding responsible computing instruction into technical computing courses including faculty expertise, time constraints, and competing pedagogical approaches. The misalignment between computing curricula competencies and responsible computing may also contribute to the reluctance of faculty to incorporate computing ethics into technical courses. This may be because the debate about the efficacy of teaching professional responsibility through the codes of ethics continues to focus primarily on the direct teaching of the revised principles and their application to contrived case studies. Instead, we propose an approach that situates professional codes of ethics within a set of narratives or stories to be considered in relation to the real-world deployment of an emerging technology, agile industrial robots. We present the *Computing Ethics Narratives* framework which offers the reading of professional ethics codes as just one of many lenses that can help students to internalize and act upon their responsibilities as both creators of future technologies and as members of multiple stakeholder groups and diverse communities.

Keywords: codes of ethics, agile robots, computer science education

1 Introduction

As the ACM/ IEEE computing curriculum has evolved over time, so have the ways in which we assess how our students are able to apply what they have learned. Part of this curriculum evolution has resulted in a move away from learning outcomes to instead creating and acquiring competencies. Competencies are defined as a collection of knowledge (what), skills (how), and dispositions (why) that are required to be a successful professional in the interrelated computing disciplines (ACM, 2020). The ACM Computing Curricula 2020 report outlines a competency-based framework that moves away from a knowledge focused model to a model that include skills and behaviors in the assessment of student learning (ACM, 2020). However, there appears to be a disconnect between the draft lists of competency statements in the CC2020

and the principles adopted in the most recent versions of the ACM and IEEE professional codes of ethics and conduct.

While academic programs are not under any obligation to shape their computing programs based on ACM curriculum recommendations, during CSAB/ABET accreditation reviews CS programs are required to demonstrate their ability to produce students that can: “Recognize professional responsibilities and make informed judgments in computing practice based on legal and ethical principles.”. Likewise, their curriculum must contain topics teaching about “Local and global impacts of computing solutions on individuals, organizations, and society.” (ABET, 2021). The CSAB/ABET requirements are broad in scope yet vague in the specifics about how programs can meet these requirements. The latest ACM/IEEE curriculum is very specific about the types of technical competencies that students should be able to demonstrate, however clear competency statements about knowledge, skills, and dispositions related to professional conduct and responsible decision-making are currently absent from almost every topic area.

We argue that the principles from the ACM and IEEE professional codes of ethics should be more explicitly reflected as core competency statements in any of the ACM/IEEE CC 202X curriculum reports. In recent professional computer science education meetings, the number of papers, birds of a feather (BoF) sessions, and symposia focusing on the teaching of computing ethics at the undergraduate level has increased dramatically (ACM SIGCSE, 2022). There are also recommendations from a recent report from the National Academies of Sciences (NAS) *Fostering Responsible Computer Science Research* that calls for universities and colleges to “enhance: 1) teaching and learning in computer science and engineering, information science, and other computing-related fields to ensure that the next generation is better equipped to understand and address ethical issues and potential societal impacts of computing and (2) humanities and social and behavioral science education to ensure that students in those fields are equipped to participate in informed discussions of potential impacts of computing research and technologies.”(NAS, 2022, p. 78).

This paper provides a framework and a robust set of materials for developing and evaluating student competencies to analyze ethical issues and the potential impact of real-world scenarios through the lens of current professional code of ethics. We describe a pedagogical framework, a narrative repository, and a variety of strategies for teaching with the professional codes of ethics that are illustrated through a module on the use of agile robots in public spaces. The module is designed to be taught within a technical course in which decision-making scenarios would be most relevant, such as an introductory robotics or embedded systems course, but could also be implemented in a single computing ethics course. We also provide suggestions for the qualitative assessment of student application of professional codes of ethics as an essential competency in their technical training.

2 Related Work

There is a growing body literature on the barriers to embedding comprehensive and evidence-based responsible computing instruction (including professional codes of ethics and conduct) into undergraduate computing courses. The first is the prevailing model of using a single stand-alone course to introduce a broad scope of ethical issues in computing to students who may have little background in social, cultural, or philosophical concepts to evaluate their own responsibilities as creators of technology. This approach is convenient because these types of courses can efficiently meet the requirements of accreditation and have an abundance of textbooks available to guide a prescribed set of instruction (Johnson, 2009; Tavani, 2015; Quinn, 2020). However, as Greer and Wolf (2020) have discussed the single course approach can be problematic in the message that it sends to both faculty and students that the course is an add-on that is necessary for only for compliance purposes as opposed to understanding that ethical issues and social impact are as fundamental to disciplinary training as instruction on algorithms, system architecture, and networks.

The alternative approach is to embed ethical and social aspects of computing into all areas of the computing curricula so that student competencies can develop from the earliest exposure of learning to program to advanced levels in technology design and development. This approach can be observed in several projects that integrate social with technical instruction and are specifically designed to target students' conceptualizations of technology and their privileged role as creators (Peck, 2017; Skirpan et al, 2018; Grosz et al, 2019; Saltz et al, 2019, Doore et al, 2020). In some cases, the integrated approach to computing ethics education can be the result of a single instructor (Henderson, 2019; Peck, 2021) or to a collective group of CS and interdisciplinary teams who produce materials and guidelines for embedding ethical and social impact content into the full spectrum of computing courses (Mozilla Foundation, 2021).

Recent studies have focused on how ethical and social responsibility are understood and used in instruction by computer science faculty. For example, Fiesler, Garrett, and Beard (2020) conducted an analysis of syllabi of stand-alone computing ethics courses to illustrate the primary topics, instructional approaches, and learning objectives. Greer and Wolf (2020) conducted qualitative interviews with CS department chairs and faculty to better understand barriers to moving from a standalone course approach to an embedded ethics approach. Their findings suggest that technical course faculty often omit discussions of computing ethics because of a

lack of expertise, perceived time constraints, or lack of appropriate teaching materials. There was also a considerable deference to authority (i.e., ABET accreditors, NAE, NAS) given as the reason for not devoting more department time and resources to helping faculty acquire the expertise needed to embed responsible computing instruction throughout the technical curriculum. This is belief is ironic as this has historically been a requirement for accredited departments demonstrate the impact of their professional ethics instruction in their course materials, assessments, and alumni reporting.

Studies investigating the efficacy of integrating professional codes of ethics into the CS curriculum have reported mixed results. Several empirical and qualitative design studies found that instructional time and resources spent applying the ACM codes of ethics as applied to ethical case studies had little to no impact on student outcomes (Peslak, 2007; McNamara, Smith, Murphy-Hill, 2018; Hedayati-Mehdiabadi, 2022). While the studies found the codes of codes themselves were not a significant factor in shaping computing students' ethical decision making when considering example case studies, there were several positive influence factors that suggested responsible computing critical thinking. These included when students were able to 1) relate to the context of a real-world story, 2) convey concern for stakeholders (beyond the developers) who might be impacted or harmed, 3) identify argument fallacies, 4) articulate a sense of professional responsibility due to technical knowledge, and 5) evaluate the significant level of harm presented by the ethical issue in the scenario (Hedayati-Mehdiabadi, 2022). Based on encouraging results from more interdisciplinary approaches to teaching responsible computing principles that focus on expansive and comprehensive approaches to computing ethics instruction (Burton, Goldsmith, and Mattei, 2018), we believe that codes of ethics can be effectively incorporated into the CS curriculum, but they must be used as a framing lens before during, and after to support the positive influence factors of ethical decision-making. What follows is a framework for using professional codes of ethics as one of many types of narratives with which to explore and internalize complex responsible computing principles and behaviors.

3 Computing Ethics Narratives

The Computing Ethics Narratives project (CEN) is an interdisciplinary digital open educational repository (OER) that combines work in computer science, philosophy, digital and computational studies, and cinema studies to curate and design curricular resources to integrate computing ethics lessons into undergraduate computer science (CS) education (Doore, Nascimento, and Cooper, 2021). The goal of the CEN project is to make computing ethics topics accessible to CS instructors through the creation of a searchable repository of *narratives* organized by social themes and technologies. The repository also includes a set of teaching *modules* featuring the wide variety of curated narratives types. The pedagogical approach is to help students develop a set of critical thinking competencies to better prepare them to be responsible creators of future technologies. The CEN framework places a priority on providing curricular resources focused on the competencies that define responsible professional behavior and the development of *ethical sensitivity or sensibility* (Bebeau, 2014). This includes students being able to not only identify potential ethical problems and potential harms in context but also being able to examine the potential for conflicting perspectives and priorities of multiple stakeholders in any given situation. The framework adopts Ricoeur's emphasis on narratives for ethical deliberation (Ricoeur, 1991; Bullock, Nascimento, Doore, 2021).

The CEN framework defines *narratives* as historical or fictional stories and perspectives involving technologies as a primary character of the plot. Individual narratives are classified by at least one or more ethical themes and technology areas. Ethical themes are organized around a taxonomy based loosely on topics found on the Berkman Klein Center for Internet & Society (Berkman Klein Center, 2022). Narratives also belong to one or more technology areas that are based on the ACM Computing Classification System (CCS) (ACM, 2012). The curated keyword searchable repository currently contains approximately 500 narratives. The repository also contains teaching modules that demonstrate how a narrative approach might be used to integrate computing ethics lessons into the CS curriculum. A *module* is a pedagogically cohesive set of narratives and instructional activities designed to enhance students' ethical sensibility on evaluating issues associated with technologies. Each module is based on a set of student competencies building knowledge, skills, and dispositions evaluating ethical issues related to a societal theme or a technology area. Each activity within the module uses several narratives curated from the online repository, which contains (e.g., film clips, short stories, news articles, blog posts, series episode clips, podcasts, etc.). To facilitate faculty using the CEN repository, the narratives provide a link to the original source, a short summary, and discussion prompts (Figures 1-3).

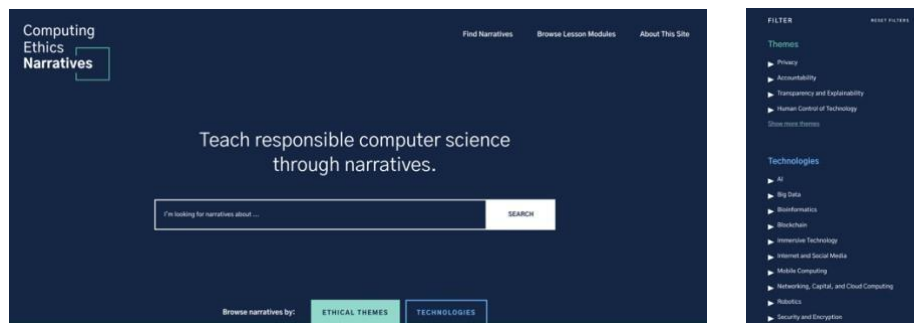


Figure 1. CEN narratives are searchable by ethical theme and technology areas.

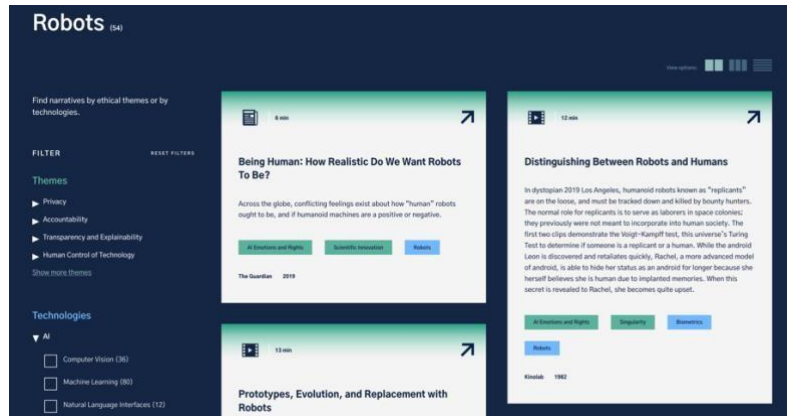


Figure 2. CEN narratives under technology area ‘robots’ showing summary of narrative.

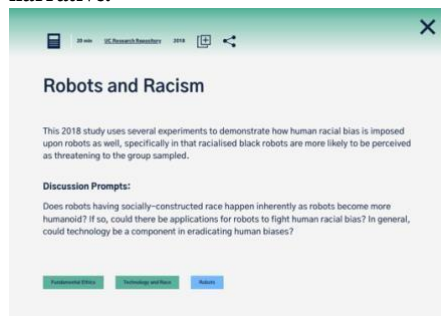


Figure 3. Narrative discussion prompts for individual reflection or group discussion.

The narratives and discussion prompts are flexible enough to be used as a part of whole class discussions, individual reflections, small lab group meetings, or remote crowdsourced responses on an LMS platform. There are filters to allow faculty to search by open access narratives as well as sources that might require subscription services through their institutional libraries. The CEN repository architecture is a database that organizes narratives and is integrated with an existing online film and television series clip repository, Kinolab (Cooper, Nascimento, and Francis, 2021) (Figure 4) that is compliant with current regulations on the use of copyrighted materials used for educational purposes (U.S. Congress, 1998).

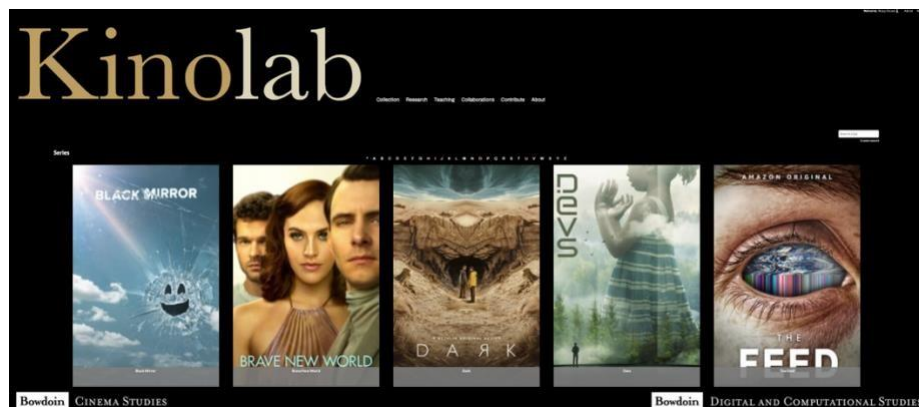


Figure 4. CEN and Kinolab connection permits viewing access to annotated film/series.

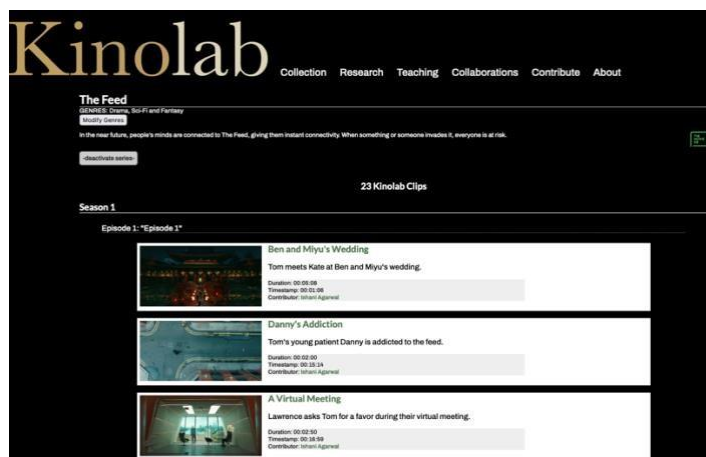


Figure 5. Kinolab clip annotations provide scene summary, duration, and timestamp.

The CEN narratives provide a framework to build modules around a central ethical theme or technology area. Each module provides a summary for instructors, the module goal and the related themes or technologies (Figure 6).

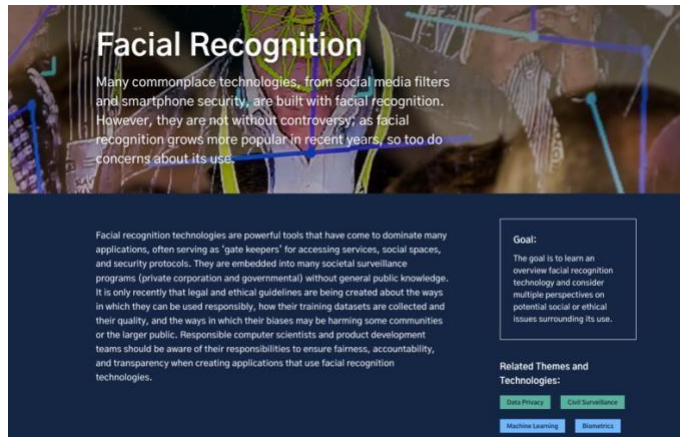


Figure 6. CEN module on Facial Recognition Systems.

In this framework, professional codes of ethics are narratives to be accessed along with the other types of teaching resources that provide explicit connections between professional responsibilities within either real-world or imaginative fiction contexts. They are not treated as narratives that are separate from students' consideration of technology as a plot character, their own role as a technology creator, or the potential impact upon a wide variety of actors (i.e. technology stakeholders) beyond their own perspective. There are a variety of ways to accomplish this integration of responsible computing codes of ethics into CEN modules. In the next section, we illustrate how professional codes of ethics serves act as narratives to considered in the real world deployment of an emerging technology into a diverse campus community.

4 Agile Robots in Public Spaces Module

Every CEN module begins with an introduction to the topic placing it in the context of the course or department objectives. The *Agile Robots in Public Spaces* module, might fit well in with course/department objectives focused on 1) students learning how to analyze the potential risk and impact of emerging technologies on diverse communities of stakeholders, 2) students being able to apply existing professional codes of ethics to real world scenarios, and 3) students learning how to create responsible computing policies and processes that reflect multiple stakeholder perspectives. In this module example, the goal was to have students apply both the ACM code of ethics and a responsible computing framework (Friedman and Hendry, 2020) to develop a set of responsible use policies for a recently acquired emerging technology on campus (i.e., a Boston Dynamic agile robot). The primary activities involved using these responsible computing resources to investigate campus stakeholder roles, values, priorities, and perspectives, and examine ethical/legal/professional responsibilities of faculty and students when conducting human-robot interaction research in public spaces on our campus.

4.1 Pre-module assessment

The *Agile Robots in Public Spaces* module begins with a pre-assessment of how much students knew about professional codes of ethics and about agile robots. This assessment is typically a link on the repository module page containing a brief online survey to collect students' pre-existing ideas about the module topic. For this module, we decided to pilot a survey that we would use as a tool in ongoing research on human-robot interactions within our own campus community. The survey was based on the new *General Attitudes Towards Robots Scale (GAToRS)* (Koverola, Kunnari, Sundvall, and Laakasuo, 2022) and we added several questions to the end of the survey that specifically asked students which principles in computing professional codes of ethics (ACM/IEEE) would be helpful in developing ethical guidelines and best practices for deploying an agile robot in public spaces. This was an in-class activity without the use of codes that also served as a mid-course assessment evaluating to what extent students had internalized earlier lessons and were able to recall and apply the ACM principles to different case studies in the course. The survey helped us assess 1) their pre-existing perceptions around potential benefits and harms of agile robots, 2) their ability to recall and apply the codes of ethics to a novel situation, and 3) their pre-module understanding of why creating responsible use policies was important enough of a topic to discuss and work through. The survey results suggested that approximately half of the students had positive personal and societal attitudes towards agile robots in general, however there were just as many students that expressed either uncertainty or strong negative feelings about these robots in both the personal and the societal impact.

4.2 Module instructional activities

We began by assigning several narratives explaining agile robot evolution and technical capabilities, highlighting their application in the Defense Advanced Research Projects Agency (DARPA) 2020 Subterranean Challenge (Bouman et al., 2020). Students learn more about the robot's perception, computation, and autonomy through this narrative. Next, students were provided a set of narratives about the adoption of agile robots by law enforcement agencies and their potential for use in public spaces. Boston Dynamics was the first company to lease their agile quadruped robot, Spot, to the Massachusetts State Police bomb squad for a period of three months, during which state police report that it was used in two incidents in addition to testing. The American Civil Liberties (ACLU) chapter of Massachusetts raised two issues with the use of Spot by the Massachusetts state police. The first involved the potential for weaponization and the second was a lack of transparency and public documentation by the state police of their intended use and internal rules for its use (Jarmanning, 2019). We also examine Boston Dynamics' explicit ethical use clause contained in its buyer's agreement specifying Spot remain a general-purpose robot without weaponization (Boston Dynamics, 2022). This is in contrast to similar robotics companies who have designed and marketed to military agencies. Ghost

Robotics executives stated that they hope to see their robots used in active combat zones assisting with tasks like scouting, targeting, and explosive defusing as early as 2022 (Brown, 2020).

The second set of narratives provided were examples of speculative narratives about the technology and its potential uses. In 2017, the dystopian science fiction television series featured violent robot dogs hunting down a band of thieves in the episode “Metalhead” and the show’s creator explicitly cited Spot as an inspiration (Herman, 2017). In February 2021, the art collective MSCHF equipped Spot with a paintball gun and invited users to control Spot and its paintball gun over the internet as part of a collective art installation called “Spot’s Rampage”. Representatives from Boston Dynamics condemned the experimental art piece on the grounds that it depicted Spot as a violent, harmful, and intimidating robot (Knight, 2021). The juxtaposition of the two narratives and the Boston Dynamic ethical user policy highlights for students how difficult it is to ensure that a technology will be used exclusively in ways that are in line with the creator’s intent.

The narrative set also included a recorded talk about ethical issues in human-robot interactions (Darling, 2015), the ACM/IEEE professional codes of ethics, and previously covered materials on Value Sensitive Design principles (Freidman and Hendry, 2020) to help students continue to use them as references in thinking critically about different perspectives and responsibilities to stakeholders in the design and deployment of emergent technologies.

The first in-class activity was to help students identify potential groups of stakeholders in the campus community who might be impacted by the deployment of an agile robot. The class identified five primary stakeholder groups: *Students*, *Faculty*, *Staff*, *Visitors*, and *Administrators*. Students were then randomly assigned to each of the five groups and told to consider what their stakeholder group’s perspectives, values, and priorities might be in the deployment of an agile robot in our campus community. They were also asked to define key concepts/terms that might need to be defined and the situation or context in which their stakeholders might encounter an agile robot on campus. They were told that there might be competing values and priorities within their own group and that should be reflected in the notes of their discussion. Finally, each group was assigned several additional narratives to consider as they began to draft a list of recommendations and guidelines that represented their stakeholder group role, values, and perspectives.

In the next class period, the student groups were given time to refine their draft recommendations and instructed to develop a set of 3-5 slides that represented their stakeholder role, values, concerns, and recommendations for responsible use policies and processes for the introduction of this emergent technology on campus. They were also asked as a group to step outside of their stakeholder roles to reconsider what principles from the ACM/IEEE codes of ethics might apply best to this use context that would support their policy recommendations. At the end of class, each group then presented their stakeholder perspectives, their recommendations, and how these recommendations aligned with ACM/IEEE codes of ethics. Despite the differences in their various stakeholder concerns and values, there was much common ground in the types of policies and processes their stakeholder groups had identified. The class

voted on the recommendations from each stakeholder group that they felt should be accepted or further refined to comprise a final set of responsible use policies applied to our lab's agile robot in research and campus outreach in public spaces.

4.3 Teaching with codes of ethics

What follows is a brief example of the student-led analysis of this issue and the application of a set of the ACM/IEEE code of ethics to this topic. The analysis activity is designed to be flexible enough for faculty to adapt it to their own pedagogical style and student engagement preferences (e.g., small group discussion, individual written analysis, role playing of stakeholder groups, etc.). The key is the activity must be active. The analysis activity can lend itself to small or large class sizes and might include lectures or student reading and discussion on the evolution of agile robot development, technical specifications of current types of agile robots, and the potential applications of these types of robots. However, the core narratives that students should be familiar with for any of the CEN modules is the ACM and IEEE codes of ethics. While they are not the central focus of the modules, they do act as a lens through which students can examine themes or technologies and consider values, principles, and behaviors of responsible computing professionals.

Students were first asked, "Are there principles in the ACM or IEEE codes of ethics that might be helpful in developing best practices and ethical guidelines for deploying an agile robot in public spaces on campus?" The response was 100% "yes". The next question asked, "How might the ACM /IEEE codes of ethics guide the lab in the potential use of an agile robot in research and teaching on our campus?". Three Articles from the ACM code of ethics were cited by the majority of the students. Students most frequently identified Article 1.2 *Avoid Harm* as the primary principle that should help the lab identify any potential harm that might be caused by an agile robot on campus. This principle commits technologists to prevent "unjustified physical or mental injury, unjustified destruction or disclosure of information, and unjustified damage to property, reputation, and the environment" (ACM, 2018). The students cited the narratives that described the controversy and public criticism of the NYPD's use of Spot in 2021 (Bushwick, 2021) and applied that a sense of harm that might be felt by members of the campus community who had concerns about Spot being deployed in public spaces. Next, they cited Article 1.1 *Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing* emphasizing that with a diverse group of stakeholders on campus it was important to get input about the rules that would be developed for use of this emerging technology. Finally, Article 1.6 *Respect Privacy* was the third principle cited in the application of the code of ethics survey. The students expressed privacy concerns about the number of cameras and sensors on the agile robot that could be used to record and store information about who had interacted with or walked near the robot while it was being used in public spaces.

For the final activity, stakeholder groups reported their recommendations for establishing lab policies that would regulate human-robot interactions and operations and how these recommendations aligned with the identified principles from ACM

code of ethics. The groups found broad consensus around most of the recommended rules, and where there was conflict, negotiations between groups took place to ensure all stakeholder concerns were recognized and addressed. The draft recommendation statements are summarized below:

Human-Robot Interactions and Operation Policies (Draft):

- Create a page on the lab website for campus community members to learn about research mission and goals about the use of Spot in assistive technology research. This page would have answers to FAQ and an anonymous form link to provide feedback about the use of the robot on campus. (ACM 1.1)
- Spot policies should be reviewed and brought into alignment with other campus policies on emerging technologies such as drones, recordings, etc. (ACM 1.1)
- There should be ‘Spot Crossing’ signs letting community members know this is an area they might encounter Spot while the lab is conducting research or demonstrations. (ACM 1.1)
- Community members should have the opportunity to ‘opt-out’ of interacting with Spot in a campus public space, meaning if they indicate they are afraid or uncomfortable, the operators should stop Spot and let them pass. (ACM1.1)
- There should be a ‘Spot’ training program that requires operators to demonstrate the safe operation of the agile robot in public spaces. (ACM 1.2)
- Spot operators should be required to work in teams of 2 or more and should have to use an operational checklist to ensure the robot is in safe working order before using it. (ACM 1.2)
- Any use of Spot in autonomous navigation mode should have two operators in close proximity in case there are problems. (ACM 1.2)
- There should be a clear expectation and lab mechanism for operators to report any unexpected incidents or malfunctions while operating Spot. (ACM 1.2)

The students recognized that to conduct research with Spot, there would be a need for some data collection and storage from the robot’s many sensors and cameras; however, there should be clear policies on the types of data that needed to be collected, the way they were stored to respect the privacy of members of the campus community.

Agile Robot Data Collection Policies (Draft):

- Spot should not have contact with children under 3 and no data should be collected or stored on children under the age of 13 based on COPPA. (ACM 1.6)

- All research data collected using Spot should be reviewed and receive approval from the IRB and should be collected using informed consent processes. (ACM 1.6)
- All data collected using Spot cameras should be processed to remove identifying information (e.g., faces) prior to storage in the research database. (ACM 1.6)

While they felt that developing a set of guidelines was a good first step, the students advocated it would also be important to conduct focus groups and/or administer surveys to representative members of these stakeholder groups to refine and validate the draft guidelines in the next academic year.

This is only one possible way to assess the ability of students to work with a code of ethics to develop responsible computing best practices and policies based on the principles in the code and a set of accompanying narratives about the use of an emerging technology. This assessment of module impact on the development of student competencies applying professional codes of ethics can happen in a variety of ways depending on the size of the class, the availability of course support, if it is administered through an in-person or online format, and the level of details that are appropriate for the intended purpose. Although this is not an empirical study of the impact of teaching with professional codes of ethics, it does suggest that when ethical codes are perceived as part of the complex set of narratives must consider navigate when evaluating their social and professional responsibilities as creators of emerging technologies, professional codes are not abstract principles. Instead, they help to provide a starting place and familiar framework to establish and internalize responsible computing behaviors and practices.

5 Conclusions and Future Work

To become responsible computing professionals, CS students require more authentic opportunities to consider the social impact of the technologies they will create. While professional codes of ethics can provide broad and proactive guidance on the responsibilities of computing professionals, helping students to internalize these principles into competencies so they can how, when, and why to act upon them is a difficult and ongoing challenge. This is especially true because the vast majority of CS faculty have not integrated these principles into own their teaching and assessment practices. This paper describes a rationale, pedagogical framework, and an open educational repository to provide faculty with opportunities to weave compelling narratives into their technical courses. This CEN framework teaches with set of narratives to bring the principles of professional responsibilities to life within the context of complex storytelling where emerging technologies and the technologists who create them, are characters in a larger narrative. Students are asked to take on multiple perspectives on the potential impact of these technologies and to consider what their responsibilities are to the larger society in creating future worlds.

Acknowledgements

The authors would like to extend their gratitude to Fernando Nascimento, Allison Cooper (CEN collaborators), members of the Harvard/MIT Cyberethics Work in Progress Group - Trystan Goetze, Kevin Milla, Jenna Donohue, Kiran Bhardwaj who provided feedback on an early version of this paper, and members of the Colby INSITE lab who help to make Spot a responsible research partner on campus.

References

- Accreditation Board for Engineering and Technology (ABET). (2021). Criteria for accrediting computing programs 2020-21. <https://www.abet.org/accreditation/accreditationcriteria/criteria-for-accrediting-computing-programs-2020-2021/>.
- Association for Computing Machinery (ACM). (2012). The 2012 ACM computing classification system. <https://www.acm.org/publications/class-2012>.
- Association for Computing Machinery (ACM). (2018). ACM code of ethics and professional conduct. <https://www.acm.org/code-of-ethics>.
- Association for Computing Machinery and IEEE Computer Society. (2021). Computing Curricula 2020: Paradigms for Global Computing Education (CC2020). DOI: 10.1145/3467967
- Association of Computing Machinery. (2022). 53rd Annual ACM Technical Symposium on Computer Science Education. March 2-5, 2022. Providence, RI. <https://sigcse2022.sigcse.org/schedule/>
- Bebeau, M. J. (2014). An evidence-based guide for ethics instruction. *Journal of Microbiology & Biology Education*, 15(2), 124-129.
- Berkman Klein Center. (2020). Topics. <https://cyber.harvard.edu/topics>
- Boston Dynamics, (2022). Boston Dynamics ethical principles. <https://www.bostondynamics.com/ethics>.
- Bouman, A., Ginting, M.F., Alatur, N., Palieri, B., Fan, D.D., Touma, T., et al. (2020). Autonomous Spot: Long-range autonomous exploration of extreme environments with legged locomotion. In *International Conference on Intelligent Robots and Systems (IROS)* pp. 2518-2525. IEEE. <https://10.0.4.85/IROS45743.2020.9341361>.
- Brown, D. (2020, December 7). The U.S. is entering a robotic future. Washington Post. <https://www.washingtonpost.com/technology/2020/12/07/robot-dogs-patrol-tyndall/>.
- Bullock, B.B., Nascimento, F.L., Doore, S.A. (March 2021). Computing ethics narratives: Teaching computing ethics and the impact of predictive algorithms. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education (SIGCSE '21)* pp. 1020-1026. ACM. <https://dl.acm.org/doi/10.1145/3408877.3432468>.
- Burton, E., Goldsmith, J., & Mattei, N. (2018). How to teach computer ethics through science fiction. *Communications of the ACM*, 61(8), 54-64.

- Bushwick, S. (2021, May 7). The NYPD's robot dog was a really bad idea: Here's what went wrong. *Scientific American*. <https://www.scientificamerican.com/article/the-nypdsrobot-dog-was-a-really-bad-idea-heres-what-went-wrong/>.
- Computing Curricula (CC). (2020). CC2020: Paradigms for global computing education. <https://www.acm.org/binaries/content/assets/education/curricularecommendations/cc2020.pdf>.
- Congress, U. S. (1998). Digital millennium copyright act. *Public Law*, 105(304), 112.
- Darling, K. (2015). Ethical issues in human-robot interactions. *The Conference*. <https://www.youtube.com/watch?v=m3gp4LFgPX0>
- Doore, S. A., Nascimento, F., & Cooper A. (2021). Computing Ethics Narratives. <https://www.computingnarratives.com/>
- Cooper, A. Nascimento, F., & Francis, D. (2021). Kinolab. <https://kinolab.org/>.
- Cooper, A., Nascimento, F., & Francis, D. (2021). Exploring Film Language with a Digital Analysis Tool: The Case of Kinolab. *DHQ: Digital Humanities Quarterly*, 15(1), 129.
- Doore, S. A., Fiesler, C., Kirkpatrick, M. S., Peck, E., & Sahami, M. (2020, February). Assignments that blend ethics and technology. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (pp. 475-476).
- Friedman, B. & Hendry, D.G. (2020). *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press.
- Fiesler, C., Garrett, N., & Beard, N. (2020, February). What do we teach when we teach tech ethics? A syllabi analysis. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (pp. 289-295).
- Gotterbarn, D.W., Brinkman, B., Flick, C., Kirkpatrick, M.S., Miller, K., Vasansky, K., Wolf, M.J. (2018). ACM code of ethics and professional conduct.
- Greer, C., & Wolf, M. J. (2020). Overcoming Barriers to Including Ethics and Social Responsibility in Computing Courses. In *Societal Challenges in the Smart Society* (pp. 131-144). Universidad de La Rioja.
- Grosz, B.J., Grant, D.G., Vrendenburgh, V., Behrends, J., Hu, L., Simmons, A. & Waldo, J. (2019). Embedded EthiCS: Integrating ethics across CS education. *Communications of the ACM*. 62(8), 54-61.
- Hedayati-Mehdiabadi, A. (2022). How do computer science students make decisions in ethical situations? Implications for teaching computing ethics based on a grounded theory study. *ACM Transactions on Computing Education*. <https://dl.acm.org/doi/10.1145/3483841>.
- Henderson, T. (2019, January). Teaching Data Ethics: We're going to ethics the heck out of this. In *Proceedings of the 3rd Conference on Computing Education Practice* (pp. 1-4).
- Herman, A. (2017, December 30). 'Black Mirror' watch: Talking to David Slade, the director of 'Metalhead'. *Ringer*.)

- <https://www.theringer.com/tv/2017/12/30/16803872/blackmirror-metalhead-david-slade>.
- Institute of Electrical and Electronics Engineers (IEEE). (2020). *IEEE Code of Ethics*. <https://www.ieee.org/about/corporate/governance/p7-8.html>.
- Jarmanning, H. (2019, November 25). Mass. State police tested out Boston Dynamics' Spot the robot dog. Civil liberties advocates want to know more. *WBUR*. <https://www.wbur.org/news/2019/11/25/boston-dynamics-robot-dog-massachusettsstate-police>.
- Johnson, D.G. (2009). *Computer Ethics (4th ed.)*. Boston. Pearson Education.
- Knight, W. (2021, February). Boston Dynamics' robot dog is not armed – in the name of art. *Wired*. <https://www.wired.com/story/boston-dynamics-robot-dog-armed-name-art/>.
- Koverola, M., Kunnari, A., Sundvall, J., Laakasuo, M. (2022). General attitudes towards robots scale (GAToRS): A new instrument for social surveys. [Unpublished manuscript]. Accepted for publication in *International Journal of Social Robotics*.
- McNamara, A., Smith, J., & Murphy-Hill, E. (2018, October). Does ACM's code of ethics change ethical decision making in software development? In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (pp. 729-733).
- Mozilla Foundation (2021). Teaching Responsible Computing Playbook. *Responsible Computer Science Challenge*. <https://foundation.mozilla.org/en/what-wefund/awards/teaching-responsible-computing-playbook/>
- National Academies of Sciences, Engineering, and Medicine (NAS). (2022). *Fostering Responsible Computer Science Research: Foundations and Practices*. National Academies Press. <https://doi.org/10.17226/26507>.
- Peck, E. (2017). The Ethical Engine: Integrating Ethical Design into Intro Computer Science. *Medium*. <https://medium.com/bucknell-hci/the-ethical-engine-integratingethicaldesign-into-intro-to-computer-science-4f9874e756af>
- Peck, E. (2021). Ethical CS. <https://ethicalcs.github.io/>
- Peslak, A. R. (2007). A review of the impact of ACM code of conduct on information technology moral judgment and intent. *Journal of computer information systems*, 47(3), 1-10. Quinn, M. J. (2020). *Ethics for the Information Age (8th ed.)* Boston: Pearson Education.
- Ricoeur, P. (1991). *A Ricoeur Reader: Reflection and Imagination*. University of Toronto Press.
- Saltz, J., Skirpan, M., Fiesler, C., Gorelick, M., Yeh, T., Heckman, R., Dewar, N. & Beard, N. (2019). Integrating ethics within machine learning courses. *ACM Transactions on Computing Education (TOCE)*, 19(4), 1-26.

- Skirpan, M., Beard, N., Bhaduri, S., Fiesler, C., & Yeh, T. (2018). Ethics education in context: A case study of novel ethics activities for the CS classroom. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education* (pp. 940-945).
- Tavani, H.T. (2015). *Ethics and Technology: Controversies, Questions, and Strategies for Ethical Computing* (5th ed.). John Wiley and Sons.

Transparency and product safety regarding medical diagnostic systems

Daria Onitiu

University of Edinburgh, United Kingdom

donitiu@ed.ac.uk

Abstract. A promising research area, revolutionising the early detection of diseases, entails the use of medical diagnostic systems for clinical decision support. However, we need to conduct a careful balance between a system’s performance and its usefulness in informing patient outcomes. This paper is focusing on the ethical challenges of medical diagnostic systems on patient autonomy, as well as notions of shared decision-making, and evidence-based medicine, suggesting how to address these issues in a regulatory landscape. It intends to scrutinise broader efforts by the U.S Food & Drug Administration on Artificial Intelligence and Machine Learning software as medical device and claims that the quality of oversight ultimately informs the safety of medical diagnostic systems as clinical decision support. Additional guidance needs to establish the conditions of a medical diagnostic system’s continuous alignment of patient values and stimulate causal effects in clinical reasoning.

Keywords: medical diagnostic systems, ethics, transparency

1 Introduction

Medical diagnostic systems transform healthcare. By way of illustration, imagine a medical diagnostic system that uses machine learning (ML) to classify between mild or advanced diabetic retinopathy. The U.S Food and Drug Administration (FDA) has recently approved such a device, which employs an algorithm to analyse images of a patient’s retina, providing an important contribution in managing a common disease in diabetic patients that leads to vision loss (FDA, 2018). Yet, medical diagnostic systems currently operate in a regulatory lacuna, and we need to think about how disease classification using ML will shape the role of a healthcare professional and the patient when engaging with such a system on the ground. This paper intends to uncover some ethical problems of medical diagnostic tools, focusing on the interplay of performance modifications and the need for transparency regarding the certification of these Artificial Intelligence (AI) tools.

We need a nuanced discussion on the impact of medical diagnostic systems on ethical principles, including patient autonomy, as well as notions of shared decision-making, and evidenced-based medicine. For instance, individual autonomy can conflict with notions of shared-decision-making when conflated by algorithmic constructions on disease classification. Moreover, I highlight that ML approaches do not necessarily improve the quality of decision-making, including the healthcare professionals acting on best available evidence. There is a risk that a medical diagnostic system, if not defined appropriately within ethical principles, can blur the line of human intervention, and disturb the role of ML-approaches as clinical decision-support.

The FDA intends to address some challenges of medical diagnostic systems and envisages a product lifecycle approach suitable for adaptive algorithms in a healthcare environment (FDA, 2019, p. 10; FDA, 2021, p. 3). Yet, we need to substantiate some aspects of this regulatory proposal, including the interplay between performance specifications and transparency requirements, focusing on medical diagnostic systems. I suggest a differentiated picture of how transparency goals, including so-called post hoc explainability methods in medical diagnostic systems, can complement FDA proposals that go beyond a system's intended use and will consider the impact of ML approaches on clinical decision-making.

An important step to verify the interplay ML-approaches with the users, healthcare professional and the patient in a clinical environment is to understand the role of potential and causal effects of medical diagnostic systems. That is, we introduce a "language" in the model to allow us to formalise the knowledge and assumptions in the data based on underlying cause-and-effect relationships (Miller, 2021). Modelling causal effects is one aspect of ensuring effective human oversight, and we need to consider this as an important safeguard to channel ethical principles within the role of verification and validation of medical diagnostic tools.

2 An outlook of the ethical challenges of medical diagnostic systems

The study of medical ethics allows us to discover the principles and processes to justify a particular course of action, including a practitioner's communication of evidence and considering patient involvement in the decision-making process (Laurie, Harmon, Porter, 2013; p. 2). Some research enumerates the ethical challenges of medical AI systems holistically (for example, Luxton, 2022). Building on this work, I provide a more nuanced picture showing how algorithmic processes can produce tensions with some core ethical principles within shared decision-making and evidence-based medicine.

2.1 ML approaches do not readily fit with patient autonomy

The principle of patient autonomy entails the healthcare professional's duty of negative and positive action to enable patients making an informed decision about their medical

care (Beauchamp and Childress, 2019, p. 104; Holm, 2022, p. 183). As argued by Christine and Kaldjian (2003, p.10), ‘communicating information about prognosis and treatment is recogni[s]ed as one of the clinical cornerstones of respecting patient autonomy’ as well as shared decision-making. With medical diagnostic tools, the position for defining patient autonomy lies in the system’s *data* about the individual patient and the model’s information-processing technique describing the healthcare professional’s associative process, judging between the benefits and risks regarding the system’s output.

Let us elaborate on this, considering the meaning of a consequentialist notion of contemporary medical ethics, which examines ‘the effect of a decision on individuals’ and whereby the individual creates ‘their own decisions as far as possible’ (Laurie, Harmon, Porter, 2016, p. 6). We are interested in how the use of algorithms for disease classification can support a patient’s welfare, such as satisfaction or wellbeing, including the actions leading to the best possible outcome. Therefore, we might want to focus on consequentialist theory as it puts us in the position to judge the relative effects of actions and importance of outcomes (Card and Smith, 2020, p.4).

Imagine a medical diagnostic system that can analyse chest x-rays and provide a prediction that is associated with the individual patient suffering from pneumonia. How do we evaluate the possible benefits and harms of this predictive process and how much weight should we place on the probabilities? Do we consider the degree of benefit and harm as a statement of the system’s automatically testing chest x-rays scans, or do we need to go further and look at the algorithm’s functional representations informing both, the doctor’s, and individual’s own value judgement? There is a risk that an agent’s actions and probable outcomes are rather associated with the system’s probabilistic account of disease classification, than related to the individual’s calculation of risk management and communication. That could, in turn, distort the way the patient perceives the system’s performance specifications and articulates own choice.

Moreover, healthcare professionals not only need to quantify shared decision making, considering the probability of benefits and harms regarding disease classification using AI, but also streamline the contribution of medical diagnostic systems in a clinical environment. This is certainly not an easy task, as ML approaches can disturb constructed relationships in reasoning.

2.2 ML approaches can disturb constructed relationships

A healthcare professional needs to conduct a delicate balance between knowing a patient’s best interests, including values and beliefs, and appraising the system’s reliability in individual circumstances (Grote, 2021, p. 337). We need to note (i) a healthcare professional’s consideration of a patient’s interests and choice (Christine and Kaldjian, 2003, p. 13) (ii) a healthcare professional’s positive action promoting patient wellbeing and utility (Beauchamp and Childress, 2019, p. 217), whereby the relationship between the doctor and the patient resembles a process of shared decision-making.

Nevertheless, it is important to note that ML approaches operate as a ‘black box’, whereby the algorithmic decision-making processes are not comprehensible to the

average user (Mittelstadt, Allo, Taddeo et al, 2016, p. 6). A healthcare professional may be impaired to have ‘a realistic understanding of the system’, beyond disclosing the system’s specificity and sensitivity (Holm, 2021, p. 183). However, is there a moral imperative for the healthcare professional to rely on a predictive model that can outperform human judgement and possibly override a patient’s demand for alternative treatment options? What this shows is that medical diagnostic systems can conflict with patient autonomy and beneficence, endorsing a notion of soft paternalism based on the system’s intended use as a disease classification tool.

Imagine a scenario where a medical diagnostic system detects pneumonia in an image of the patient’s chest x-ray and the healthcare professional must decide on the underlying factors that influenced the classification outcome. What follows is that medical diagnostic systems could give away new interpretations of patient autonomy and beneficence in clinical decision-making. A medical diagnostic system that operates as a ‘black box’ could minimise a healthcare professional’s discretion in exercising clinical judgement, which includes those intuitive assumptions about the appropriate treatment recommendations that reflect his or her experience and the patient’s best interests. There is a risk that the use of AI can steer positive action to the degree that the value attached to probabilistic judgements remains unverified by both the healthcare professional and patient.

We need to determine what are the harm-inducing conditions that go well beyond a system’s functional use, and which shape a healthcare professional’s risk communication and management in a process of shared decision-making. I suggest that a healthcare professional must judge a system’s confidence level based on the degree of information that allows exercising clinical discretion to balance ethical principles, as well as utilising clinical expertise based on scientific evidence. Nevertheless, another issue concerning the use of AI for decision support is that ML approaches do not necessarily promote evidenced-based outcomes, as I will show below.

2.3 ML approaches do not necessarily improve evidenced-based outcomes

Evidence-based medicine concerns the process of decision-making that combines ‘clinical expertise and patient values’ with ‘best available evidence’ (Burlacu, Iftene, Busoiu et al, 2020, p. 191). For example, a medical diagnostic system is argued to provide evidenced-based modalities in imaging, supporting clinicians ‘in integrating ever-increasing loads of medical knowledge and patient data into routine care’ (Scott, 2018, p. 44). However, medical diagnostic systems do not necessarily promote evidence-based outcomes. Google Health conducted an interesting study examining the impact of deep learning approaches on the detection of diabetic retinopathy within 11 clinics across provinces in Thailand (Beede, Baylor, Hersch et al, 2020). The report illustrates a discrepancy between the system’s accuracy in a laboratory and real-life environment and highlights that the deployment of medical diagnostic systems needs to be tailored to the clinical workflow, such as considering data quality and socio-economic factors in specific healthcare contexts (Beede, Baylor, Hersch et al 2020; Douglas Heaven 2020).

3 Medical diagnostic systems and FDA proposals: performance versus transparency

The previous discussion identified some challenges of medical diagnostic systems regarding decision-making, and the need for more guidance in considering patient autonomy, the notion of shared decision-making, and the deployment of these tools within evidenced-based statements. Therefore, we should specify the kind of verification and validation requirements required to tackle these challenges.

The FDA proposals focusing on the role of ML approaches in software as medical device establish a suitable starting point to investigate these issues. It intends to provide an approach regarding the so-called ‘update problem’ of medical devices by using ML/DL approaches that are iterative and adaptive to a healthcare environment (Gilbert, Fenech, Hirsch et al, 2021, p. 2; Gerke, Babic, Evgeniou et al, 2020, p. 1; FDA 2019, p. 3).

The FDA examines the safety and effectiveness regarding software changes in their Discussion Paper and the Action Plan, which I shall call the “FDA guidance” in the remaining part of this discussion (FDA, 2019; FDA 2021). The FDA guidance, providing an important premarket assurance regarding software modifications, illustrates the Predetermined Change Control Plan that includes Pre-Specifications (SPS) and Algorithmic Change Protocols (ACPs) (FDA, 2019, p. 10; FDA, 2021, p. 3). The SPS and the ACPs both illustrate a shift from dealing with a software’s significant changes to types of modifications that are ‘anticipated’ (FDA, 2019, p.10; FDA, 2021, p.1). In addition, the FDA guidance stipulates enhanced transparency requirements based on manufacturer’s real -world performance monitoring (FDA, 2019, p. 9; FDA, 2021, p. 1).

The FDA guidance, endorsing a notion of safety and performance of medical AI devices, including medical diagnostic systems, stands in sharp contrast to transparency. Transparency is commonly defined as the scrutability of algorithmic decision-making, being closely aligned to explainability (Mittelstadt, Allo, Taddeo et al, 2016, p. 6). As acknowledged by Mittelstadt, Allo, Taddeo et al (2016, p. 6), the debate of transparency of algorithms is not new. Nevertheless, I assume to provide a different perspective on transparency in medical diagnostic systems, which includes an extension of a system’s intended purpose to the application of ethical principles within clinical decision-making. Whilst we can agree that diagnostic skill can be examined in standard settings, such as comparing clinician judgement and the system’s specificity and sensitivity, there are important variations in clinical judgement that go well beyond academic ability (see also, Chan, Gentzkow, Yu 2021). Take the example of a medical diagnostic system that offers the same classification to similar cases, but obliges the health care professional to act differently, based on the patient’s own needs, values, and preferences. What I intend to show is that further guidance needs to specify that disease classification is a coordinated process that shapes how the doctor and the patient perceive the system’s computational interpretation of an underlying disease.

3.1 Patient autonomy and ex ante usability

An important aspect of the FDA guidance is the connection between transparency and usability using a ‘patient-centered approach’ (FDA, 2021, p.4). A system’s usability can be defined as ‘the extent to which a [ML] system can be used to achieve specified goals with effectiveness, efficiency, and patient satisfaction in multiple healthcare environments’ (Cutillo, Sharma, Foschini et al, 2020, p. 1).

Imagine a medical diagnostic system with improved performance in classifying images of diabetic retinopathy after the manufacturer retrained the algorithm on real-world data (see FDA, 2019, p.18). The first aspect for manufacturers is to ensure the system’s usability and document performance improvement, including analytical and clinical validation in the ACP regarding the system’s use (FDA, 2019, p. 18). Moreover, manufacturers need to ensure that any modifications and anticipated changes in the system’s performance are transparent to the users (FDA, 2021, p. 14). However, there is a discrepancy between the role of usability in performance modifications and usability and transparency, whereby the former is constrained to the system’s analytical and clinical performance *within* the system’s intended use. A manufacturer documenting a modified algorithm that can determine high-confident cases based on real-world data, proves the medical diagnostic system’s ‘quality of use’, considering the requirements regarding device labelling and real-world performance monitoring (Bevan, 1995). We need a more structured approach to define unanticipated impacts of medical diagnostic systems when interacting with the patients, including his or her expectations about future treatment, perception of symptoms and the role of AI in clinical judgement.

The FDA needs to consider another important aspect of usability that intends to investigate whether the medical diagnostic system promotes patient-centered outcomes. Cabitza and Zeitoun (2019, p. 161) illustrate this aspect of usability very well and distinguish between ‘statistical validity’ and the system’s usability to verify ‘the extent to which physicians can relate to the AI, attach some clinical meaning to its advice and integrate its use in their daily workflows and routines’. This interpretation of usability is appropriate to ‘fill the semantic gap’ regarding the expert’s interpretation of medical imaging using AI (Holzinger 2016, p. 122). I would add that FDA guidance might need to safeguard the system’s alignment with patient values, including autonomy to offer a comprehensive statement about a system’s usability. Accordingly, my view of usability envisages how the healthcare professional can attach different values based on the system’s probability of harm and benefits (see also, Savulescu and Wilkinson, 2019). In other words, how does the algorithm’s outlook help healthcare professionals to act in favor of the patient’s interests and values when considering the recommendation of treatment options? Does the algorithmic decision-making support the patient to act within his or her best interest?

If we want to tackle these questions, we could argue that the degree of human oversight can shape a system’s intended use beyond a medical diagnostic tool’s clinical and analytical performance. For example, a healthcare professional who is dealing with a system that classifies specific complications of diabetic retinopathy needs to weigh up different values informing risk communication and management to the patient,

rather than when dealing with mild forms of diabetic retinopathy that require the patient to come back for re-screening in the future. The FDA considers this scenario as entailing the manufacturer's real-world performance monitoring and updating device labelling, provided that the user's degree of human oversight is not based on a system's limitations or unchanged performance requirements (see also, FDA, 2019, p. 16). However, what is missing is that the manufacturer should also consider the way AI tools shape diagnostic decisions for various stages of disease classification tasks and how these findings can comply or conflict with ACPs and SPS.

The outcome is the need for a requirement of ex ante usability documenting potential risks that goes beyond the manufacturer documenting anticipated changes, such as updating ACPs using performance data. A requirement of ex ante usability would require manufacturers to update performance modifications, as well as technical safeguards, enabling different agents to foresee the degree of intervention with a medical diagnostic system, maximising an individual's choice. I will articulate the role of technical safeguards in this context in Section 3.2.

3.2 Shared decision-making and post hoc explainability

Another aspect I mentioned in the previous section is that manufacturers need to ensure that any modifications of the system's performance are transparent to the users (FDA, 2021, p. 14). This notion of transparency, whilst tied to the system's modifications, should arguably include the inscrutability of the medical diagnostic tool when it is 'actionable to the user' (COCIR, 2021, p. 19; FDA, 2019, p. 10). However, we need to provide a more comprehensive picture of the role of performance and explainability of medical diagnostic tools and analyse the way that interplay should influence the FDA approach to AI medical device regulation.

An important question arising from the notion of shared-decision making is the practitioner's and patient's involvement to act upon a system's output. The FDA guidance seems to investigate this by focusing on the device modifications that a system needs to show an '[a]ppropriate level of transparency (clarity) of the output and the algorithm aimed at users' (FDA, 2019, p. 10). The FDA, the United Kingdom's Medicines, and Healthcare products Regulatory Agency (MHRA) and Health Canada paper on Good Machine Learning Practices (GMLP) explicitly mention the 'need for the human interpretability of the model outputs' (FDA, Health Canada, MHRA 2021, Principle 7).

Post hoc explainability methods in medical imaging offer an important perspective concerning the degree of insight regarding the algorithmic decision-making. For example, a popular post hoc explainability method entails saliency maps in medical imaging tasks, highlighting the input pixels that contribute to the output (Seçkin Ayhan, Kümmerle, Kühlwein et al, 2022, p. 1). By way of illustration, a saliency map operating on an AI model to detect pneumonia can localise the regions of the image made for the prediction (see also, Da Silva, 2020).

However, explainability methods are not a source of justifying the system's reliability in individual circumstances (Lipton 2016, p. 41-42). For example, Imagine the saliency map focused on the pixel value in the image, rather than the aspects of the

image correlating with the underlying disease (Ghassemi, Oakden-Rayner, Beam, 2021, p. 746). Research by Arun, Gaw, Sing et al (2021, p.1) revealed that saliency maps do not assist clinicians to localise the underlying factors for disease classification. Accordingly, it would be naïve for us to assume that transparency safeguards and explainability ensure a system's continuous use regarding anticipated changes from the outset. However, we need then to ask ourselves how these technical safeguards correspond to 'the need for a manufacturer's transparency to users about the functioning of AI/ML-based devices [and] the benefits, risks, and limitations of these devices' and how can user perspectives inform the evaluation of the medical diagnostic system (FDA, 2021, p.5-6).

We must move away from a conception of post explainability methods enabling human observation of the algorithmic process and emphasise its role to support a healthcare professional's positive action towards an observable result. Let me elaborate on this using an example of a saliency map which outlines *some* features in a localised region in an image. Suppose now that the practitioner acknowledges that the model's localization of an area within the chest x-ray is the *deciding factor* describing the patient's pain and suffering of a disease. What this shows is that the model's relative feature importance for the output only gains significance when exhausted by the mutual interaction with the user, the healthcare professional and patient.

Future work needs to inform the role of post hoc explainability methods to inform the model's reconciliation with clinical decision-making, rather than concordance with human judgement based on the system's functionality in a clinical context. FDA guidance needs to establish post hoc explainability as a process-based verification method that requires specific training for clinicians when dealing with visualisation methods including saliency maps and translates medical diagnostic systems within the ambit of shared-decision making regarding diagnostic tasks.

Another aspect related to the example above is the representation of knowledge and expertise regarding an underlying disease as being a deciding factor for the model's output. The healthcare professional's current expertise about disease classification, including the diagnosis and treatment of pneumonia would contribute to his or her positive action to observe the algorithms' classification. Accordingly, I am going to analyse how clinical evaluation is another aspect for assessing probabilities for the system's alignment with evidenced-based medicine in Section 3.3.

3.3 Evidenced-based outcomes and clinical evaluation

The FDA, MHRA, and Health Canada's GMLP document recognises that datasets need to be 'representative of the intended patient population' and users need to be informed about the 'performance of the model for appropriate subgroups, characteristics of the data used to train and test the model, acceptable inputs, known limitations, user interface interpretation, and clinical workflow integration of the model' (FDA, Health Canada, MHRA, 2021, Principle 3, Principle 9).

Having said that, many FDA- approved medical AI devices are based on retrospective studies, making it more difficult for operators to understand the system's

nuances applied to individual cases (Wu, Wu, Danshjou et al, 2021, p. 582). Hence, it is argued that manufacturers need to consider validation studies considering prospective randomised studies to assess actual clinical outcomes, as well as comparisons between clinicians' performances with and without the medical diagnostic system (Wu, Wu, Danshjou et al, 2021, p. 582-583). The idea of randomized control trial is to 'unmask vulnerabilities' such as generalizability of the system's performance and inherent limitations, such as overfitting, to define patient-centered goals and outcomes (Wu, Wu, Danshjou et al, 2021, p. 582-583).

Therefore, the first aspect defining the role of clinical validation of AI systems is to achieve the 'empirically rigor' 'to maximize the benefits' of a medical diagnostic tool, whilst 'offsetting potential effects' regarding a system's implementation (McCradden, Anderson, Stephenson et al, 2022, p. 9). This requires the FDA to establish the parameters for constituting a sufficient demonstration of evidence and confirming the system's reliability in a clinical setting. The work done by researchers in developing reporting guidelines on clinical trials on AI tools, such as the SPIRIT-and CONSORT AI extension (Liu, Cruz Rivera, Moher et al, 2020; Cruz Ribera, Liu, Chan et al, 2020) is helpful to establish transparency of AI-based recommendations but more good practices need to deal with evaluation considering adaptive algorithms (Liu, Cruz Rivera, Moher et al, 2020, p. 1371). Moreover, manufacturers need to ensure that the clinical validation of an AI system considers how the information collected in a prospective study is contextualised in a clinical setting (Liu, Glocker, McCradden, Ghassemi et al 2022, p. 385). We need to address this question from the perspective of clinical evaluation as well as the post-market monitoring, to ensure that the system's performance ensures clinical outcomes are respected (Wu, Wu, Danshjou et al 2021, p. 583).

However, another issue is the inherent risks related to the misuse of medical diagnostic tools when interacting with a patient. By way of illustration, imagine a medical diagnostic system trained within a specific sub-population in San Francisco, United States and used on a different population with different socio-economic factors. What this shows is that we must pay attention to the 'more insidious failures, such as an algorithm that gives racially biased recommendations because it was trained with subtly biased data' (Kaushal and Altman, 2019, p. 62). I believe that the FDA needs to build up its statement in Principle 9 of the GMLP and elaborate on the risks of how medical diagnostic systems can give rise to "off-label" use (FDA, Health Canada, MHRA, 2021, Principle 9). Accordingly, we need more guidelines on how users need to be informed about the risks of a system's performance bias in a clinical setting and how unintended uses can be monitored during a system's deployment, considering the role of practitioners to ensure effective risk management and communication of probability of harm when these AI tools operate on the ground.

4 Setting the tone for the role of causal effects in medical diagnostic systems

An important point, which surfaced in my discussion concerning the verification and validation of medical diagnostic systems, is that clinicians and patients need to have more control of the tool's operation as decision-support. I established that we need a requirement of ex ante usability, which entails technical safeguards to enable a degree of human intervention regarding potential and unanticipated changes in the system's operation. In addition, I elaborated on the role of technical safeguards, including post hoc explainability, and how the practitioner and patient can realistically assess the system's confidence within a specific setting. Finally, I highlighted that the manufacturer needs to consider how different social actors contextualise a system's predictions to ensure evidenced-based outcomes. I will be closing the argument by specifying how we should discuss the degree of human control in the future.

Modeling causal effects and so-called 'what if' scenarios (Holzinger, 2021) can safeguard the role of medical diagnostic systems in clinical decision-making. Causal effects are not inherent in the algorithmic process (Pearl and Mackenzie, 2018). For example, a common argument regarding ML- predictions in healthcare is that the healthcare professional should rather know *why* the tool is making certain predictions than relying solely on algorithmic correlations (Holzinger, Carrington, Müller, 2020). We can investigate whether a model's simulation of causal effects can stimulate decision-making when acting within the interests of the patient. Whilst a comprehensive discussion would exceed the scope of this paper, I suggest extending my investigation by discussing the FDA approach and making three recommendations, based on the role of human control and intervention with medical diagnostic systems:

First, we need to identify the role of patient autonomy to justify certain actions when a ML system is performing certain tasks. Ex-ante usability is a requirement that needs to shape information duties, as well as device changes. In doing so, we must not underestimate the degree of one's own action to assess the contours of individual choice. That means a practitioner can *create* the patient's manifestation of a disease based on the system's association with the patterns in the data. Ex-ante usability is a requirement for manufacturers to engage with some boundary work and investigate how users, healthcare professionals and patients engage with a different set of scenarios. One way in doing so is the use of counterfactual explanations in medical imaging, which allows the individual to stimulate decision-making focusing on alternative scenarios (Vermeire, Brughmans, Goethals et al 2022). Counterfactuals allow us to quantify likelihood of harm by isolating individual circumstances, including preconditions (Glocker, Musolesi, Richens et al 2021, p. 3). This allows us to scrutinise and define those outcomes which are most closely related to the actions promoting wellbeing and the patient's interests.

Second, we need to steer a healthcare professional's positive action to achieve reconciliation concerning the use of medical diagnostic tools as clinical decision-support. Therefore, an important step for us would be examining the way healthcare professionals assess confidence levels with reference to the AI, including what parameters allow a healthcare professional to build up a reasonable justification for

positive action, such as recommending the patient treatment options. Here, I would specify that counterfactuals in medical diagnostic systems could direct the individual to engage with his or her perceptual judgement, based on possible reconstructions of the underlying factors, and navigating through the system's classificatory purpose.

Finally, we see the role of causal effects to define the need for expert knowledge in clinical evaluation (Glocker, Musolesi, Richens et al 2021, p.1). It is argued that randomized control trials allow the evaluation of causal hypothesis (Prosperi, Guo, Sperrin et al 2020, p. 369). Nevertheless, London (2019, p. 17) argues that the extent domain experts scrutinise those causal relationships 'derives from experience and precedes our ability to understand why interventions work'. What follows is that the aim in understanding causal effects in disease classification is based on dealing with uncertainties informing both, clinical evaluation, as well as risk communication regarding the use of AI on the ground. Hence, what we need are metrics for reasonable causal conclusions guiding randomized study.¹ The role of decision-making here is crucial to inform randomized study including the level of human expertise and 'how the outputs of the AI system were used to contribute to decision-making or other elements of clinical practice' (emphasis added by author) (Liu, Cruz Rivera, Moher, 2020, p. 1371). What needs to be added; however, is that we need further multidisciplinary engagement that discusses the delicate process that should exemplify both, the benefits of explanations stimulating causal thinking between various users, healthcare professionals and novice operators and risks of the use of AI as decision support where one cannot untangle causation from correlation.²

5 Conclusion

The FDA is facing some big questions when dealing with the future of ML innovation in medical devices operating on the ground. This discussion aims to scrutinise the proposals and identify whether and how the FDA approach can leverage a patient-centered approach more broadly, considering the challenges of medical diagnostic systems implementing patient autonomy, as well as notions of shared decision-making, and evidence-based medicine. Defining transparency in how medical diagnostic systems stimulate clinical reasoning and patient values is the next step for regulators to ensure the role of medical AI tools as a reliable and safe decision-support on the ground.

¹ Indeed, another area concerns causal inference using observational data when a randomised study is not practicable (Hernán and Robins, 2016).

² London (2019, p. 17) uses the example of healthcare professionals prescribing aspirin as a drug that could relieve pain 'for nearly a century without understanding the mechanism through which it works'.

Acknowledgement

The work benefitted from the research undertaken at the UKRI Research Node on Governance & Regulation within the Trustworthy Autonomous Systems programme. The work also benefitted from the author's research stay at the Stanford Center for AI Safety.

References

- Arun, N. Gaw, N. Singh, P. Chang, K. Aggarwal, M. Chen, B. Hoebel, K. Gupta, S. Patel, J. Gidwani, M. Adebayo, J. Li, MD. Kalpathy-Cramer, J. (2021). Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. *Radiology: Artificial Intelligence*, 3 (6), pp. 1-12.
- Beauchamp, TL. Childress, JF. (2019). *Principles of Biomedical Ethics*. Oxford University Press.
- Beede, E. Hersch, F. Iurchenko, A. Wilcox, L. Ruamviboonsuk, P. Vardoulakis, L. (2020). A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1-12.
- Bevan, N. (1995). Measuring usability as quality of use. *Software Quality Journal*, 4 (2), pp.115-130.
- Burlacu, A. Iftene, A. Busoiu, E. Cogean, D. Covic, A. (2020). Challenging the supremacy of evidence-based medicine through artificial intelligence: the time has come for a change of paradigms, *Nephrology, Dialysis, Transplantation*, 35 (2), pp. 191-194.
- Cabitza, F. Zeitoun, J-D. (2019). The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence, *Annals of Translational Medicine*, 7 (8), pp. 1-9.
- Card, D. Smith, NA. (2020) On Consequentialism and Fairness. *Frontiers in Artificial Intelligence*, 3, pp.1-11.
- Chan Jr, DC. Gentzkow, M. Yu, C. (November 2021, revised May 2021). *SELECTION WITH VARIATION IN DIAGNOSTIC SKILL: EVIDENCE FROM RADIOLOGISTS*. www.nber.org/system/files/working_papers/w26467/w26467.pdf.
- Christine, PJ. Kaldjian, LC. (2013). Communicating Evidence in Shared Decision Making. *The Virtual Mentor*, 15 (1), pp. 9-17.
- Cruz Rivera, S. Liu, X. Chan, AW. Denniston, AK. Calvert, MJ. The SPIRIT-AI and CONSORT-AI Working Group, SPIRIT-AI and CONSORT-AI Steering Group, SPIRIT-AI and CONSORT-AI Consensus Group. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nature Medicine*, 26 (9), pp-1351-1363.
- Cuttilo, CM. Sharma, KR. Foschini, L. Kundu, S. Mackintosh, M. Mandl, KD. (2020). Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency, *NPJ Digital Medicine*, 3 (1), pp. 1-5.
- Da Silva, M. (8 October 2020 GitHub). *Interpretable Deep Learning Part II: Visual Interpretability with Attribution Methods*. <https://metrics-lab.github.io/2020/10/08/visual-interpretability-with-attribution-methods.html>.
- Douglas, HW. (27 April 2020 MIT Technology Review). *Google's medical AI was super accurate in a lab. Real life was a different story*. <https://www.technologyreview.com/2020/04/27/1000658/google-medical-ai-accurate-lab-real-life-clinic-covid-diabetes-retina-disease/>.

- FDA. (11 April 2018 FDA Press Release). FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. <https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye>.
- FDA. (2019). Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback. <https://www.fda.gov/media/122535/download>
- FDA. (2021). Artificial Intelligence/Machine Learning (AI/ML)-Based: Software as a Medical Device (SaMD) Action Plan. <https://www.fda.gov/media/145022/download>
- FDA., Health Canada., MHRA. (2021). Good Machine Learning Practice for Medical Device Development: Guiding Principles. <https://www.fda.gov/media/153486/download>
- Gerke, S. Babic, B. Evgeniou, T. Cohen, IG. (2020). The need for a system view to regulate artificial intelligence/ machine learning-based software as medical device. *NPJ Digital Medicine*, 3 (1), pp. 1-4.
- Ghassemi, M. Oakden-Rayner, L. Beam, AL. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet*, 3 (11), pp. 745- 750.
- Gilbert, S., Fenech, M., Hirsch M., Upadhyay S., Biasiucci A., Starlinger J. (2021). Algorithm Change Protocols in the Regulation of Adaptive Machine Learning-Based Medical Device. *Journal of Medical Internet Research*, 23 (10), 1-8.
- Glocker, B. Musolesi, M. Richens, J. Uhler, C. (2021). Causality in digital medicine. *Nature Communications*, 12 (1), pp- 1-6.
- Grote, T. (2021). Trustworthy medical AI systems need to know when they don't know. *Journal of Medical Ethics*, 47 (5), pp. 337-338.
- Hernán, MA Robins, JM. (2016). Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology*, 183 (6), pp. 758-564.
- Holm, S. 2022. Handle with care: Assessing performance measures of medical AI for shared clinical decision-making. *Bioethics*, 36 (2), pp- 178-186.
- Holzibger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop?.*Brain Informatics*, 3 (2), pp. 191-131.
- Holzinger, A. (2021). Explainable AI and Multi-Modal Causability in Medicine. *De Gruyter Oldenburg*, 19 (3), 171-179
- Holzinger, A. Carrington, A. Müller, H. (2020). Measuring the Quality of Explanations: The System's Causability Scale. *Künstliche Intelligenz (Oldenburg)*, 34 (2), pp. 193-198.
- Laurie, GT. Harmon, HE. Porter, G. (2016). *Law & Medical Ethics*. Oxford University Press.
- Lipton, ZC. (2016). The Mythos of Model Interpretability. *Communications of the ACM*, 61 (1), pp. 36-43.
- Liu, X. Cruz Rivera, S. Moher, D. Calvert, MJ. Denniston, AK. The SPIRIT-AI and CONSORT-AI Working Group. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine*, 26 (9), pp. 1364-1374.

- Liu, X. Glocker, B. McCradden, MM. Ghaessemi, M. Denniston, AK. Oaken-Rayner, L. (2022). The medical algorithmic audit, *The Lancet*, 4 (5), pp. 384-397.
- London, AJ. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *The Hastings Center Report*, 49 (1), pp. 15-21.
- Luxton, DD. (2022). AI decision-support: a dystopian future of machine paternalism?. *Journal of Medical Ethics*, 48 (8), pp. 232-233.
- McCradden, MD. Anderson, JA. Stephenson, EA. Drysdale, E. Erdman, L. Goldenberg, A. Zlotnik Shaul, R. (2022). A Research Ethics Framework for the Clinical Translation of Healthcare Machine Learning, *The American Journal of Bioethics*, 22 (5), pp. 8-22.
- Miller, K. (16 March 2021 Stanford Hai). *Should AI Models Be Explainable? That depends.* <https://hai.stanford.edu/news/should-ai-models-be-explainable-depends>.
- Mittelstadt, BD. Allo, P. Taddeo, M. Wachter, S. Floridi, L. (2016). The ethics of algorithms: mapping the debate. *Big Data & Society*, 3 (2), pp. 1- 21.
- Pearl, J., Mackenzie D. (2018). Mind over Data. *Significance*. 15 (4), 6-7.
- Prosperi, M. Yi, Guo. Y, Sperrin, M. Koopman, JS. Min, JS. He, X. Xing, R. Shannan, W. Mo, B. Iain, E. Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2 (7), pp. 369-375.
- The European Coordination Committee of the Radiological, Electromedical and Healthcare IT Industry (COCIR). (2021). Artificial Intelligence in EU Medical Device Legislation. www.cocir.org/media-centre/publications/article/cocir-analysis-on-ai-in-medical-device-legislation-may-2021.html.
- Scott, IA. (2018). Machine Learning and Evidence-Based Medicine. *Annals of Internal Medicine*, 169 (1), pp. 44-46.
- Seçkin Ayhan, M. Kümmerle, LB. Kühlwein, L. Inhoffen, W. Aliyeva, G. Ziemssen, F. Berens, P. (2022). Clinical validation of saliency maps for understanding deep neural networks in ophthalmology. *Medical Imaging*, 77, pp.1-29.
- Vermiere, T. Brughmans, D. Goethals, S. Barbossa de Oliveira, RM. Martens, D. (2022). Explainable image classification with evidence counterfactual. *Pattern Analysis and Applications*, 25, pp. 315-335.
- Wu, E. Wu, K. Daneshjou, R. Quyang, D. Ho, DE. Zou, J. (2021). How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals, *Nature*, 27 (4), pp. 582- 584.

An Interdisciplinary Approach to European Trustworthy Digital Environments

Martin Griesbacher and Hristina Velanova

University of Graz, Graz, Austria

martin.griesbacher@uni-graz.at

Abstract. The paper presents a dedicated interdisciplinary cybersecurity approach, which integrates the consideration of European values and fundamental rights. It goes beyond a technological perspective on information security by reformulating the objective of cybersecurity to the goal of establishing European trustworthy digital environments. The two main components of the approach are: 1.) A broader view on digital assets (data, financial, knowledge, and reputational assets) and organisational vulnerability layers (soft-/hardware, physical, social legal, and ethical layer) and 2.) the added attention to the protection of the integrity of individuals (physical, mental, material, social, legal, and ethical integrity). The paper also identifies further connections to tools for the advancement of human factor-related cybersecurity (e.g., in cyber risk assessment, awareness, and training measures). The presented cybersecurity approach exemplifies the value of perspectives from social science and humanities to provide a dedicated human factors-oriented cybersecurity approach which also takes into account European values and fundamental rights.

Keywords: cybersecurity, trustworthiness, ethics, organizational cybersecurity

1 Introduction: A European interdisciplinary approach to cybersecurity

While technological innovation is advancing with an unprecedented pace, so do new uncertainties and more complex threats and risks related to these new technologies arise. Currently, one of the greatest challenges today's information society is confronted with is how to build secure ICT infrastructures and how businesses can establish themselves as trustworthy players. Undoubtedly, both secure infrastructure as well as solid cybersecurity strategies are a necessary precondition for organizations in signalling trustworthiness to existing and prospective consumers as well as to the business community in general. Nevertheless, as numerous scandals and incidents worldwide in the past years suggested, cybersecurity, or as we argue, a purely technical understanding of cybersecurity, is not sufficient for building a trustworthy digital environment. For this, a much broader perspective is needed, one which

extends the technical scope of cybersecurity by considering European norms and values to meet the needs for a trustworthy digital environment of all entities involved.

This article presents and integrates the outcomes of two European research projects, which shed light on the complex interdependencies between technological designs of cybersecurity and societally legitimized ethical expectations in building up trustworthy digital environments from a social science and humanities (SSH) perspective.¹ The goal of this article is to support the development of trustworthy digital environments by identifying the interfaces of ethical considerations to the concept of cybersecurity and the connections between conceptual and practical elements for enhancing cyber resilience². For that, we first start with a brief conceptual analysis of trust and trustworthiness and their relevance for cybersecurity and the digital environment. We then discuss the ethical requirements for trustworthy digital environments in a European context and present a criteria catalogue for trustworthy ICT products and services with well-defined core areas of trustworthiness. In a next step, the article integrates the discussion of these core areas for digital trustworthiness into a humancentred definition of cybersecurity. Through this definition, cybersecurity moves away from the purely technical realm towards a human-centric concept that strongly values individual as well as organizational integrity. In the conclusion and outlook we discuss key interfaces of the presented interdisciplinary cybersecurity approach to the practical implementation of measures to improve human factor-related cybersecurity (e.g., risk assessment, training and awareness).

2 Conceptual relationships between trust, trustworthiness and cybersecurity

As we already pointed out, amidst the fast-paced digital transformation businesses face two main challenges. First, they must assert themselves as trustworthy players. Second, to do so, they have to build trustworthy ICT infrastructure that would remain secure and robust against the rising number of risks and threats in the digital environment. Here we examine the relationship between the key concepts of trust, trustworthiness, and cybersecurity.

Trust is considered to be an attitude, a belief, one has towards another person that she/he will act according to her/his expectations that shape a particular situation. This makes trust a three-place predicate where A (the trustor) trusts B (the trustee) to do x (Hardin 2002, 9; Baier 1986, 236; Luhmann 1980, 27). From here it follows that trust

¹ The projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833923 (SOTER) and No 731711 (TRUESSEC.eu).

² Cyber resilience can be defined as the "ability of an organisation to continue to carry out its mission by anticipating and adapting to cyber threats and other relevant changes in the environment by withstanding, containing and rapidly recovering from cyber incidents" (FSB 2018, 9).

is context-dependent, namely, it relates to the realization of a particular task. Several requirements must be fulfilled so that one can talk about trust. Trust is always accompanied with risks and uncertainties regarding the behaviour of the trustee. This

means that the trustor gives the trustee a discretionary power to affect her interests with the risk that the trustee might abuse that power (Hardin 2002, 11). In this way, the trustor makes herself vulnerable to the trustee. Trust also comes with an attitude of optimism that the trustee is competent to do what she is entrusted to do (McLeod 2021). Trust also presupposes expectations and confidence that the trustee will behave as expected. This is an important requirement as it clearly distinguishes trust from distrust. Additionally, from a sociological perspective, trust is always influenced by societal conditions. While dynamics in public opinion can swiftly change individual opinions, trust remains an important precondition for the functioning of societies (see Luhman 2017).

Trustworthiness is closely related to trust. If trust is an attitude the trustor forms about the trustee, trustworthiness presents a quality that is the object of that attitude and that at the same time provides sufficient reasons that would justify the attitude of trust. Like trust, trustworthiness is also considered as a three-place relation where B is trustworthy for A regarding the performance of x. Based on this, if B is trustworthy for A regarding x, then it is rational for A to trust B (Nickel, Franssen and Kroes 2010, 431). Shifting the attention from trust to trustworthiness supports investigations of context-depending conditions which can be made explicit for a conscious decision to trust.

The discussion about trust and trustworthiness does not end here. Scholars have also investigated the motives that motivate the trustee to act in a trustworthy manner. For some the trustee is motivated by her self-interest to act trustworthy and hence she encapsulates the trustor's interests into her own (Hardin 2002, 3-5). For others it is the goodwill of the trustee that is considered as a motivation. The trustor counts on the trustee to act out of goodwill towards her and to do what she is entrusted to do (Baier 1986; Jones 1996). There are also scholars who see moral commitment, moral obligation and virtue as a motivation for the trustee to be trustworthy. For instance, the trustor might place her trust in a stranger by relying on the stranger's moral commitment and moral integrity (McLeod 2021).

Having briefly described the concepts of trust and trustworthiness, we will now elaborate on how they are related to and relevant for cybersecurity. The increased presence of business to consumer (B2C) as well as business to business (B2B) interactions in the digital realm raises new challenges to building and maintaining trust and trustworthiness in digital contexts. The risks, threats and vulnerabilities seem to be even higher and more complex in the digital environment. These include ransomware, malware, cryptojacking, threats against availability and integrity, threats against data, e-mail related threats such as phishing, whaling or vishing or non-malicious threats that result from human errors and misconfigurations but also from physical disasters that target IT infrastructures (ENISA Threat Landscape 2021). The threats target the information and the data that is exchanged or created in B2C and B2B interactions, the information and communication technology that is used, the network via which the interactions take place as well as the human factor, that is, the

people, who directly or indirectly interact with the technology or are affected by it such as business employees or clients. Such a vivid threat landscape demands greater protection of the vulnerable assets. A good cybersecurity strategy can contribute to reducing these threats and at the same time instil a feeling of trust among all stakeholders. In that regard, cybersecurity can be seen as a necessary condition for building and maintaining trust and at the same time as a prerequisite for businesses to signal trustworthiness in the digital environment. But at the same time, cybersecurity measures alone might not be sufficient to convince individuals and the public to engage in the use of an ICT product, service or process. Here, the dependency on societal contexts is crucial. ICT offered within the European Union will have to consider the relevant normative context. This includes not only the requirement to be compliant with existing regulations but also to take into consideration ethical expectations which often go beyond compliance (see also Stelzer & Veljanova 2017). In summary, the implementation and adoption of ICT should be guided by the objective of being part of a European trustworthy digital environment. After all, today any engagement with ICT comes along with the risk to be harmed (e.g., loss of privacy or financial resources).

3 Core areas and criteria for a trustworthy digital environment

In the previous section we argued that cybersecurity is a necessary condition for building and maintaining trust, but also that the objective of a trustworthy digital environment goes beyond cybersecurity measures. In this section we ask the question what requirements an organization has to fulfil to be considered trustworthy so that it would be rational, for instance, for consumers, to trust it. For this purpose, we suggest going beyond a purely technical understanding of cybersecurity and more into a humancentred one. This implies that we should analyse cybersecurity in relation to other values that directly or indirectly contribute to its attainment. It is precisely this aspect that we consider an added value of this paper as it does not offer a mere technical analysis of cybersecurity, but it attempts to look into the non-technical realm as well.

The H2020 project TRUESSEC.eu provided an example of such an analysis. The project focuses on strengthening consumer trust in new and emerging ICT products and services by encouraging the adoption of assurance and certification processes that consider multidisciplinary aspects while ensuring the protection of European fundamental rights and the respect of European values in the digital environment. As part of the project an interdisciplinary team consisting of ethicists, data protection experts, sociologists, computer scientists and business representatives produced a tool in the form of a Criteria Catalogue whose main goal is to support consumers in assessing the trustworthiness of businesses, ICT products, services or processes they are using or planning to use based on a set of values, criteria and indicators. At the same time, the Criteria Catalogue serves as a compass to businesses to assess their own trustworthiness and the trustworthiness of their ICT products and services (Stelzer et al., 2018; Stelzer & Veljanova, 2020).

The TRUESSEC.eu Criteria Catalogue follows a hierarchical structure and consists of three elements: core areas, criteria and indicators of trustworthiness that can contribute towards building trustworthy ICT products and services and thus towards greater trust in the digital environment. The core areas reflect those values that should be considered in the design and use of ICT products and services. They are defined based on the European values as stated in particular in Art. 2 and Art. 3 of the *Treaty on European Union* (2012) such as human dignity, freedom, equality, respect for human rights, non-discrimination, justice, equality and solidarity. Additionally, the core areas also draw from the European fundamental rights such as the right to the integrity of the person, respect for private and family life, protection of personal data, the freedom of expression and information, the freedom to conduct a business, nondiscrimination as enshrined in the *Charter of Fundamental Rights of the European Union* (2012). Based on this framework we identified six core areas of trustworthiness: transparency, privacy, anti-discrimination, autonomy, respect, and protection. The goal of the core areas is not to point out what businesses should do to remain compliant, as this does not suffice for enhancing trustworthiness. Their goal rather is to go beyond compliance and suggest ways how this post-compliance could be accomplished (Stelzer et al., 2018). This is for the simple reason that compliance is indeed necessary in the digital, however, it is not sufficient to guide the society in the right direction (Floridi, 2018).

The six core areas serve as an important basis for the development of criteria for trustworthiness as they indicate what criteria are needed so that the core areas are appropriately addressed. The Criteria Catalogue is comprised of twelve criteria. Examples include information, user-friendly consent, enhanced control mechanisms, privacy commitment, transparent processing of personal data, law enforcement declaration and appropriate dispute resolution (Stelzer et al., 2018). This interdisciplinary approach showcases in an exemplary way, how ethical considerations can extend the established discourse of technical security standards by refocusing on the question what actually makes an ICT product, service or process trustworthy in the European ethical context.

4 Understanding organizational cybersecurity in the context of trustworthy digital environments

While the question of the criteria catalogue addresses more the relationship between an individual and a concrete ICT product, service or process, the implementation and governance of cybersecurity strategies for organisations need its own interdisciplinary approach, which is compatible with the idea of European trustworthy digital environments. Here we again deploy a SSH perspective to systematically identify relevant phenomena for the integrity of organisations and individuals that go beyond a mere technological perspective on information security (see also Griesbacher et al., 2022; Renwick et al., 2020).

4.1 Digital Assets and Vulnerability Layers

For a definition of cybersecurity with special consideration of human factor-based aspects we can use the “more representative definition of Cyber Security” by Schatz et al. (Schatz et al., 2017) as a starting point, which is based on a semantic analysis of a wide scope of existing definitions in industry, government and academia. According to them, cybersecurity comprises “the approach and actions associated with security risk management processes followed by organizations and states to protect confidentiality, integrity and availability of data and assets used in cyber space.” Being derived from mostly tech-based publications, this definition addresses important aspects of cybersecurity, but lacks a broader and more systematic understanding of human factorrelated digital assets and vulnerabilities for being able to capture cybersecurity as a socio-technical phenomenon. This incorporates the widely recognized insight that human behaviour and technology are always deeply connected (Carlton et al. 2019; D’Arcy et al. 2009; Siponen and Vance 2010; Whitmann 2018). Considering the sociotechnical connection, we expand central concepts of the definition of cybersecurity by human factor-based elements. First, we integrate a broader view on the question of the integrity of organizational digital assets. *Primarily, any professional organization will want to keep complete control of digital assets in its organizational environment*, and thereby maintain organizational integrity (see Figure 1). We can identify four main types of digital assets to be protected for organizational digital integrity:

1. data: all types of digitally stored information (e.g., personal information of customers)
2. financial: all digital assets which can be directly exchanged into any payment currency
3. knowledge: know-how and intellectual property which is digitally stored or can be accessed digitally (for knowledge risks see North et al. 2019)
4. reputational: the trust attributed to the organization in cyber space (e.g., user ratings)

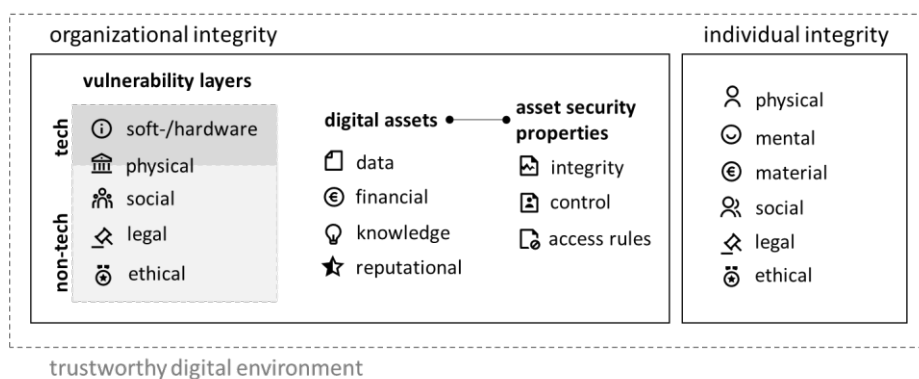


Figure 1: Interdisciplinary approach to trustworthy digital environments

Protecting digital assets includes (1.) Ensuring the integrity of the contents of these assets; (2.) Enforcing access rules to these assets (incl. questions of confidentiality and availability), (3.) Controlling the rights to change, transfer or copy of the asset. While this definition of protection goals is based on assumptions widely accepted in information security research (see Eckert 2018), it deviates from the established CIA triad (Confidentiality, Integrity, and Availability). This does not imply the intention to replace but to complement cybersecurity efforts. The CIA triad is optimized towards data and technological systems and does not sufficiently cover the specific characteristics of other type of digital assets. For example, reputational assets of an organisation cannot be controlled the same way as data within the organisations' systems. Scoring of their services or the organisation itself is often processed on third party systems (e.g., ratings in Google search). Focussing on the protection goal of "access rules" would mean here, that only users are participating in the scoring, which can be considered as legitimate (e.g., only users which really are customers of the organisation or have bought its products or services).

The risk for unauthorised or unintended changes of security properties of digital assets, from now on called security incidents, can be located in different vulnerability layers of an organisation. These layers are functioning as interfaces of an organisation to its digital assets. For cybersecurity risk assessment and management, Ganin et al. (2017) identify three main vulnerability domains, the physical (hardware), information (software) and social (personnel) domain. Most of the vulnerabilities associated with the information domain have to be covered by specialized information security personnel (see the "tech" area in dark grey in Figure 1). In this paper, however, we focus on vulnerabilities associated with human factor-based cybersecurity issues (see the "non-tech" area in Figure 1). In this area, the vulnerability to cybersecurity incidents is at least to a certain extend directly connected to human behaviour. For incidents in the physical layer this could mean an employee giving an intruder disguised as a contractor of a delivery company access to hardware. The social layer includes vulnerabilities that arise most prominently from social engineering attacks, but also from negligent or malevolent behaviour of employees within an organisation. Due to the continually increasing regulation in the digital domain (e.g. GDPR, NIS directive) and the growing issue of online disinformation (see Martens et al. 2018) the approach needs to be extended by two human factor-based vulnerability layers: the legal and the ethical layer. At a basic level, legal and ethical threats are not based on technology, but can rather be attributed to legal and social dynamics. They can be a primary cause for the loss of digital assets. As example, an externally orchestrated online 'review bombing' of an organisation's products, or services, (see Tomaselli et al., 2021) can in turn lead to a loss of reputational assets or can have secondary consequences such as the organisation being subsequently targeted due to bad press or specific allegations. Accordingly, within the security domain, the issue of handling personal data and the public behaviour of employees has increased to be an important issue for cyber resilience, especially as reputational damage is sometimes of immeasurable value to organisations, and in turn hard to recover once breached.

To prevent losses of organizational integrity, cybersecurity should be addressed on all identified vulnerability layers (software, physical, social, legal & ethical). All layers must be secured with targeted security measures. Loss of integrity within one layer might bypass security measures on other layers (e.g. strong encryption as technical measure bypassed via stolen passwords or decryption keys of employees, acquired in social engineering attacks). Therefore, in this interdisciplinary approach, cybersecurity goes beyond the narrower approach of information security. For example, corresponding losses of integrity caused by security incidents can be the cause of further damages to an organisations' integrity (e.g. undue loss of financial or reputational assets).

4.2 Combining organizational and individual integrity

As European cybersecurity research and innovation efforts should be conceptualized within the relevant legal and ethical frameworks, European values and fundamental rights must be considered as integral part of any effort for improving cybersecurity. In our interdisciplinary approach this is represented by additionally considering the integrity of any individual involved in cybersecurity measures (incl. the physical, mental, material, social, legal, and ethical integrity). Cybersecurity "made in Europe" therefore should put efforts to protect the organizational as well as the individual integrity together to implement and support not just in a stricter sense secure but an overall trustworthy digital environment.

Understanding cybersecurity in the context of European values and norms (esp. regarding fundamental rights) extends the attention in cybersecurity governance to the potential harms to individual integrity. As many incidents not only concern the integrity of an organisation, cybersecurity should also target the protection of the integrity of any individual involved not only in the case of an incident (e.g., in the case of theft of customer data) but also regarding the potential effects of security measures themselves (e.g., monitoring/surveillance vs. privacy).

How can we define the integrity of an individual in a European normative context? The presented criteria catalogue in section 3 presented six core areas for trustworthiness based on a set of European values and fundamental rights. From these normative concepts we can derive the main dimensions of individual integrity. An individual might be harmed physically (e.g., the general understanding of security and safety), mentally (e.g., psychological stress), materially (e.g., the theft of financial assets), socially (e.g., the disruption of societal participation), legally (e.g., the loss of personal data) and ethically (e.g., the violation of legitimate normative expectations). With this, we can define individual integrity as a situational attribute, where the probability of damage to the physical, mental, material, social, legal and ethical integrity of all individuals involved (e.g., customers or employees) is as low as possible. Interacting with ICT products, services and process within a European context demands the demonstration of a trustworthy digital environment, so that individuals can make itself vulnerable and confidently engage, interact, and depend on ICT.

In summary, all activities of an organisation targeted at improving cybersecurity need to consider causing potential harm to involved entities, in regard to all individual integrity attributes (physical/bodily; mental/psychological; material/financial; social; legal/rights; ethical standards). With this approach we integrate ethical considerations into a (holistic) interdisciplinary cybersecurity approach which enable new avenues for certification and standardization. The interdisciplinary approach allows to place any cybersecurity action within an organisation in the context of a trustworthy digital environment strategy. The main goal is not just to implement exhaustive controlling and monitoring instruments, but to establish and communicate a trustworthy environment for insiders and outsiders (e.g., customers). This is understood as a dedicated European approach to cybersecurity, based on the respect of European values and fundamental rights.

5 Conclusion and Outlook

The goal of this paper is to present an interdisciplinary approach to cybersecurity to complement established technical approaches to information security. We argued, that by applying SSH perspectives we can identify a wider set of digital assets and vulnerabilities of organisations in the digital domain and also integrate European values and fundamental rights in cybersecurity considerations. In this interdisciplinary approach, the implementation of protective measures for digital assets needs to respect the integrity of all potentially involved individuals. This allows organisations not only to highlight (obligatory) security features. It also enables the presentation of a trustworthy digital environment, which explicitly acknowledges European values and fundamental rights.

The interdisciplinary approach to cybersecurity is not only an academic exercise in systematic conceptualization. It can also be seen as a fundamental cornerstone to improve human factor-related cybersecurity methods and standardisation. Currently, in information security standards and threat taxonomies no specialized attention is given to non-technological aspects. For example, the ISO/IEC 27000 family does include only a few details on training requirements for general employees (ISO/IEC, 2018) and the widely used threat taxonomies of MITRE have no dedicated list of threats relevant for personnel outside of information security. The latter would be relevant, because general employees would need to know especially only about those threats, where their behaviour can make a difference between theoretical risk and actual incident.

By identifying the non-technological vulnerability layers (physical, social, ethical, and legal) it is possible to identify relevant cyber threats for the different layers. A first collection of human factor-related threats has already been made available by the SOTER project (see Venegas Mayoral et al., 2022). Examples from this “Human Factors Cyber Threat Taxonomy” are “Bypassing Physical Security” (physical layer), “Employee and IT security mismatch” (social layer), “GDPR-related customer issues” (legal layer), or “Fake online reviews and review bombing” (ethical layer). Knowledge on and monitoring of human factor-related cyber threats also enables a

dedicated approach to cyber risk assessment (see Renwick et al., 2021). The practical process involved would be to estimate the probability of incidents from different types of threats and their potential impact (see the definition of cyber risks in FSB, 2018).

Additionally, there is also an avenue for improving cybersecurity training for general employees. A pedagogically founded and systematic cybersecurity competence training framework (see Griesbacher et al., 2020) allows to design training and awareness measures for specific skills, knowledge, and attitudes relevant for cybersecurity (e.g., digital information competence or incident handling). The competences also include elements connected to the monitoring of the vulnerability layers (e.g., the ability of organisation members to recognize threats) and they can be related to specific threats in the taxonomy. By documenting the cybersecurity training elements in reference to a competence catalogue (ibid.) and the threat taxonomy new avenues for monitoring necessary awareness and skills of organisation members, reporting of human factorrelated cybersecurity measures (e.g., for auditing or certification purposes), and cyber threat information sharing can be explored.

The domains briefly addressed in this conclusion play a crucial role for preventing and mitigating cybersecurity incidents. The presented interdisciplinary cybersecurity approach can be seen as a key element of a systematic and connected set of human factor-related cybersecurity tools for the advancement of cyber risk assessments, awareness and training measures. It also exemplifies the value of SSH perspectives to provide a dedicated human factors-oriented cybersecurity approach which also takes into account European values and fundamental rights.

References

- Baier, A. (1986). Trust and Antitrust. *Ethics*, 96(2), 231-260.
- Carlton, M., Levy, Y., & Ramim, M. (2019). Mitigating cyber attacks through the measurement of non-IT professionals' cybersecurity skills. *Information & Computer Security*, 27(1), 101-121.
- D'Arcy, J., Hovav, A., & Galletta, D. (2009). User awareness of security countermeasures and its impact on information systems misuse: a deterrence approach, *Information Systems Research*, 20(1), 79-98.
- European Union Agency for Cybersecurity [ENISA] (2021). *ENISA Threat Landscape 2021*, April 2020 to mid-July 2021. Retrieved from: <https://www.enisa.europa.eu/publications/enisa-threat-landscape-2021>
- Eckert, C. (2018). *IT-Sicherheit. Konzepte – Verfahren – Protokolle*. Berlin, Boston: De Gruyter.
- European Union (2012). *Consolidated Version of the Treaty on European Union and the Treaty on the Functioning of the European Union*. OJ C 326, 0001–0390. Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A12012M%2FTXT>.

- European Union (2012). *Charter of Fundamental Rights of the European Union*. OJ C 326, 391-407. Retrieved from: <https://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=celex%3A12012P%2FTXT>.
- Financial Stability Board [FSB] (2018). *Cyber Lexicon*. Retrieved from <https://www.fsb.org/2018/11/cyber-lexicon/>
- Floridi, L. (2018). Soft Ethics and the Governance of the Digital. *Philosophy and Technology*, 31, 1-8. <https://doi.org/10.1007/s13347-018-0303-9>
- Ganin, A. A., Quach P., Panwar M., Collier Z. A., Keisler J. M., Marchese D., & Linkov I. (2017). Multicriteria Decision Framework for Cybersecurity Risk Assessment and Management. *Risk Analysis*, 40(1), 1-27.
- Griesbacher, E.-M., Griesbacher, M., Rabel, P., & Renwick, R. (2020). *D6.4 Training Modules Compilation (II), Deliverable 6.4 of the Horizon 2020 project SOTER*. Retrieved from: <https://cordis.europa.eu/project/id/833923/results>
- Griesbacher, M., & Griesbacher, E.-M. (2022). *D6.7 Report on Training Actions (III), Deliverable 6.7 of the Horizon 2020 project SOTER*. Retrieved from: <https://cordis.europa.eu/project/id/833923/results>
- Hardin, R. (2002). *Trust and Trustworthiness*. New York: Russell Sage Foundation.
- International Organization for Standardization [ISO] (2018). *Information technology – Security techniques – Information security management systems – Overview and vocabulary (ISO/IEC 27000:2018)*. Retrieved from: <https://standards.iso.org/ittf/PubliclyAvailableStandards/index.html>
- Jones, K. (1996). Trust as an Affective Attitude. *Ethics*, 107(1), 4-25.
- Luhmann, N. (1980). Trust: A Mechanism for the Reduction of Social Complexity, in Luhmann, N., *Trust and Power*. New York: Wiley.
- Luhmann, N. (2017). *Trust and Power*, edited by Morgner, C. and King, M., Polity Press, Newark.
- Martens, B., Aguiar, L., Gomez-Herrera, E., & Mueller-Langer, F. (2018). The digital transformation of news media and the rise of disinformation and fake news. *JRC Digital Economy Working Paper 2018-02*. Seville, Spain: European Commission.
- McLeod, C. (2015). Trust. *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Retrieved from: <https://plato.stanford.edu/archives/fall2015/entries/trust/>
- Nickel, P. J., Franssen M., & Kroes P. (2010). Can We Make Sense of the Notion of Trustworthy Technology?. *Knowledge, Technology & Policy*, 23, 429–444.
- North K., de Carvalho A. B., Maria A., Durst S., Carvalho J. A., Gräslund K., & Thalmann S., (2019). Information and knowledge risks in supply chain interactions of SMEs. An exploratory study. *Proceedings 10th Conference Professional Knowledge Management*, Potsdam, 268-277.
- Renwick, R., Griesbacher, M., Griesbacher, E.-M., & Rabel, P. (2020). *D2.1 Mapping of human behaviour related threats and mitigation measures (I), Deliverable 2.1 of the Horizon 2020 project SOTER*. Retrieved from: <https://cordis.europa.eu/project/id/833923/results>

- Renwick, R., Jordan, E., Wadhwa, K., Venegas Mayoral, E., Segura Gonzalez, A., Griesbacher, E.-M., & Griesbacher, M. (2021). *D2.2 Mapping of human behaviour related threats and mitigation measures (II), Deliverable 2.2 of the Horizon 2020 project SOTER*. Retrieved from: <https://cordis.europa.eu/project/id/833923/results>
- Siponen, M., & Vance, A. (2010). Neutralization: new insights into the problem of employee information systems security policy violations, *MIS Quarterly*, 34(3), 487-502.
- Stelzer, H., & Veljanova H. (2017). *TRUESSEC.eu D4.2. Support Study: Ethical Issues*. Retrieved from: <https://cordis.europa.eu/project/id/731711/results>
- Stelzer, H., E. Staudegger, H. Veljanova, A. Haselbacher, S. Reichmann, M. Griesbacher, J.M. del Álamo, D. Guamán, & J. Kingsbury (2018). *TRUESSEC.eu, D7.2 Cybersecurity and privacy Criteria Catalogue for assurance and certification*. Retrieved from: <https://cordis.europa.eu/project/id/731711/results>
- Stelzer, H., & Veljanova, H. (2020). Developing an Ethical Compass for Big Data. *Big Data and Democracy*, Macnish, K., Galliot, J. (ed.), Edinburgh Scholarship Online. DOI: 10.3366/edinburgh/9781474463522.001.0001
- Schatz, D., Bashroush, R., & Wall, J. (2017). Towards a more representative definition of Cyber Security. *The Journal of Digital Forensics, Security and Law* 12, (2), 53-74.
- Tomaselli, V., Cantone, G. G., & Mazzeo, V. (2021). The polarising effect of Review Bomb, *arXiv:2104.01140v1*. Retrieved from: <https://arxiv.org/abs/2104.01140>
- Venegas Mayoral, E., Trujillo Donoso, L., Griesbacher, M., Segura González, A., Renwick, R., Rabel, P. (2022). *D5.2 White Paper on Cybersecurity standards, Deliverable 5.2 of the Horizon 2020 project SOTER*. Retrieved from: <https://cordis.europa.eu/project/id/833923/results>
- Whitman, M. E., & Mattord, H. J. (2018). *Principals of Information Security*, Cengage Learning, Boston, MA.

Computing Apples and Oranges? Implications of Incommensurability for (Fair) Machine Learning

Otto Sahlgren^{*[0000-0001-7789-2009]} and Arto Laitinen^{*[0000-0002-4514-7298]}

^{*} Tampere University, Tampere, Finland otto.sahlgren@tuni.fi

arto.laitinen@tuni.fi

Abstract. In this paper, we discuss value relations and questions of incommensurability and incomparability in the context of machine learning and fairness therein. We examine three stances and consider their implications for machine learning supported decision-making and the pursuit of fair algorithms using a hypothetical example from recruitment.

Keywords: fair machine learning, philosophy, incommensurability, incomparability, value theory

1 Introduction

The value of machine learning (ML) systems lies in part in their capacity to discover patterns in data, to detect subtle (dis)similarities between data items, and to provide decision-makers information to help them make good decisions. Consider ML-based recruitment systems, for example, which allow for precise comparison and ranking of candidates in terms of their merits and overall “hiring-worthiness”. However, to enable human decision-makers to make justified decisions, such systems should arguably track evaluative differences amongst candidates with respect to their “hiring-worthiness”, which raises questions of comparability and commensurability. Recruitment can also have ethical objectives aside choosing the candidate with the most merits, for example, fairness. Similar questions arise here, as one should be able to compare competing ways to improve algorithms’ fairness and to evaluate individuals’ claims to fair treatment. For example, in Finnish Law, an individual belonging to an underrepresented group can be selected from among applicants that have “roughly the same qualifications”, which raises the question of how “rough similarity” should be understood.

The present work is motivated by philosophical questions related to

(in)comparability. The nature of value relations is significant in both cases described above – namely, in cases where we compare options with the help of ML systems and where we compare different ML systems in terms of ethical value. These topics are discussed in philosophical debates on (in)comparability and value

(in)commensurability (see Chang, 1997; Elson, 2017; Hsieh, 2005) and recently also in the context of AI (see Dobbe et al., 2021; Fleisher, 2021; Goodman, 2021).

We consider the implications of three philosophical accounts of (in)commensurability and (in)comparability for the abovementioned cases.

The structure of the paper is as follows. In section 1, we review three philosophical accounts of (in)comparability and (in)commensurability and consider their implications for ML-supported decision-making. In Section 2, we consider further implications for building fairness-sensitive algorithms, a topic discussed under the umbrella of “fair ML”. We distinguish three cases where incommensurability and incomparability pose both theoretical and practical challenges: conducting positive discrimination (or “affirmative action”) with algorithms (section 2.1.), implementing individual fairness measures (section 2.2), and addressing trade-offs between statistical operationalizations of fairness (section 2.3). Throughout the paper, we use the context of ML-supported recruitment to illustrate the relevant questions and challenges.

2 Machine-Supported Decisions, Incomparability, and Incommensurability

The Trichotomy Thesis of value relations states that one of three relations – worseness, betterness, or equality – holds between any two items (e.g., options, value-bearers, and preferences) that are comparable. Debates on (in)commensurability and (in)comparability revolve around paradigmatic cases where we find ourselves amiss when determining which relation obtains, if any. For example, can Mozart and Michelangelo be compared in terms of creativity? If so, is the value relation in question captured by the trichotomy, or is there another relation?

The terms “incommensurability” and “incomparability” are used with slightly different meanings in the general debate (see Hsieh & Andresson, 2021). The seeming failure of comparison to yield any positive result in terms of betterness, worseness, or equality is called “incommensurability” by some and “incomparability” by others (and some use these terms interchangeably); we will not follow either of these usages. We will use the term “incommensurability” to refer to two things lacking a common measure (i.e., cardinal measuring is not possible) and “formal incomparability” to refer to two items lacking a covering value with respect to which they can be compared (e.g., the number 54 cannot be compared to the color green in terms of tastiness). We call the cases where two items are formally comparable, but neither is better, and yet they are not exactly equal, “puzzle cases”.

We maintain that cases of formal incomparability are rather rare. In standard cases discussed in the literature, there is often a covering value (perhaps non-cardinal) that allows for formal comparability even in the puzzle cases (e.g., creativity in the case of Mozart and Michelangelo). Even though one can admit the covering value applies to both (Mozart and Michelangelo are arguably creative, while perhaps not exactly equally creative), the outcome of comparison remains puzzling.

We illustrate these questions with the following example:

A technology company wants to hire an ethicist. The chosen recruit should be “hiring-worthy” (i.e., possess a set of attributes, merits, and skills that make them a good technology ethicist and employee). “Hiring-worthiness” is the “covering value” here with respect to which the candidates are compared in multiple “component respects” (e.g., different skills and attributes). An initial screening shows that three roughly similarly merited candidates – Alex, Bill, and Connie – stand out from the crowd. Alex and Bill are academic philosophers with well-cited publications in technology ethics. Connie is a computer scientist with experience from ethical development in top companies in the industry. The recruiters decide to use a ML system to precisely rank the candidates. The system suggests that Alex is slightly better than Bill – perhaps, because Alex has published one paper more than Bill. Before the recruiters proceed to compare Alex and Connie, they ponder: is it even possible to rank the two? They have different degrees and backgrounds. Academia differs drastically from industry. Would they be comparing “apples and oranges”? Could the system, even in principle, help settle the choice between Alex and Connie?

The previous hypothetical is analogous to Derek Parfit’s (1987) famous example of comparing two poets and a novelist in terms of their literary merits which Parfit used to challenge the Trichotomy Thesis. In our case, if the Trichotomy Thesis is true, Alex’s being equal to Connie and Bill’s being worse than Alex would imply that Bill is worse than Connie (due to transitivity). But should one publication make the difference? The Thesis also implies that Alex and Connie are either equal or one is better. But how would we determine this in puzzling cases which involve apples and oranges, poets and novelists, or philosophers and computer scientists?

In the next three subsections we discuss three different responses to the puzzle cases. Some accounts deny the existence of such cases, some suggest they can be explained without rejecting the Trichotomy Thesis (Regan, 1977; Elson, 2017), and some grant the existence of a fourth relation, such as “parity” (Chang, 2002), “rough quality” (Parfit, 1987), or some type of “indeterminacy” or “imprecision”. These are all loaded terms with their own connotations, and we will try to do justice to each while maintaining some terminological clarity. We will discuss the implications of each stance for ML-supported decision-making.

2.1 Eliminativism

It might seem that neither Alex or Connie is better than the other, or that nor are they equally “hiring-worthy”. So-called *eliminativists* would maintain that this is a mere illusion. They argue that no two options objects are ever “apples and oranges” in a fundamental sense, and that the Trichotomy Thesis is true of any two items (Regan,

1997). Our epistemic limitations are the source of any apparent indeterminacy in comparison and ranking. The eliminativist would thus claim that the relations between Alex, Bill, and Connie in our example can all be accounted for with the three standard relations: each candidate is either better than, worse than, or exactly equal to another. For the eliminativist, the ML system can help determine which relations hold between the candidates provided that the relevant concepts (“hiring-worthiness”) are clearly operationalized in the language of mathematics and that the model is good at tracking relevant the recruiters’ preferences (e.g., the value contributed by publications). The eliminativist has no principled objection against ML systems’ capacity to rank Alex and Connie *correctly* with respect to “hiring-worthiness” (even though existing systems might be limited in many contingent ways).

The eliminativist view has been criticized for many reasons, however. For one, epistemic limitations (e.g., ignorance) might not exhaust the problems with “puzzling cases”. The parameters for choosing between two options are for the decision-maker to decide, the choice can yet be a “hard choice” (Chang, 2017). For example, one has firstperson authority in making the hard choice between a career in academic philosophy or in software development. Both are desirable in their own ways, yet they would still seem comparable in light of their contributions to a good life. Furthermore, while MLsupported decision-making is premised on the notion that value-bearers and preferences can in principle be mathematically represented and that human goals and tasks can be translated into computational problems with reward and loss functions, it might be that all “goals and purposes simply cannot be represented as the maximization of the expected value of a scalar reward” (Goodman, 2021, 5). Perhaps, “hiring-worthiness” does not lend itself to simple and exact quantification. ML systems would seem to provide a sense “quasi-precision” when dealing with such concepts and translating human goals into system specifications, respectively (see Dobbe et al., 2021).

2.2 “Parity”

Alternative accounts draw on the idea of there being “neighborhoods” of value and suggest that the Trichotomy Thesis does not capture the entire scope of value relations. Perhaps most influentially, Ruth Chang claims that “between two evaluatively very different items” (e.g., Alex and Connie, a poet and a novelist), “a small unidimensional difference cannot trigger incomparability where before there was comparability” (2002, 673). This claim is motivated with the so-called “Chaining Argument”: Suppose there is another candidate Don, a philosopher similar to Alex and Bill but clearly worse than both in all component respects (e.g., publication, work experience, and so on). If Alex and Bill are in the same neighborhood of value as Connie, Don should plausibly be worse than Connie. If this is the case, however, Connie is comparable to Don and thereby also comparable to Alex and Bill because the philosopher candidates form a “chain”, a sequence from worse to better philosophers. Hence Alex, Bill, and Connie are formally comparable. However, the kicker comes in Chang’s suggestion that our best philosopher, Alex, is not exactly equal to Connie. They are rather “on a par”: formally comparable, not exactly equal,

but neither is better than the other¹. If they were exactly equal, “sweetening” Alex by making any small improvement (e.g., one more publication) would make sweetened “Alex+” better than Connie. When we compare Alex and Alex+, the latter is better. Nonetheless, Chang suggests that sweetening would not make a difference when it comes to Alex+ and Connie – they are not equal but Alex+ is not better. Both Alex and Alex+ are “on a par” with Connie.

Chang suggests “parity” deviates from the standard trichotomy of comparative relations. Whether it actually does is debated (see Hsieh & Anderson, 2021), but for present purposes we will stay truthful to Chang’s suggestion that it does. Now, if

“parity” is a true comparative relation, ML systems should arguably be able to track such relations were they to estimate value relations, support human judgment, and provide justification for decisions. A challenge looms for existing ML methods, however, because parity is an intransitive relation: Alex is on a par with Connie, who is on a par with Alex+, but Alex is not on a par with Alex+ but is worse. The traditional trichotomy (betterness, worseness, and equality) is still there, however, because the fourth alternative is only added for puzzling cases (with apples and oranges). Regardless, allowing for “parity” in ML would require a way to implement this fourth value relation. For example, the output ranking would exhibit some type of a partial or perhaps incomplete ordering where a given rank (e.g., the top-*k* candidates) can contain items that are in a relation of “relaxed” or “imprecise” equality (because they are “on a par”) as well as items that are exactly equal (because they are evaluatively identical).

2.3 “Clumpiness”

The idea of “parity” does not please all theorists as it requires accepting the intransitivity of some value relations. In the ML setting, this can also prove challenging for precise ranking, for example. Is there a way, then, to accept the Trichotomy Thesis and thus avoid this possible issue without simultaneously falling into the eliminativist trap of “quasi-precision”? Hsieh (2005) suggests so, claiming that some values can be “clumpy”. For example, the property of “hiring-worthiness” could be understood as a range property² within which we find “clumps” of “hiring-worthiness” similar to the clumps of “excellence” that different grades seek to track as evaluative categories for student coursework. For Hsieh, the number of clumps depends on the “resolution” of comparison which “specifies the degree to which possession of each of the relevant respects of the covering consideration sorts an item into one clump or another” (2005,

¹ Derek Parfit (1987) has suggested similar view according to which the items can be “roughly equal”. The difference between “rough equality” and “parity” is debated: according to one understanding, the former is “invoked to allow for comparability among alternatives that display the same respects” and the latter “to allow for comparability between alternatives that are different in the respects that they display” (Hsieh & Anderson, 2021, S2.2).

² A range property is a property which comes in degrees (e.g., a continuous scale along which to measure it) yet with respect to which we can specify a range and an object can either fall into that range (or not) in a binary sense. For example, an essay that is perfect in all evaluated respects falls perfectly within the range of essays worthy of the best grade. However, another essay that lacks slightly in certain respects can yet be equally worthy of the best grade.

184). Now, whereas Chang would suggest that items in a “clump” are perhaps rather “on a par”, Hsieh suggests they are in fact exactly equal. This has the happy consequence that the Trichotomy Thesis can be preserved without having to accept the eliminativist view that equal candidates must have the same real-valued “hiringworthiness”. Due to the fixed resolution of comparison, the “top-notch” candidates, for example, are exactly equal in terms of hiring-worthiness. Yet this does not prevent one from scoring items (including candidates) on continuous scales nor “binning” them into clumps different ways. Different resolutions can serve practical aims: one can start with a coarser resolution of two clumps (e.g., “top-notch” and “unsuitable” candidates), and by increasing the resolution one can create new clumps (e.g., “moderately suitable”).

Note that all comparisons need not be clumpy – there is room for totally ordered rankings and ML models which seek to track precise comparative value relations and generate corresponding rankings, respectively. Furthermore, even in cases where a target variable *is* supposed to track a clumpy value (“hiring-worthiness”), a continuous target variable provides one resolution of comparison, albeit a very fine-grained one. With clumpy values (e.g., hiring-worthiness), human decision-makers can build clusters, partial orderings, or output classes based on the real-valued total ranking of candidates, similar to how a teacher can determine which individuals should receive the best or worst grade after first observing and ranking a set of student essays. The important implication for ML models is that, when the target variable ought to track a clumpy value, the number and boundaries of the clusters, output classes, “bins”, or ranks employed to evaluative items should reflect the appropriate resolution and outcome of (correct) evaluative judgment. The set of top-*k* candidates recommended to the recruiter, for example, should track the clump of “top-notch” candidates.

The key differences between eliminativism, and the “parity” and “clumpiness” views can thus be stated as follows: An eliminativist using the recruitment system would consider two candidates with different predicted values (e.g., Alex is 0.9 hiring-worthy, Bill is 0.89 hiring-worthy) as either better or worse than each other (as the ranked output would suggest). Were the eliminativist to have an ideal classifier, they could trust that the output ranking tracks the comparative value relations between the ranked items. A proponent of “parity” relations could consider Alex and Bill as unequal, whereas Alex would be on a par with Connie. Proponents of this view would thereby have to rely on partial order rankings. A proponent of clumpiness would require a similar approach. However, they would suggest that all candidates belonging to the same neighborhood of overall “hiring-worthiness” (e.g., the top-*k* candidates) should be viewed as exactly equal in the evaluative sense (despite the 0.1 difference between Alex and Bill). This is because two candidates having the same features (e.g., identical merits) is necessary and sufficient for exact equality in a descriptive sense, yet not necessary for exact equality in the evaluative sense (albeit sufficient). A ranking of candidates *qua* feature vectors can be totally ordered, respectively, but an evaluative ranking of those candidates depends on the resolution which in turn specifies the clumps wherein candidates are evaluated as exactly equal. In other words, whereas “parity” would require a partially ordered ranking which allows for

both equality and “parity” within a given rank, rankings with respect to clumpy values should have to allow for exact equality within a given rank in a way that does not imply numeric identity in real-valued outputs.

3 Fair Algorithms: Positive Discrimination, Similarity, and “Hard Choices” Concerning Fairness

The previous cases concerned comparability within the context of a single ML model seeking to track “hiring-worthiness”. Often, however, ML-supported decision-making is guided also by ethical considerations related to fairness, for example. In this section, we consider three cases where questions of incommensurability and incomparability pose not only theoretical but practical challenges for designing algorithms in a fairness-sensitive manner. The first case concerns positive discrimination and its permissibility in (non-)automated recruitment. The second case concerns the measurement and implementation of individual fairness in ML systems. The third case concerns choices regarding trade-offs between multiple fairness targets in ML.

3.1 Imprecise Equality and Positive Discrimination with Algorithms

Recruitment policies can purposefully seek to promote disadvantaged groups’ access to employment. One instrument for doing so is *positive discrimination* (or “positive action” or “affirmative action”) where candidates from underrepresented protected groups (e.g., women) are favored over candidates from overrepresented groups (e.g., men). If applied with care, so-called “bias mitigation methods” developed for correcting discriminatory biases in ML systems (Mehrabi et al., 2021) could prove useful for such purposes in ML-supported recruitment as well. Importantly, however, justification of positive action often requires, among other things, that the selection process does not give a merely arbitrary or “disproportional advantage to members of the relevant group. In Finland, for example, the Non-Discrimination Act requires that all candidates are initially treated on an equal basis, and states that “an individual belonging to an underrepresented group can be selected from among applicants that have roughly the same qualifications” (Non-Discrimination Ombudsman of Finland, N.D.). As noted above, there is notable disagreement regarding how “rough sameness” should be interpreted. We will first consider what the formal setting of positive action implies for the use of bias mitigation methods after which we consider the notion of “rough sameness” through the theoretical lenses described in the previous section.

Positive Action with Algorithms

What kinds of bias mitigation methods would capture the “spirit of positive action” as described above? We suggest that at least three formal conditions need to be satisfied

for a hiring decision to be considered “positive action” (in a “neutral” sense in which we can still ask whether it is justified):

- (*Formally Equal Assessment*): The selection process involves a formally equal assessment of a candidate from an underrepresented group (C_{UNDER}) and a candidate from an overrepresented group (C_{COVER}); their “hiring-worthiness” is assessed based on identical criteria. This implies the use of a single model with features representing component factors of “hiring-worthiness” and a single “cutoff” point (i.e., a decision-threshold).
- (*Absolute Sufficiency*): The chosen candidate is considered sufficiently hiringworthy in an independent, non-comparative sense. Depending on the case, the level of absolute sufficiency (i.e., the decision-threshold) can be decided either prior to or after ranking the relevant candidates.
- (*Imprecise Equality*): The favored candidate C_{UNDER} should be “roughly equal” in terms of their qualifications when compared to non-favored candidate C_{COVER} .

In other words, regardless of its justification, for recruitment to instantiate positive action at all, candidates C_{UNDER} and C_{COVER} should be “roughly equal” and “in the neighborhood” of what is required from each candidate in a non-comparative sense. Absolute Sufficiency is to be assessed with respect to a single covering value and by employing identical criteria that (hopefully) track the component factors of that value. For example, C_{UNDER} “making the cut” with lesser-than-sufficient qualifications would not qualify for positive action because it would not satisfy Absolute Sufficiency nor Imprecise Equality. Note that, in cases where Absolute Sufficiency has necessary conditions (call these *Criteria-First Cases*), a candidate that satisfies those conditions is plausibly “clearly better” than one that does not. In other cases, such as when the level of absolute sufficiency can be set only after ranking the candidates, the best candidate takes the spot (call these *Ranking-First Cases*). In Aggregation Cases, C_{UNDER} and C_{COVER} need not be on the same side of the decision-threshold, however, even though the recruiter cannot choose C_{UNDER} if C_{COVER} is somehow “clearly better”.

Positive action involves taking protected attributes into account in selection after the initial assessment has taken place. Consider now, that there are at least three general different ways to incorporate information about protected group-membership into MLsupported decision-making (see Hellman, 2019; Mehrabi et al., 2021). First, one might use different decision-thresholds for members of different protected groups within the model. Second, one might use different models for those groups entirely. Third, one could change the output labels of members of the underrepresented group from negative to positive when they are close to the decision-threshold (Kamiran et al., 2012).

We suggest that none of these methods satisfy all three conditions described above. Consider the first approach. Here, recruiters would employ similar component factors for comparison but with different “cut-off points” for C_{UNDER} and C_{COVER} . The level of Absolute Sufficiency would thereby differ across groups, implying that different weights are given to component factors. If so, the condition of Formally Equal

Assessment is violated. In the second approach, “hiring-worthiness” would be evaluated with different models for C_{UNDER} and C_{COVER} . This means the covering values used for evaluation are non-identical and the Formally Equal Assessment condition is again violated³. The third approach comes closest the formal setting assumed in cases of positive action: candidates are evaluated based on the same model and there is a single decision-threshold. However, it lacks a way to resolve cases where there are two candidates – C_{UNDER} and C_{COVER} – and a single available position. Positive action will be at best an artefact of choosing C_{UNDER} as opposed to an explicit aim implemented in the method’s processing logic in a strict sense. That is, the comparative dimension of positive action – “if two candidates are within a range R , choose C_{UNDER} over C_{COVER} ” – remains uncaptured by the method.

Positive Discrimination and “Rough Sameness”

Let us now consider Ranking-First Cases to examine how theoretical stances concerning comparative relations bear on (the possibility of) positive action. First note that eliminativists and other proponents of the Trichotomy Thesis often consider a choice permissible only if there are normative reasons to choose A over B (e.g., A is better) or if A and B are exactly equal. For them, the boundaries of “rough sameness” with respect to qualifications will remain rather arbitrary and thus they would consider it irrational to choose C_{UNDER} over C_{COVER} in case the former is ranked below the latter. The eliminativist can, of course, concede the plausible claim that there are other normative reasons (e.g., the value of equity) that override rational choice based on qualifications alone.

A proponent of clumpy values could contend, however, that “roughly the same qualifications” merely means that C_{UNDER} and C_{COVER} have to belong to the same clump of “hiring-worthiness”. As C_{UNDER} and C_{COVER} would hence be exactly equal, choosing either one of them is permissible in light of rational choice based on qualifications alone (contra the eliminativist) and the instantiation of positive action is merely an artefact of choosing C_{UNDER} . However, one could question whether “rough sameness” should actually be understood as an even softer requirement: could it not suffice that C_{UNDER} belongs merely to the clump below the one including C_{COVER} ?

Alternatively, if “rough sameness” is what Chang means by “parity”, positive action is possible when C_{UNDER} and C_{COVER} are “on a par” – neither candidate is actually better than the other, nor are they exactly equal. Both “parity” and “clumpiness” can thereby lead to the conclusion that one never hires a *de facto* worse candidate when implementing positive action.⁴ While we do not discuss these issues

³ An open question is, however, whether the non-identical models can be commensurate as models of overall “hiring-worthiness”.

⁴ If parity obtains between two options, one will lack so-called “given reasons” (i.e., reasons grounded in normative facts) for choosing between them. One has no *prima facie* reason to favor either, making the decision a “hard choice” (Chang, 2017). Chang argues that normative commitments can create “will-based reasons” that function as tie-breakers. Positive action could thus be understood as a normative commitment to equality that creates a reason for choosing between candidates that are “on a par”. *Mutatis mutandis*, the same holds for the clumpiness view.

further, we note that each stance has implications for how positive action ought to be understood as a non-moralized category of acts, and how it should be implemented in ML-supported decision-making.

3.2 Tracking Evaluative Features: The Case of Individual Fairness

Individual fairness (IF) as an approach to algorithmic fairness draws on the principle of formal equality: fairness requires treating similar individuals similarly (Dwork et al. 2012). Individual fairness is measured with similarity-metrics that estimate the similarity (or distance) between individuals in terms of some set of attributes which typically exclude an individual's protected status, for example. If individuals who are similar according to the metric receive different outputs, the algorithm is considered unfair.

Similarity metrics ought to track moral values and (dis)similarities which are relevant from the perspective of fairness. Will Fleisher (2021) considers incommensurability to pose a challenge because “a similarity metric requires that it be possible to aggregate the moral values, or evaluate them together, in a straightforward way” (2021, 17). Fleisher notes that some “moral values are incommensurable” and “cannot be evaluated on a common measure, i.e., they cannot be straightforwardly aggregated or exchanged” (Ibid., 3). Indeed, “similarity” in IF approaches is ambiguous between exact, “descriptive” similarity (being qualitatively identical) and similarity that is relevant from the perspective of fairness and ethics. However, the relationship between these two types of similarities can be complex – a (dis)similarity of the former kind may or may not equate to (dis)similarity of the latter kind. As we considered above, it might be that “sweetening” one candidate would not render them better than another candidate, even though the descriptive distance between them would increase as a result (see Fleisher, 2021). Likewise, if values are clumpy, for example, a small difference between two “top-notch” candidates will not make a relevant difference in terms of fairness.

Estimating “fairness-relevant” distance and possible puzzles in how evaluative features behave (e.g., “parity” and “clumpiness”) requires human value judgements. Fleisher notes, however, that appealing to human arbiters in evaluating similarities between individuals, while promising, can be problematic due to “implicit biases in [human] judgment” (2021, 2). Fleisher is correct, but the cure may nonetheless be adding more human arbiters, and the hope that different humans have different biases. In most cases, cognitive and socio-cultural diversity among those arbiters can be beneficial to ensure “diversity of biases” and that the arbiters arrive at correct judgments regarding how similarity ought to be measured. Even if we accept that measurements of similarity will always reflect prior moral judgments and biases (Fleisher, 2021, 2), it is fine (and inevitable) that judgements about fairness rely on evaluative or moral judgements insofar as such an equilibrium is achieved.⁵

⁵ Fleisher also considers a possibility we already discussed above: using partial order rankings instead of (or in addition to) similarity-metrics.

3.3 “The Impossibility of Fairness” and “Hard Choices”

A final challenge relates to trade-offs in implementing multiple operationalizations of algorithmic fairness. As Dobbe and colleagues note, “normative concerns of comparable significance and scope must be rendered commensurable in order for a responsible tradeoff to be struck and translated to a system’s specification” (2021, 4). However, trade-offs between different fairness criteria are a prime example of cases where we seem to be lacking such commensurate options. Various fairness definitions prescribe equalizing some statistical metric (e.g., positive predictions, error rates) across individuals or groups in the model (Verma & Rubin, 2018). But many of these metrics cannot be equalized simultaneously except in highly contrived cases (Kleinberg et al., 2017). Trade-offs are taken to suggest “the impossibility of fairness” due to fairness definitions’ representing irreconcilable moral views.

To make justifiable decisions concerning trade-offs, we should be sensitive to reasons for choosing one fairness metric over another. However, some consider the decision regarding proper measurement and implementation of fairness “a hard choice” in virtue of the options being incommensurate: “certain alternatives are neither better, worse nor equal to one another with respect to fairness” (Goodman, 2021, 7). Rival options are rather “on a par”. Value incommensurability thus poses a challenge for justifiably deciding which fairness criteria to implement in algorithms. But if such a decision requires choosing between incommensurate options, what do we do?

Goodman’s proposal draws on Chang’s (2017) view according to which “will-based reasons” created through normative commitment can function as tie-breakers:

“[n]othing in the world will tell us the correct answer” regarding proper measures of algorithmic fairness; “[i]nstead, we must *commit*” (Goodman, 2021, 7). Dobbe and colleagues (2021) propose another solution. Recognizing that ML systems typically affect numerous stakeholders with different interests, they suggest that incommensurability and hard choices become *political* issues as AI systems’ normative capacities cannot be evaluated and measured by the same standards by different stakeholders. We note that these solutions are not mutually exclusive – whether it is individual persons or collectives making significant decisions, the cold reality of compromises is equally faced by both. Extending Chang’s solution to “hard choices” between incommensurable options, one could argue that collective wills can create tiebreaking “will-based reasons” through normative commitment in a manner analogous to the individual case.

If trade-offs are exclusively “hard choices”, there are no normative facts or secondorder principles concerning distributive fairness that provide reasons to resolve tradeoffs in a manner *R* purely because *R*-ing is the right thing to do. We should resort to proceduralism (e.g., fair democratic decision-making procedures) or will-based reasons (e.g., resolution through normative commitment), for example. While these might be independently justifiable and desirable approaches, we note that equating trade-offs with Changian “hard choices” denies the possibility of there being “given” normative reasons that speak in favor of resolving trade-offs in one way over another. For instance, while a fair democratic process might lead to the decision to choose the

option that creates the largest benefit to those who are worst off, one could also argue that prioritarian regard should guide choices concerning trade-offs because it is what fairness requires.

Furthermore, it is noted in some works that certain fairness definitions are in principle compatible albeit practically in conflict due to contingent states-of-affairs (e.g., differences in the distribution of attributes across subpopulations) (Binns, 2020).

Some trade-offs can thus arise “merely” in virtue of our present, contingent circumstances. This suggests, however, that they can be resolved in the long-term, at least in principle. We would need a “transitional approach” and incremental improvements to get to a place where multiple fairness definitions can be simultaneously satisfied. If this is true, it leaves open the possibility that different fairness metrics for algorithms may in theory be commensurate with respect to a covering value: Fairness or Justice *qua* an ethico-political value. Different operationalizations of fairness in ML could be understood as comparable *qua* factors that contribute to a multidimensional covering value understood as overall fairness (or not). Competing fairness measures are applicants for the job of the best conception of Fairness (or at least one component of it), as it were. As we are not ideal judges with direct access to ideals of Justice and Fairness, we do not merely choose or commit to any of the candidates. We test them, build new ones, compare their pros and cons, and hopefully end up with the best conception of algorithmic fairness so far. Ultimately, the right approach to resolving trade-offs will be one that best reflects the constitutive aim of conceptions of algorithmic fairness; what “Fairness with a capital F” or “Justice with a capital J” in fact requires. Insofar as feasibility constraints and long-term effects might prevent us from achieving the best solution straight away, for the time being, one might have to settle for the second-best.

4 Conclusions

In this paper, we sought to motivate questions related to incommensurability, incomparability, and “hard choices” in the context of ML-supported decision-making. We reviewed three stances on comparability and value relations, discussing their implication for ranking and ordering items with ML. Each stance, we suggested, has implications concerning whether and how ML-generated rankings can track value relations. We also discussed fairness in ML, noting how similar puzzles arise in the context of building fair algorithms. Here, we located challenges relating to positive discrimination and how it might be pursued with algorithms and for determining whether and how to implement one or several fairness measures in ML systems. While the fundamental questions concerning the nature of value relations remain unsolved, we hope to have shed light on their practical significance for designing and using ML systems, suggesting also possible ways forward.

References

- Binns, R. (2020). On the apparent conflict between individual and group fairness. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 514-524.
- Chang, R. (1997). Introduction. *Incommensurability, Incomparability, and Practical Reason*, R. Chang (ed.), Cambridge: Harvard University Press.
- (2002). The Possibility of Parity. *Ethics*, 112, 659–688.
- (2005). Parity, Interval Value, and Choice. *Ethics*, 115, 331–350.
- (2017). Hard choices. *Journal of the American Philosophical Association*, 3(1), 1-21.
- Dobbe, R., Gilbert, T. K., & Mintz, Y. (2021). Hard choices in artificial intelligence. *Artificial Intelligence*, 300, 103555. <https://doi.org/10.1016/j.artint.2021.103555>.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214-226.
- Elson, L. (2017). Incommensurability as Vagueness: A Burden-Shifting Argument. *Theoria*, 83 (4), pp. 341–363.
- Fleisher, W. (2021). What’s Fair about Individual Fairness?. *AIES '21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 480–490.
- Goodman, B. (2021). Hard Choices and Hard Limits for Artificial Intelligence. *arXiv preprint arXiv:2105.07852*.
- Hellman, D. (2020). Measuring algorithmic fairness. *Virginia Law Review*, 106(4), 811-866.
- Hsieh, N. H. (2005). Equality, clumpiness and incomparability. *Utilitas*, 17(2), 180-204.
- Hsieh, N-H. & Andersson, H. (2021). Incommensurable Values. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.). <https://plato.stanford.edu/archives/fall2021/entries/value-incommensurable/>. Accessed 10.5.2022.
- Kamiran, F., Karim, A., & Zhang, X. (2012). Decision theory for discrimination-aware classification. *2012 IEEE 12th International Conference on Data Mining*, 924-929. IEEE.
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6), 1-35.
- Non-Discrimination Ombudsman of Finland. (N.D). *Positive action*. [Website]. <https://syrjinta.fi/en/positive-action>. Accessed 9.5.2022.
- Parfit, D. (1987). *Reasons and Persons*. Oxford: Oxford University Press.
- Regan, D. (1997). Value, Comparability, and Choice. *Incommensurability, Incomparability and Practical Reason*, In R. Chang (ed.). Cambridge: Harvard University Press.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), 1-7. IEEE.

Trust and Explainable AI: Promises and Limitations

Sara Blanco¹

¹ Eberhard Karls Universität Tübingen, Tübingen, Germany
sara.blanco@uni-tuebingen.de

Abstract. Trustworthiness is widely quoted as a key property to enabling the effective deployment of AI. However, it is not obvious how to achieve it. The literature often assumes that explanations lead to trust. This claim has been referred to as the *Explainability-Trust Hypothesis* (ET). The link between trust and explanations is complex, and I argue that taking ET for granted is problematic. By pointing out the epistemic limitations of ET, I aim to shed light on crucial aspects of trustworthy AI: which kind of trust it requires and which kind of explanations can help to achieve it. This clarification will help us to understand until which point it is fair to establish a connection between trust and explanations.

Keywords: Trust, Explainability, Artificial Intelligence.

1 Introduction

Trustworthiness is widely quoted as a key property to enabling the effective deployment of AI (HLEGAI, 2019; Leslie, 2019). The main problem with AI is that the most successful systems that have been developed so far are neural networks (NN). Neural networks are computing systems in which the information is processed by going through nodes organised in layers. These nodes produce non-linear equations as an output, which will go through further nodes until reaching the output layer, which provides a result. Because of their computational complexity, this kind of systems are black boxes.¹ That is, a specific outcome cannot be traced back to its correspondent input, and it is not comprehensible to humans how the system arrived from the former to the latter. Against the backdrop of this situation, arises the dilemma of whether we can (or should) trust technology whose workings are not fully understood. A possible solution to it lies in explainable AI (XAI). For the purposes of this article, I use XAI in

¹ By black boxes I understand systems that are functional, structural and/or run opaque (Creel, 2020). Different authors (Erasmus et al., 2020; Krishnan, 2020; Pérez, 2019; Zednik, 2021) identify different types of opacity and which of them are more undesirable for different purposes. Here I use the notion of black box in a broad sense, considering that any system whose outputs cannot be explained or made sense of, is a black box.

a wide sense, including both the systems that are interpretable by design and the ones to which an additional system has been applied in order to provide explanations (e.g. LIME, Grad-CAM). The idea behind AI is that providing explanations for the results of a system will make it more trustworthy. This has been called the *Explainability-Trust Hypothesis* (ET) (Kästner et al., 2021). However, the relationship between explanations and trust is complex, and I argue that taking ET for granted is problematic.

The main goal of this paper is to clarify under which circumstances ET holds and to point out its limitations. I will argue that even though explanations may sometimes lead to trust, they do not guarantee it (contrary to what the literature suggests). My analysis focuses on epistemological problems of ET that are usually overlooked and need to be taken into account when using explanations as a way to achieve trust. Although it has been noted that ET expresses an epistemological claim, and not an empirical one (Kästner et al., 2021), the details of this epistemic nature and its implications have not yet been explored. I elaborate on this issue with the aim to uncover how trust can be achieved.

In the literature, it is often assumed that ET has a universal character. There are two sides to this assumption: that every truster reacts equally to the same explanation, and that every explanation changes the beliefs of the truster. I will show that both assumptions fail. Because ET is formulated in a simple manner, neither the target truster or the kind of explanations that it refers to are specified. This vagueness makes ET epistemically limited. According to ET, trust can be reached through explanations. That is, explaining to the truster how a system works will provide the truster good, justifying reasons to believe the system is trustworthy. However, only some explanations in some circumstances will achieve this goal. Before attempting to ensure trust through explanations, as suggested by ET, we need to clarify how explanations ensure trust, and most importantly, how explanations ensure the appropriate trust that is demanded in the AI field.

The paper proceeds as follows: In section 1, I present ET, its background and its influence in the literature. In section 2, I point to the epistemic limitations of ET. I show that the connection between explanations and trust is contingent and dependent on the epistemic context of the truster. In section 3, I make my case against ET and explain how its limitations lead to its rejection. In section 4, I briefly conclude.

2 Background and formulation

Explainability is the property of a system whose outputs can be explained. That is, it is possible to provide explanations that allow humans to make sense of the workings of the system. Explainable AI (XAI) is opposed to the concept of black-box, which refers to opaque systems whose workings are not understandable for human beings.

The *Explainability-Trust Hypothesis* (ET) refers to the idea that explanations lead to trust. Following (Kästner et al., 2021):²

² Definition 1 is a literal quote from Kästner et al. (2021)'s paper, except for the words “potential truster”, which I have used instead of the original “stakeholder”. The

Explainability-Trust Hypothesis: Explainability is a suitable mean (1)
for facilitating trust in a potential truster.

In other words, that the trust of a potential truster can be achieved through explanations. For example, imagine a doctor who is first presented with a new tumor detection NN. They may distrust the system because they ignore how it works. But after receiving an explanation, if the doctor is convinced and approves the NN mechanism, they will believe that the NN is trustworthy.

ET is a very intuitive claim that has been assumed, implicitly and explicitly, by many researchers. In the *Ethics guidelines for trustworthy AI* (HLEGAI, 2019), explicability³ appears as a key ethical principle. We could say, that their version of ET reads as follows: “Explicability is crucial for building and maintaining users’ trust in AI systems” (p. 13).

Parallely to (Kästner et al., 2021), I point to formulations of the same idea:

- (a) “We argue that explaining predictions is an important aspect in getting humans to trust and use machine learning effectively, if the explanations are faithful and intelligible” (Ribeiro et al., 2016, pp. 1135–1136).
- (b) “[...] one of the main goals of explanation is to establish trust of people [...]” (Miller, 2019)
- (c) “The big advantage of such systems [*explainable systems*] would include not only explainability, but [...] Most of all, this would increase acceptance and trust [...]” (Goebel et al., 2018, p. 297).
- (d) “[...] there is a need to explain [...] so that users and decision makers can develop appropriate trust [...]” (Hoffman, Klein, et al., 2018, p. 197).

We can see that associating explainability with an increment of trust is common in the literature. However, the relationship between trust and explanations is complex. It has been noted that ET expresses an epistemic relationship and not an empirical one (Kästner et al., 2021). Despite being a keen observation, Kästner et al. do not go deeper into it. My reading of ET is that it is a claim that does not connect empirical facts. Instead, it talks about knowledge: how someone can come to trust a system (or how someone can get to know that a system is trustworthy).

Trust is a rich concept that has been prominently treated in philosophy during the last decades (Baier, 1986; Gambetta, 1988; Hardin, 2002; Hawley, 2014; Jones, 1996). Only concerning interpersonal scenarios, it opened a fruitful debate about its foundation and nature. Considering an artificial system as a possible trustee adds even more layers

reason for this choice is to widen the scope of ET. This way, ET can be applied also in scenarios which involve users, policymakers, developers, etc.

³ Authors such as Floridi et al. (2018) emphasize the distinction between explainability and explicability, considering the latter a wider concept. I will come back to this terminological note in section 2.2, but in general, it is not my aim to engage with this terminological debate here.

to the discussion. Because of this, a hypothesis as ET is highly ambitious: it proposes a course of action to achieve a very controversial concept such as trust. How does this course of action work? The answer lies again in the subject and their perception of the world. According to ET, if the truster counts with explanations about how the system works, they will understand the system and therefore they will trust it. This means, that ET captures a relationship between belief-states of the truster. That is the reason why ET is not as simple as it may seem. It says something about how a truster comes to trust, which is a subjective process. However, it offers an objective⁴ answer: explanations. This tension between subjectivity and objectivity is problematic. ET connects something objective (explanations) to something subjective (trust) without characterizing the subject. This leads to a false universalization, from which epistemic limitations follow. I proceed to unpack such limitations.

3 Epistemic limitations

So far, I have presented ET as an intuitive claim widely supported by the literature. However, a closer analysis of this hypothesis reveals nuances that are usually overlooked. In the literature, it is often assumed that ET has a universal character. There are two sides to this assumption: that every truster reacts equally to the same explanation, and that every explanation changes the beliefs of the truster. In this section, I show how both assumptions fail.

3.1 Trust and background knowledge

It is recurrent across the literature to state that explainability leads to an increment of trust in the system. However, it is usually not specified in which contexts this can happen. This omission contributes to the mental image of the ideal subject, who has all the previous knowledge needed to understand the explanation at hand. This way, the features and workings of explanations are discussed with not much attention paid to the target user. Guidotti et al. (2018) offer a good overview of the most popular methods for making AI explainable. The vast majority of them ignore the background knowledge of the user as a relevant factor to take into account when developing and evaluating explanations. Few studies can be found where it is specified that the explanations to which they refer are targeted at a certain demographic of the population. One of these rare examples are Ribeiro et al. (2016), who specify that only graduate students were selected for their study, precisely because previous knowledge of the field was necessary. In Hoffman, Mueller, et al. (2018)'s words, "what counts as an explanation depends on what the learner/user needs, what knowledge the user (learner) already has, and especially the user's goals" (p. 3). Background knowledge is a generally assumed requisite when talking about which kinds of explanations are effective and which are not in XAI.

⁴ Objective as opposed to subjective: an explanation is a set of sentences external to their receptor, not an internal state of them (as it is a belief, for example).

Taking for granted the background knowledge of the truster suggests by omission that every person reacts equally to the same explanation. Trust is a part of the reaction of the truster. Thus, one could infer that a common explanation should bring about the same trust in every truster. This is a dangerous assumption that ought to be qualified. It can be the case that the same explanation induces different belief changes depending on the epistemic background of the truster. For example, imagine a hospital in which it is being decided whether to use an AI to diagnose pancreatic cancer. After explaining that this AI is a NN and how it operates on the patients' scans, the doctors of the hospital are convinced of its benefits, while the patients remain indifferent. The doctors had the previous knowledge required to understand the explanations, but the patients did not even know what a NN is. Not every truster will have the same background knowledge. In the absence of such knowledge, the explanation will not alter the beliefs of the truster and hence will be ineffective.

Reflecting on the paradoxical relationship between trust and transparency pointed out by O'Neill (2002), Nguyen (2021) highlights the role of expertise in certain explanations. Nguyen argues that some explanations (or justifications) simply cannot be adapted to the general public. There are domains, to which Nguyen refers as cognitive islands, that require a high level of expertise to fully grasp. When justifying decisions in these areas, the reasons provided by the experts are not epistemically accessible to people outside the field due to their lack of background knowledge. This does not make the justification less valid. However, the experts may be pressured to find a more accessible alternative, even if that alternative is not the real reason behind the actions that are being justified (or even worse, the actions of the experts become tailored⁵ to the kind of justifications that they know would be accepted). Nguyen makes this point to bring to light the fact that the need for transparency can become surveillance, rather than trust. In what concerns this paper, I find that he makes a good point putting the focus on the existence of explanations that, despite being valid, are ineffective for some purposes; for example, the understanding of the general public. This affects ET directly since the idea is to use explanations as a mean to increase trust. However, ineffective explanations won't achieve this goal. In fields such as AI, finding effective explanations is challenging due to the need for expertise of the truster in different areas. I have shown that ET does not hold in every context. Paying attention to whom the explanations are aimed is key when the purpose is to achieve their trust.

⁵ Nguyen illustrates this point with several examples. A particularly clear one pictures the case of a philosophy department that is expected to perform assessment. That is, to prove that certain learning outcomes have been achieved. This could be evaluated by another philosophy department, that could rate writing samples of the students from the first department. However, a more transparent method (to the general public) consists in checking the salaries of the students who have graduated from that philosophy department. If the second justification is the one chosen to prove the success of the department, such department could start to substitute metaphysics courses by entrepreneurship ones. Would this measure translate into students who will graduate with a sounder knowledge and skills in philosophy? Not really, but a non-expert party could justify that this it is the case.

In conclusion, the effectiveness of an explanation is dependent on the background knowledge of its target, and so is the potential trust such explanation can inspire. It is not possible to draw a single path that leads to trust in every truster. Explainability contributes to trust *only* to the trusters to whom explanations are effective. Therefore, explanations are not a universal solution for the trust problem: they are a context dependent answer for a universal question.

3.2 Trust and bad explanations

The second ET assumption refers to the idea that every explanation changes the beliefs of the truster. In other words, that making AI explainable is enough to increase the trust of the trusters. However, not every explanation suffices. It can be the case that an explanation fails to achieve trust because it is a bad explanation. An explanation needs to meet certain criteria in order to qualify as a good explanation. In this paper, I call *good explanation*, broadly speaking, to explanation that brings understanding. However, what makes a good explanation (especially a scientific one) is a complex matter. Different criteria have been proposed, such as Ruben (1990)'s *particular causation*, Deutsch (2009)'s *hard-to-varyness*, or Woodward (2004)'s *manipulationism*. This discussion goes beyond the scope of this paper, even though it is an interesting topic for further research linked to XAI and ET.

Hoffman, Mueller, et al. (2018) make a keen distinction between the goodness and the satisfaction of an explanation (pp. 4-5). In their words,

“While an explanation might be deemed good [. . .], it may at the same time not be adequate or satisfying to users-in-context. Explanation Satisfaction is defined as the degree to which users feel that they understand the AI system or process being explained to them. Compared to Goodness [defined earlier in the paper as a property that could be evaluated a priori just by checking features such as clarity and precision], satisfaction is a contextualized, a posteriori judgment of explanations” (p. 5, text inside brackets mine).

While I find the use of the term “goodness” slightly misleading in this context,⁶ the distinction is spot on. In the quote above is clear the importance of the reaction of the truster to an explanation. An explanation can be good (if it meets certain criteria)⁷ but ineffective (or in Hoffman, Mueller, et al. (2018)'s terminology, unsatisfactory). It is in these cases that context becomes crucial. This brings us back to the previous subsection, where I focused on the importance of background knowledge. Other qualities, such as

⁶ I would rather use “valid”, since what is commonly understood for “good” matches the definition that Hoffman, Mueller, et al. (2018) give of “satisfaction”.

⁷ See Appendix A in Hoffman, Mueller, et al. (2018)'s paper.

the clarity of the explanation, its scope⁸ or its connection to the evidence, also play a role in the effect of the explanation on the truster.

Let me illustrate this point with an example. Imagine we have a system that is able to calculate the statistic interactions between features in a random forest trained to predict cervical cancer, given some risk factors (Molnar, 2019, pp. 149–150). A student asks about which are the most correlated risk factors. One could reply as follows:

Figure 1 shows the interaction strength (H-statistic) for each feature with all other features for a random forest predicting the probability of cervical cancer. (i)

That explanation states true facts and it is connected to evidence. However, it would not be rare that the student remained confused. At first sight, it is not clear which is the factor with the highest relative interaction effect with all other features: the years that the patient used hormonal contraceptives, the number of pregnancies she has gone through, her age or when her first sexual intercourse was. The explanation is not clear and the student is not satisfied with it. Alternatively, one could offer the following explanation:

Figure 2 shows the interaction strength (H-statistic) for each feature with all other features for a random forest predicting the probability of cervical cancer. The years on hormonal contraceptives has the highest relative interaction effect with all other features, followed by the number of pregnancies. (ii)

Both explanations refer to the same data. However, explanation (ii) does it in a clearer way, and underlines the answers that the student is more interested in (it tells them which are the two features with the highest relative interaction with the rest). Explanation (i) is likely to be ineffective, even if it tells no lies. Its lack of clarity makes it so, at least for explanatory purposes. We can conclude that understanding is more relevant than validity when labelling explanations. An explanation that the learner does not understand won't provoke any effect on them. Because of this, when defining the goals of XAI, authors as Pérez (2019) propose to switch from explanations to understanding. According to him, “it is necessary to explore a naturalistic approach to the way in which context and background knowledge mold explanations before moving to a discussion of how similar devices can be used, and have been used, in understanding AI models” (pp. 447-448). Other authors such as Floridi et al. (2018) distinguish between *explainability* and *intelligibility*. While explainability refers to the availability of explanations, intelligibility points to the understanding by the recipient of the explanation. Under this terminology, an explanation like (i) in the example above would bring explainability to the study, but not intelligibility. Following this line of

⁸ If the scope of an explanation is too wide, it is a bad explanation due to its imprecision. An explanation that could explain everything, does not really explain anything.

thought, what Floridi demands for AI is *explicability* rather than explainability. For him, explicability comprehends both intelligibility and accountability, and not necessarily an explanation (if a system is intelligible and accountable even not being explainable, explanations would not be necessary). I won't go further in this terminological debate, but it is an idea to keep in mind when discussing XAI and what is exactly the goal of providing explanations.

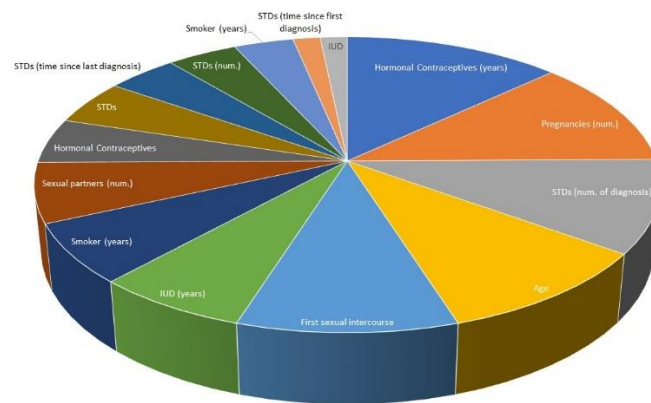


Figure 1. Graphic made with data from the UCI Machine Learning repository (see Fernandes et al. (2017))

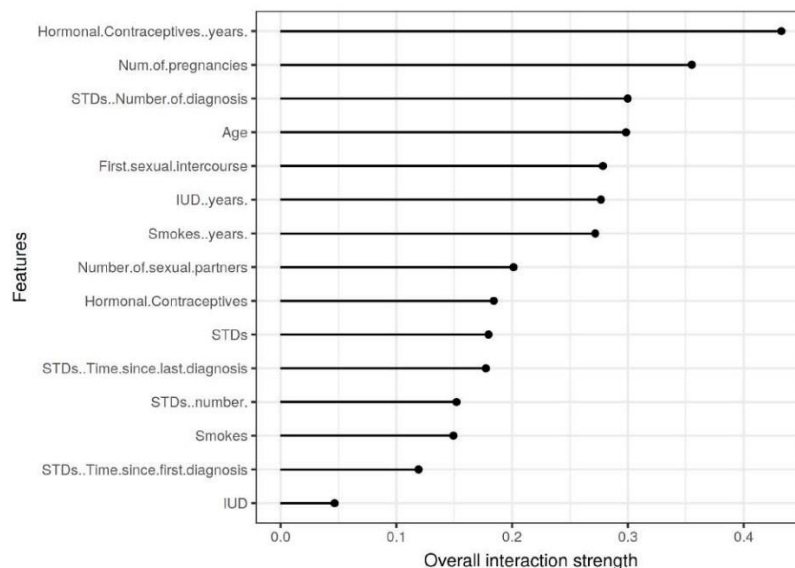


Figure 2. Graphic taken from Molnar (2019).

In sum, the quality of the explanation needs to be taken into account when pondering its effect on the trust of the truster. Not whatever kind of explanation will increase trust in every truster. In line with the previous subsection, I have shown that context plays a crucial role in trust. Therefore, the context that surrounds an explanation needs to be taken into account when considering its potential to increase trust.

4 Rejection of ET

So far, I have shown that ET exhibits epistemic limitations. Partly, these limitations come from the vagueness with which trust is usually referred to in the AI field. Having a look back at the quotations supporting ET in section 1.1, we can see that “trust” is used rather as an umbrella term.

When presenting ET in section 1, I already pointed out how the tension between objectivity and subjectivity in trust is precisely why ET fails. In the same vein as most ethics guidelines, ET proposes a universal way to achieve something subjective as trust partly is. The second part of the problem with ET is its vagueness regarding trust. ET proposes explanations as a possible way to achieve trust, but which kind of trust is aimed to be reached through explanations is unclear. ET fails due to its vagueness: about to which kind of explanations it refers to, to whom these explanations are directed at and about which kind of trust it targets. Still, ET remains quite an intuitive claim and it would not be fair to label it as completely false. However, it is significantly incomplete, which leads me to reject it. If we wanted to save something from ET for future attempts of approaching the connection between trust and explainability, there are three main tasks that need to be resolved:

- (1) Pay attention to the context of the truster and to their background knowledge when providing explanations.
- (2) Be aware that only effective explanations will lead to trust, and that not any explanation qualifies for this means.
- (3) Provide a characterization of trust which makes clear what trust means and which kind of trust is the target of explanations in AI.

This is an interesting line of research that I would like to pursue in future research. By tackling the tasks above, the aim is to restructure ET and to provide an alternative claim able to overcome the epistemic limitations of the former. ET is often used as an argument in favor of XAI. Providing a flawed argument like ET opens the door to unfair critiques to XAI. Or even worse, it could lead to the development of ineffective XAI (which fails to increment trust). A fix for this potential ineffectiveness could be working on the kind of explanations that XAI offers and its suitability to the potential trusters they are aimed for. That is a fix that is worth working on.

5 Conclusion

In this paper, I have pointed at the epistemic limitations of ET, which arise from its intended universality. Just saying that explanations lead to trust takes for granted the effectiveness of the explanations. Such effectiveness is contingent and dependent on the background knowledge of the potential trusters and the kind of explanation that is being offered. In other words, ET is vague regarding which explanations it talks about, and to whom they should be directed. Furthermore, ET is a claim that points towards a possible way to achieve trust, which is in itself quite a debated concept. Without a proper account of trust and a specification of which kind of trust can be achieved through explanations, ET is a limited hypothesis. These limitations are too much to take for granted for ET to hold. Because of this, ET needs to be rejected. There is a need for an alternative claim that connects trust and explainability more accurately, in order to properly argue for XAI. Working on a better hypothesis and characterizations of these concepts would help to avoid the development of explainable systems that fail to increment trust. This opens a way to further research in the topic.

Acknowledgments

Material from this paper was presented at the Issues in XAI conference held at the Faculty of Technology, Policy and Management, in Delft; at the Ethics of Trust and Expertise conference held at the American University of Armenia, in Yerevan; at the V Congreso de Postgrado de la SLMFCE held at the Facultad de Filosofía y Letras, in Valladolid; and at the Neurotechnology Meets AI conference held at the Institute of Ethics, History and Medicine at Ludwig-Maximilians-Universität, in Munich.

The paper has benefited significantly from the feedback from the members of the AITE project: Karoline Reindhardt, Eric Raidl, Oliver Buchholz and Thomas Grote. I would like to thank as well to the members of the IZEW-Forschungskolloquium for their constructive comments.

SB is funded by the Baden-Württemberg Foundation (program “Verantwortliche Künstliche Intelligenz”) as part of the project AITE (Artificial Intelligence, Trustworthiness and Explainability). SB is also supported by the Deutsche Forschungsgemeinschaft (BE5601/4-1; Cluster of Excellence ‘Machine Learning—New Perspectives for Science’, EXC 2064, project number 390727645).

References

- Baier, A. (1986). Trust and Antitrust. *Ethics*, 96(2), 231–260. <https://doi.org/10.1086/292745>
- Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science*, 87(4), 568–589. <https://doi.org/10.1086/709729>
- Deutsch, D. (2009). *A new way to explain explanation*. TED Talk. https://www.ted.com/talks/david_deutsch_a_new_way_to_explain_explanation
- Erasmus, A., Brunet, T. D. P., & Fisher, E. (2020). What is Interpretability? *Philosophy and Technology*, 34, 833–862. <https://doi.org/10.1007/s13347-020-00435-2>
- Fernandes, K., Cardoso, J. S., & Fernandes, J. (2017). *Transfer Learning with Partial Observability Applied to Cervical Cancer Screening BT - Pattern Recognition and Image Analysis* (L. A. Alexandre, J. Salvador Sánchez, & J. M. F. Rodrigues (eds.)); pp. 243–250). Springer International Publishing.
- Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- Gambetta, D. (1988). Can we trust trust. In D. Gambetta (Ed.), *Trust. Making and breaking cooperative relations* (pp. 213–237). Basil Blackwell.
- Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P., & Holzinger, A. (2018). Explainable AI: The new 42? *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11015 LNCS, 295–303. https://doi.org/10.1007/978-3-319-99740-7_21
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, 51(93). <https://doi.org/10.1145/3236009>
- Hardin, R. (2002). *Trust and Trustworthiness*. Russell Sage Foundation.
- Hawley, K. (2014). Trust, distrust and commitment. *Nous*, 48(1), 1–20. <https://doi.org/10.1111/nous.12000>
- HLEGAI. (2019). *Ethics guidelines for trustworthy AI*.
- Hoffman, R. R., Klein, G., & Mueller, S. T. (2018). Explaining Explanation For “Explainable Ai.” *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 62(1), 197–201. <https://doi.org/10.1177/1541931218621047>
- Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). *Metrics for Explainable AI: Challenges and Prospects*.
- Jones, K. (1996). Trust as an Affective Attitude. *Ethics*, 107(1), 4–25. <https://doi.org/10.1086/233694>

- Kästner, L., Langer, M., Lazar, V., Schomäcker, A., Speith, T., & Sterz, S. (2021). On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness. *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, 169–175. <https://doi.org/10.1109/REW53955.2021.00031>
- Krishnan, M. (2020). Against Interpretability: a Critical Examination of the Interpretability Problem in Machine Learning. *Philosophy and Technology*, 33(3), 487–502. <https://doi.org/10.1007/s13347-019-00372-9>
- Leslie, D. (2019). *Understanding artificial intelligence ethics and safety*. <https://doi.org/10.5281/zenodo.3240529>
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/J.ARTINT.2018.07.007>
- Molnar, C. (2019). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. <https://christophm.github.io/interpretable-ml-book>
- Nguyen, C. T. (2022). Transparency is Surveillance. *Philosophy and Phenomenological Research*, 1–31. <https://doi.org/10.1111/phpr.12823>
- O’Neill, O. (2002). A question of trust. In *The BBC Reith Lectures*. Cambridge University Press.
- Páez, A. (2019). The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds and Machines*, 29(3), 441–459. <https://doi.org/10.1007/S11023-019-09502-W>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13–17, 1135–1144. <https://doi.org/10.1145/2939672.2939778>
- Ruben, D.-H. (1990). *Explaining Explanation* (T. Honderic (ed.)). Routledge.
- Woodward, J. (2004). Making Things Happen: A Theory of Causal Explanation. In *Oxford Studies in the Philosophy of Science*. Oxford University Press. <https://doi.org/10.1093/0195155270.001.0001>
- Zednik, C. (2021). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy and Technology*, 34(2), 265–288. <https://doi.org/10.1007/S13347-019-00382-7>

Technical Debt is an Ethical Issue

J Paul Gibson¹[0000-0003-0474-0666], Massamaesso Narouwa¹, Damian Gordon²[0000-0002-3875-4065], Dymrna O’Sullivan², Jonathan.Turner², and Michael Collins²[0000-0002-2034-2185]

¹Telecom Sud Paris, Evry, France

²TUD, Dublin

Ireland paul.gibson@telecom-sudparis.eu

Abstract. We introduce the problem of technical debt, with particular focus on critical infrastructure, and put forward our view that this is a digital ethics issue. We propose that the software engineering process must adapt its current notion of technical debt – focusing on technical costs – to include the potential cost to society if the technical debt is not addressed, and the cost of analysing, modelling and understanding this ethical debt. Finally, we provide an overview of the development of educational material – based on a collection of technical debt case studies - in order to teach about technical debt and its ethical implications.

Keywords: Technical debt, digital ethics, critical infrastructure, teaching

1 Introduction

Ward Cunningham first used the term technical debt to describe the trade-off between quality of code and time to deliver it (Cunningham, 2019). High-quality code costs more - in the short-term - to develop and deploy than low quality code. However, lowquality code costs more in the long-term due to higher maintenance costs. Thus, the delivery of low quality software incurs a debt that must be repaid later. From a purely pragmatic view-point, it is not always good business to make such repayments (Buschmann, 2011). We propose that technical debt should not just be considered a technological or business (Conroy, 2012) issue: it is an ethical issue.

The paper’s main contributions are to raise awareness of the ethical issues that arise out of technical debt, and to report on a number of case studies that we carried out with relation to this issue. Our focus is on critical infrastructure, as the potential impact to society of not paying debt in such systems is very significant. However, we will also address technical debt whose impact is less general, but just as significant to individuals and members of specific groups. We also report on the development of an educational “brick” for teaching software engineers about ethics and how the software life-cycle can be improved through the integration of ethical requirements.

2 Technical Debt in Critical Infrastructure

Technical debt is most commonly used in reference to ICT systems (Tom et al., 2013), but the concept, and problem, is much older than that: eg, there is significant future cost associated with badly maintained infrastructure for critical systems such as transport, power, water, housing, health, education, etc. (Heimo & Holvitie, 2020). In our modern world, such infrastructure is dependent on underlying software systems (Rinaldi et al., 2001), which also incur significant debt if they are not maintained. Most of the research in this area has examined the risk associated with not keeping such infrastructure secure, and more recently there has been discussion of the ethical issues with respect to security (Grady et al., 2021)

Although these aspects are important, a secure system is not guaranteed to function correctly. Poor quality software, and associated architecture is a major source of technical debt (Ernst et al., 2015) that requires specific tools and methods to be effectively managed (Avgeriou et al., 2020). Unfortunately, metrics for technical debt focus on the cost of maintaining the system, rather than the potential cost to society if the systems are not properly maintained. Just as the technical debt builds up interest over time, the potential risk to society may also grow dramatically; we refer to this as ethical debt.

Heimo and Holvitie have proposed an alternative definition of ethical debt to be the subset of technical debt associated with the direct cost of delaying properly analysing and understanding the ethical issues associated with the system currently under development (Heimo & Holvitie, 2020). A similar definition has been proposed when developing AI systems (Petrozzino, 2021). We shall widen this definition to include the potential cost to society if the technical debt is not paid.

3 Case Studies - Work in Progress

Our research objective is to investigate different ways in which the software development process could be improved by more rigorous modelling of ethical debt. As such, we are motivated and influenced by previous work on the social and human centric aspects of software engineering (Tamburri et al., 2013, Spinola et al., 2019). We are following a case-study driven approach. There are a large number of technical debt case studies from which to choose, and our goal is not to do a literature review, such as seen in (Alfayez et al., 2020, Tom et al., 2013). Rather, we wish to re-evaluate some well-documented case studies from the point of view of ethical debt.

We have chosen not to consider the technical and ethical debt arising out of the use of AI (Bogner et al., 2021), but instead wish to focus on more traditional software systems. We wish to provide guidelines on improving the software process through better management of ethical debt, much in the way that different researchers have proposed better management of technical debt (Lenarduzzi, et al. 2021, Codabux & Williams, 2013).

3.1 Choice of Case Studies

Following feedback from colleagues, and comments of the initial reviewers of the paper, we realised the importance of choosing the case studies that would give us most insight into the problem of the ethical issues arising out of technical debt.

First, we need to be clear what we mean by technical debt: it arises from a deliberate choice to defer a technical cost in the short term, but which must be paid in the long term. Implicit in this decision is that the technical debt will be managed and repaid. There are many different examples, with a range of consequences, eg:

1. Knowing that the system will not function correctly at some moment in the future, but waiting for nearer the time to fix it.
2. Knowing that the system does not function correctly for a small number of users, but it is deemed too costly to fix it just for them. Perhaps, in the future, they will allocate time and resources to fix this.
3. Knowing that the system architecture will not scale in the long-term, but waiting to fix this when the need arises
4. Knowing the code is insecure, but it functions correctly, and there is a decision to add new functionality rather than address the security issues
5. Knowing that code depends on packages/technologies which are out of the developer's control, but which they don't fully understand.
5. Knowing that the code has not been properly tested, but deploying it anyway with the plan to fix any issues that occur as they arise.
6. Knowing that the code is not well-documented, but delivering it even when it may be difficult to maintain.

We do not claim that all legacy systems have significant technical debt, but we do think that there is technical debt (of more or less importance) in all such systems. We agree with Monaghan and Bass (2020) that "By positioning Legacy within the context of Technical Debt, practitioners have a more concrete understanding of the state of the systems they maintain". Incurring technical debt is not necessarily unethical, but it may be in some cases. Furthermore, failing to properly manage the debt may also have important ethical consequences, and so incur significant ethical debt.

3.2 Mishandling of Dates – Y2K and beyond

The issue of dates is well-documented in the domain of software. The Y2K problem was widely reported (Williams et al., 2014) with worldwide concern over the possibility of critical bugs. Y2K was not the first such reported problem – issues with leap years were known much earlier (Neumann, 1992) – and it will not be the last, as we wait for Y2K38 (Okabe et al., 2020). Recently, more than 20 bugs have been reported with respect to Y2K22 (Neumann, 2022), including significant issues for Microsoft Exchange, Honda Clocks and Google Chrome users. Also, there has been some concern over the mishandling of leap seconds (Burnicki, 2015) in GPS systems (Anumasula et al., 2018).

We consider these to be examples of technical debt, as the issues were known at development time, and decisions were taken to wait until later to address them. The potential impact of not adequately addressing this debt before the strict deadlines is very significant, and could be considered critical. For example, the cost of fixing Y2K bugs is estimated to have been about 100 billion dollars worldwide; however, the cost to society of not fixing the bugs would have been much more (Best 2003). There are many reported examples of the consequences of failure to fix Y2K bugs on time, for example in critical health service code (Thimbleby, 2021). Thus, we consider this type of time-based technical debt to also be a serious ethical issue.

3.3 Open Source – Log4J

There is significant reliance on open source software, and difficulty in tracking the complex (inter-)dependencies (Bauer et al., 2020). The Log4J bug (Olbrich et al., 2020) is a good example of how a bug in a small, but significant, open source framework (used, in this case, for logging) can have far-reaching consequences (Srinivasa et al., 2022).

The technical debt, in this case is not specifically in the Log4J code, which was mostly written and maintained by a single person. The technical debt is in the other systems that re-used the Log4J framework without fully analysing the dependency on the code and the risks of the code being insecure.

3.4 GDPR Compliance – Legacy Systems in The Health Domain

There are significant technical costs in “Retrofitting GDPR Compliance onto Legacy Databases” (Agarwal et al., 2021). Furthermore, verifying compliance with the different GDPR requirements is a significant challenge (Li et al., 2019). This may, in part, explain why GDPR non-compliance is a significant problem in healthcare (Agyei et al., 2020).

Would it be fair to blame this on technical debt ? This depends on whether GDPR was already ‘on the horizon’ when the legacy systems were being developed. If so, it could be said that there was a deliberate decision to incur technical debt with respect to future GDPR requirements. If not, then the question is not so clear cut. Perhaps the developers were aware of the security issues but, as there was no legal requirement to resolve them, they may have chosen to ignore them. In this case the future technical debt is an ethical issue. However, if the developers were unaware of such issues then this should not be considered to be technical debt.

There is evidence that companies prefer to pay fines for GDPR non compliance than to pay off the technical cost of fixing noncompliance (Grant & Crowther, 2016). This is a clear ethical issue, particularly in critical domains such as health.

3.5 Accessibility

As with GDPR compliance, in many countries there are legal requirements for software with respect to accessibility. Unfortunately, these requirements are often not

even met for public (web) services such as health (Alajarmeh, 2021), voting (Takahiro, et al., 2016) and education (Oswal & Hewett, 2013). Could this be another case of it being less costly to pay the fines than to pay the technical debt ? This merits further investigation, and is clearly an ethical (as well as legal) issue.

3.6 Agile Methods

There is some evidence that more agile development processes give rise to more technical debt (Behutiv et al., 2017), and so this is an hypothesis that merits further, more rigorous, investigation. Agile development may also give rise to more ethical debt, with respect to the cost of performing the ethical analysis (Judy, 2009). Agile often leads to a 'Move Fast and Break Things' mentality, whose legal implications have been reported (Simon Chesterman, 2021). As software development is becoming even more agile - moving towards continuous integration and continuous deployment - it is clear that updating the software development process and life-cycle to include ethical aspects is critical.

3.7 Security

Significant security vulnerabilities can arise when security-critical code is deployed without having been properly tested. Three well-known examples of this are: OpenSSL's Heartbleed (Durumeric et al., 2014), Apple's "goto fail" (Bland, 2014), and the INTEL AMT Vulnerability (Ermolov, M., & Goryachy, M., 2017).

However, after brief analysis it was clear that this is not an example of technical debt as we define it: in none of these cases was there evidence of a deliberate decision taken to deploy code that was not properly tested. The problem was that the companies misbelieved that the deployed code had been rigorously tested. Thus, there was an issue with their software development process and incompetence, but this is not a good example of technical debt.

4 Future Work

4.1 Educational Brick(s) for Teaching about Technical and Ethical Debt

The development of the educational material will follow the approach taken in the EU Ethics4EU project (Gibson et al., 2021, Curley et al., 2021) which has focused on the pedagogic aspects of digital ethics, and the production of autonomous educational bricks. We believe that the chosen case studies will be highly motivational for the students, and provide open-ended problems for educators to leverage in teaching to a wide range of learners (Stavarakakis et al., 2021). We are currently developing project work based around the Log4J case study, and university web site accessibility issues.

4.2 Integrating Ethical Debt Analysis and Management in the Software Process

A long-term goal of our research is to propose methods for incorporating ethical analysis of technical debt into the software development process (at all stages in the lifecycle). As stated by Gotterbarn (1991), “The ethical problems faced by the software engineer involve: the end product, the process of developing that product, and the human interactions in the development of the product.”

We are also motivated by Biabie et al. (2022), who propose that ethical issues should be included in the requirements phase of the software lifecycle. Previous research on management of technical debt during software development also suggests that the issue needs to be addressed throughout the life-cycle (Zengyang et al., 2015 and Rios et al., 2018).

5 Conclusions

We have shown, through a number of examples, that technical debt is often an ethical issue. There is an urgent need to educate engineers about technical and ethical debt, particularly with respect to critical system infrastructure.

Acknowledgements

The work reported in this paper was partially funded by the EU Erasmus+ transnational project Ethics4EU (<http://ethics4eu.eu>).

References

- Agarwal, A., George, M., Jeyaraj, A., & Schwarzkopf, M. (2021). Retrofitting GDPR compliance onto legacy databases. *Proceedings of the VLDB Endowment*, 15(4), 958-970.
- Agyei, E. E. Y. F., & Oinas-Kukkonen, H. (2020, April). GDPR and Systems for Health Behavior Change: A Systematic Review. In *International Conference on Persuasive Technology* (pp. 234-246). Springer, Cham.
- Alajarmeh, N. (2021). Evaluating the accessibility of public health websites: an exploratory cross-country study. *Universal access in the information society*, 1-19. Reem Alfayez, Wesam Alwehaibi, Robert Winn, Elaine Venson, and Barry Boehm. A systematic literature review of technical debt prioritization. In *Proceedings of the 3rd International Conference on Technical Debt*, pages 1–10, 2020. DOI = 10.1145/3387906.3388630
- Anumasula, R., Mohanan, S., Rathour, H. K., Agarwal, P. K., & Baba, K. V. S. (2018, December). Indian Experience on Impact of Leap Second in Synchrophasor Systems. In *2018 20th National Power Systems Conference (NPSC)* (pp. 1-6). IEEE.

- Paris C Avgeriou, Davide Taibi, Apostolos Ampatzoglou, Francesca Arcelli Fontana, Terese Besker, Alexandros Chatzigeorgiou, Valentina Lenarduzzi, Antonio Martini, Nasia Moschou, Ilaria Pigazzini, et al. An overview and comparison of technical debt measurement tools. *IEEE Software*, 2020. DOI = 10.1109/MS.2020.3024958
- Andreas Bauer, Nikolay Harutyunyan, Dirk Riehle, and Georg-Daniel Schwarz. Challenges of tracking and documenting open source dependencies in products: A case study. *Open Source Systems*, 582:25, 2020. DOI = 10.1007/978-3-030-47240-5_3
- Behutiye, W. N., Rodríguez, P., Oivo, M., & Tosun, A. (2017). Analyzing the concept of technical debt in the context of agile software development: A systematic literature review. *Information and Software Technology*, 82, 139-158.
- Best, K. (2003). Revisiting the Y2K bug: language wars over networking the global order. *Television & New Media*, 4(3), 297-319.
- Biable, S. E., Garcia, N. M., Midekso, D., & Pombo, N. (2022). Ethical Issues in Software Requirements Engineering. *Software*, 1(1), 31-52.
- Mike Bland. Finding more than one worm in the apple. *Communications of the ACM*, 57(7):58– 64, 2014. DOI = 10.1145/2622628.2622630
- Justus Bogner, Roberto Verdecchia, and Ilias Gerostathopoulos. Characterizing technical debt and antipatterns in AI-based systems: A systematic mapping study. *arXiv preprint arXiv:2103.09783*, 2021. DOI = 10.1109/TechDebt52882.2021.00016
- Burnicki, M. (2015, February). Technical aspects of leap second propagation and evaluation. In *Requirements for UTC and Civil Timekeeping on Earth Colloquium. Science and Technology Series (Vol. 115)*.
- Frank Buschmann. To pay or not to pay technical debt. *IEEE software*, 28(6):29–31, 2011. DOI = 10.1109/MS.2011.150
- Simon Chesterman. 'Move Fast and Break Things': Law, Technology, and the Problem of Speed. *Singapore Academy of Law Journal*, 33:5–23, 2021. DOI = 10.1145/3244026
- Zadia
- Codabux and Byron Williams. Managing technical debt: An industrial case study. In *2013 4th International Workshop on Managing Technical Debt (MTD)*, pages 8–15. IEEE, 2013. DOI = 10.1109/MTD.2013.6608672
- Patrick Conroy. Technical debt: Where are the shareholders' interests? *IEEE Software*, 29(6):88– 88, 2012. DOI = 10.1109/MS.2012.166
- Curley, D. Gordon, I. Stavrakakis, A. Becevel, J.P. Gibson, and D. O'Sullivan. Adaptable and reusable educational bricks for teaching computer science ethics. In *EDULEARN21 Proceedings, 13th International Conference on Education and New Learning Technologies*, page 1991. IATED, 5-6 July, 2021. DOI = 10.21125/edulearn.2021.0456
- Ward Cunningham. The WyCash portfolio management system. *ACM SIGPLAN OOPS Messenger*, 4(2):29–30. DOI = 10.1145/157710.157715
- Zakir Durumeric, Frank Li, James Kasten, Johanna Amann, Jethro Beekman, Mathias Payer, Nicolas Weaver, David Adrian, Vern Paxson, Michael Bailey, et al. The matter of

- Heartbleed. In Proceedings of the 2014 conference on internet measurement conference, pages 475–488, 2014. DOI = 10.1145/2663716.2663755
- Ermolov, M., & Goryachy, M. (2017). How to hack a turned-off computer, or running unsigned code in intel management engine. Black Hat Europe.
- Neil A Ernst, Stephany Bellomo, Ipek Ozkaya, Robert L Nord, and Ian Gorton. Measure it? manage it? ignore it? software practitioners and technical debt. In Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering, pages 50–60, 2015. DOI = 10.1145/2786805.2786848
- J Paul Gibson, Yael Jacob, Damian Gordon, and Dymrna OSullivan. Developing an educational brick for digital ethics. In Moving technology ethics at the forefront of society, organisations and governments, ETHICOMP, pages 29– 37. Universidad de La Rioja, 2021.
- Gotterbarn, D. (1991, May). Ethical considerations in software engineering. In Proceedings of the 13th international conference on Software engineering (pp. 266-274).
- Caitlin Grady, Sarah Rajtmajer, and Lauren Dennis. When smart systems fail: the ethics of cyberphysical critical infrastructure risk. IEEE Transactions on Technology and Society, 2021. DOI = 10.1109/TTS.2021.3058605
- Olli I Heimo and Johannes Holvitie. Ethical debt in is development. comparing technical and ethical debt. ETHICOMP 2020, pages 29–30, 2020.
- Jim Horning and Peter G Neumann. Risks of neglecting infrastructure. Communications of the ACM, 51(6):112– 112, 2008.
- Ken H Judy. Agile principles and ethical conduct. In 2009 42nd Hawaii International Conference on System Sciences, pages 1–8. IEEE, 2009.
- Valentina Lenarduzzi, Terese Besker, Davide Taibi, Antonio Martini, and Francesca Arcelli Fontana. A systematic literature review on technical debt prioritization: Strategies, processes, factors, and tools. Journal of Systems and Software, 171:110827, 2021.
- Li, Zengyang, Paris Avgeriou, and Peng Liang. "A systematic mapping study on technical debt and its management." Journal of Systems and Software 101 (2015): 193-220. Li, Ze Shi, Colin Werner, and Neil Ernst. "Continuous requirements: An example using GDPR." 2019 IEEE 27th International Requirements Engineering Conference Workshops (REW). IEEE, 2019.
- Miura, Takahiro, et al. "Accessibility, efficacy, and improvements in voting methodology for visually impaired persons using a web-based electronic ballot system." Proceedings of the 8th Indian Conference on Human Computer Interaction. 2016.
- Monaghan, B. D., & Bass, J. M. (2020, November). Redefining legacy: a technical debt perspective. In International Conference on Product-Focused Software Process Improvement (pp. 254-269). Springer, Cham.
- Peter G Neumann. Leap-year problems. Communications of the ACM, 35(6):162– 163, 1992. DOI = 10.1145/1349026.1349047
- Neumann, P. G. (2022). Risks to the Public. ACM SIGSOFT Software Engineering Notes, 47(2), 4-7

- Ryo Okabe, Jun Yabuki, and Masakatsu Toyama. Avoiding year 2038 problem on 32-bit linux by rewinding time on clock synchronization. In 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), volume 1, pages 1019–1022. IEEE, 2020. DOI = 10.1109/ETFA46521.2020.9212079
- Steffen M Olbrich, Daniela S Cruzes, and Dag IK Sjøberg. Are all code smells harmful? a study of god classes and brain classes in the evolution of three open source systems. In 2010 IEEE International Conference on Software Maintenance, pages 1–10. IEEE, 2010. DOI = 10.1109/ICSM.2010.5609564
- Catherine Petrozzino. Who pays for ethical debt in AI? AI and Ethics, pages 1–4, 2021. DOI = 10.1007/s43681-020-00030-3
- Srinivasa, S., Pedersen, J. M., & Vasilomanolakis, E. (2022). Deceptive directories and “vulnerable” logs: a honeypot study of the LDAP and log4j attack landscape. In IEEE EuroS&P Work shop on Active Defense and Deception (AD&D). IEEE.
- Steven M Rinaldi, James P Peerenboom, and Terrence K Kelly. Identifying, understanding, and analyzing critical infrastructure interdependencies. IEEE control systems magazine, 21(6):11–25, 2001. DOI = 10.1109/37.969131
- Rios, N., de Mendonça Neto, M. G., & Spínola, R. O. (2018). A tertiary study on technical debt: Types, management strategies, research trends, and base information for practitioners. Information and Software Technology, 102, 117-145.
- Rodrigo O Spinola, Nico Zazworka, Antonio Vetro, Forrest Shull, and Carolyn Seaman. Understanding automated and human-based technical debt identification approaches-a two phase study. Journal of the Brazilian Computer Society, 25(1):1–21, 2019. DOI = 10.1186/s13173-019-0087-5
- Stavarakakis, I., Gordon, D., Tierney, B., Becevel, A., Murphy, E., Dodig-Crnkovic, G., ... & O’Sullivan, D. (2021). The teaching of computer ethics on computer science and related degree programmes. a European survey. International Journal of Ethics Education, 1-29.
- Damian A Tamburri, Philippe Kruchten, Patricia Lago, and Hans van Vliet. What is social debt in software engineering? In 2013 6th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE), pages 93–96. IEEE, 2013. DOI = 10.1109/CHASE.2013.6614739
- Thimbleby, Harold. Fix IT: See and solve the problems of digital healthcare. Oxford University Press, 2021.
- Edith Tom, Aybuke Aurum, and Richard Vidgen. An exploration of technical debt. Journal of Systems and Software, 86(6):1498–1516, 2013. DOI = 10.1016/j.jss.2012.12.052
- S Mitchell Williams. An international investigation of associations between societal variables and the amount of disclosure on information technology and communication problems: The case of y2k. The International Journal of Accounting, 39(1):71-92, 2004. DOI = 10.1016/j.intacc.2003.12.003

EKIP: designing a practical framework for embedding ethics in AI software development

Rebecca Raper¹ and Mark Coeckelbergh¹

¹The University of Vienna, Vienna, Austria

rebecca.raper@univie.ac.at

Abstract. As Artificial Intelligence (AI) systems are used more within society, there is the pressing need to ensure that they are developed in an ethical way. Several notable attempts to ensure that AI is developed ethically have been advanced. Though each approach has its merit in terms of offering guidance for ethical AI development, none, yet offer a rigorous, practical approach to ensure that AI systems are developed in the right way. Furthermore, these approaches seem to be out of sync with current actual software development methodologies. EKIP is a three-year-long FFG-sponsored project which aims to address this issue by offering a *practical* framework for software developers. Inspired by *Ethical by Design* (Mulvenna et al., 2017) – that ethics should be placed into the design phase of software development – EKIP’s purpose is to explore more practical approaches for AI software developers to embed ethics into their systems. The aim of this paper is to outline the motivation behind EKIP, by giving an overview of the attempts that have been made to embed ethics in AI so far, and by introducing a ‘methodological gap’. It is argued that if we are to effectively embed ethics in AI systems, we need to consider new frameworks that integrate with the traditional software development techniques, such as ‘waterfall’ and ‘Agile’. An *ethical requirements*-based approach is described which presents a twist on traditional software development methodologies. To conclude, suggestions are made for future work, and there is a plea for further research to advance this topic.

Keywords: AI Ethics, AI Frameworks, Ethical AI, EKIP, Software Development Lifecycle, Software Development, Waterfall, ML-Ops, Kanban, Requirements Engineering

1 Background and Introduction: Ethical AI

Defining *Ethical Artificial Intelligence* (AI) as the pursuit to design, create and implement AI systems that adhere to ethical standards (Coeckelbergh, 2020), the discipline has gained an increasing amount of traction in recent years. This is partially owing to various high-profile incidents concerning the use (or misuse) of AI systems. For instance, in March 2020, it was reported that Cambridge Analytica was collecting and processing vast quantities of data to form decisions about the types of

advertisements an individual might be targeted with, or what articles they see online (BBC News, 2020). The controversy and objections concerned the fact that AI profiling was being used to target and manipulate individuals with material relating to a political referendum campaign, highlighting the vast amount of influence AI decision-making can have. Also in 2020, A-Level result allocation in the UK had a similar story, with algorithms used to assign grades to students showing to be prejudiced against those from certain demographics or poorer backgrounds (The Guardian, 2020) – the implication being that if you were from a poorer background, you were less likely to get into the university of your choice – ultimately reducing social mobility and career fulfilment. There have been similar cases in the United States, where AI profiling used by the police to forecast criminal behaviour, has been shown to be prejudiced against those from minority groups (Wired, 2020). Similarly, facial recognition technology used in procedures such as interviews, or for identity verification, has been shown to be less effective for users with darker skin (The Independent, 2021). As AI is being used more and more in society, to deal with economic, social, and technological problems, there is the concern that these identified types of incidents are widespread and that they are having a significant effect on individual lives and communities.

Other implications of AI systems concern issues such as *privacy* – if an AI system is making inferences about our personality characteristics, to what extent do we still have privacy? Another issue is *trust* – with algorithms having such bad press (whether this be qualified or not), how can (and should) public perception of AI be developed so that they can realise the true benefits that AI can bring? *Autonomy* and *control* are other issues – who gets to decide how AI algorithms are applied, and then decide the eventual fate of individuals based upon the algorithmic decisions? This then leads to questions of delegated *power*, and *political influence* (Coeckelbergh, 2022). There is also a piece for *education*. With AI systems affecting everybody's lives (in some way), and with Artificial Intelligence itself due to become increasingly powerful (Bostrom, 1998) everybody should be equipped to understand how it works and the impact it has on their day-to-day lives, so they can appropriately interrogate, ask questions and engage in social and political debates.

It is these such issues that have led to calls for more Ethical AI and have led to increased academic attention to the area. Institutes such as 'The Ethical AI Institute' in the UK, and 'The Institute for Ethical AI and Machine Learning' in Germany, have been established in recognition of this problem - the idea being that interdisciplinary collaboration (i.e. between science and the humanities (Tasioulas, 2021)) will help to resolve some of these issues. There are calls for immediate action to be taken.

2 Attempts to deal with the issues so far

Various attempts have been made to address the ethical issues pertaining to AI, without the need to call a moratorium on AI use/development in general. One of the first attempts made was to create a set of guidelines or principles for AI developers to follow. Different bodies have different principles (see Floridi and Cowls, 2021) but

broadly speaking, the principles can be reduced to themes such as *Transparency*, *Trust*, *Autonomy* and *Explainability*, namely, an AI is only deemed ethical if it matches these principles - so is Trustworthy, Grants Human Autonomy etc. In line with the principles, Ethical Standards have been developed to highlight what needs to be done to ensure that the principles are adhered to. For instance, a system might only be regarded as 'explainable' if it can be explained to stakeholders in a certain way. Under this approach, only systems that meet these standards can be said to be ethical, and therefore worthy of approval. The problem is: with such standards in place, it's difficult for developers to understand (without extensive training) (a) what the specific standards mean or (b) what *actually* needs to be done to meet them. Requests such as 'explain to stakeholders' can be ambiguous when notions such as 'explainability' are still very much open to philosophical debate (Arrieta et al, 2020). With the very pressing need to ensure that AI systems are developed in the right way now (rather than in 10 years' time after the terms have gone through rigorous philosophical discussion) another methodology is required that prescribes to developers how they can create their AI systems to meet such standards.

'Ethical by Design' (Mulvenna et. al, 2017) is one such methodology that has been proposed to ensure this, with the suggestion being that ethics be integrated into the AI software development process from the very start. This gained a significant amount of traction, with the IEEE adopting an 'Ethically Aligned Design' (How, 2018) manifesto to aim to achieve this. However, how to create AI systems that meet the various list of standards is still left very much open. Still – developers know they must do something to make their systems more ethical, but it is not obvious what.

One proposal - 'Value-sensitive design' (see Umbrello, 2018, Friedman, 2004 and Van den Hoven, 2007) - aims at determining a set of values at the beginning of the development process from which to work upon. Stakeholders (and end users) undergo a series of interviews to assess *what is important to them*, and then this information is used to inform the design of the AI system.

However, there are problems with such an approach. Namely, no end of user interaction will be able to fully capture every individual value, and even if they were, how do we then deal with the fact that individual values may be conflicting? An additional mechanism is required on top of this approach to determine which values take priority, and without strict oversight, this then faces the possibility of becoming nothing more than a series of prejudices. More fundamentally, however, the Value by Design approach neglects the foresight and academic rigor that has gone into the development of principles and standards such as those developed by the IEEE in the first place. There needs to be some way to bridge the gap between AI system design and the need for adherence to principles such as transparency, fairness etc.

More recently, AI Auditing (Mokander and Floridi, 2021) has been offered as a verification method to ensure that AI systems meet ethical standards previously set out. So, AI systems can be assessed against a checklist of criteria for satisfaction of the standards. However, though a good mechanism to assess that AI systems are doing what they should, there is still a *methodological gap* insofar as something is required to outline to developers how to make their AI systems ethical.

3 A ‘methodological gap’ in Ethical AI

In practice, we can see that we have *top-down* principles or standards that prescribe how AI systems ought to be developed (i.e. with *transparency* etc). We also have *horizontal* methodologies that advise that such principles should be incorporated into the AI system at the start of AI system creation (i.e. Ethical by design). We also have a technique to assess that AI systems have been created at the appropriate ethical level (AI auditing), and some techniques to show us that stakeholder values are important in the design of AI systems. Schematically, we can represent the interplay of all the different approaches by a diagram as per Figure 1:

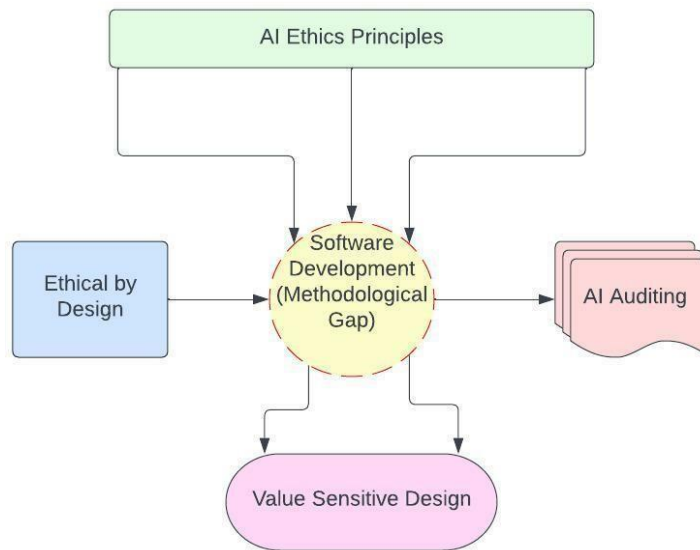


Fig. 1. The current state of AI Ethics

We can see that although we have mechanisms to provide oversight of AI systems (namely, AI Ethical Principles), and a technique to prescribe when we ought to be incorporating ethics into our systems (Ethical by Design), and how to assess them (AI Auditing), there is still space to understand how we ought to be developing the AI systems to ensure that both the principles are adhered to, and that the relevant appropriate standards are met. **This** is the methodological gap.

Describing this problem in terms of a ‘methodological gap’ allows us to begin to envisage what needs to be done to rectify the problems associated in the first chapter of this paper: namely, AI bias, discrimination, privacy and broad impacts to humans. What is required is a methodology that takes note of the already determined AI principles (i.e.

that it's important not to infringe upon autonomy), assess these against real-life potential impacts to stakeholders (i.e. my autonomy might be infringed if AI decides what I'm watching on TV tonight), and then puts mechanisms in place to protect against these. Schematically again, this need can be represented as per Figure 2.

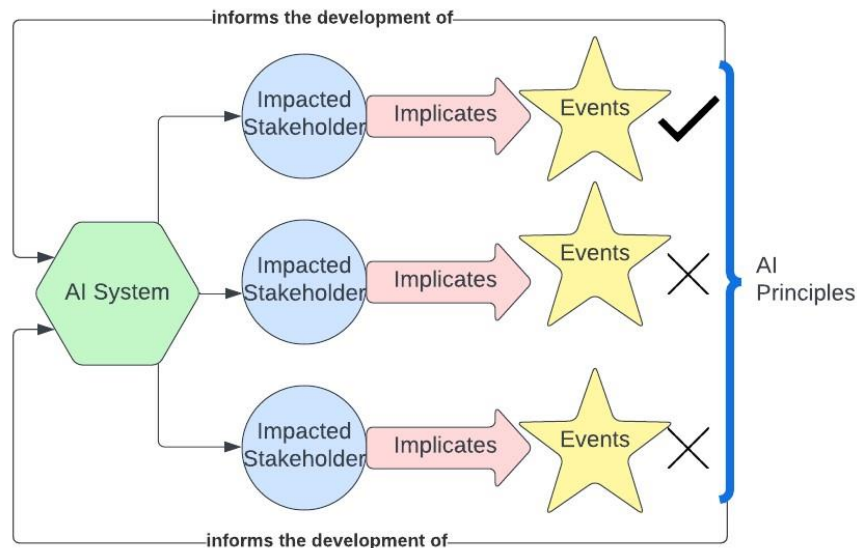


Fig. 2. A need for integration of current Ethical AI techniques.

As can be seen from the diagram, this approach still appeals to the ‘Ethical by Design’ methodology, since what ultimately informs the development of the AI system, is either an infringement or adherence to AI Ethical Principles (indicated by the cross or tick on the diagram): the system being ethical insofar as it adheres to Ethical Principles (such as those previously discussed). However, the way in which this approach differs is that the principles act as a *guide* as opposed to enforcement for how AI systems ought to be developed. In other words, the AI developer’s autonomy is preserved, since they are not forced to develop their systems in a certain way (See Raper et al, 2022, for some benefits of preserving developer autonomy). Up until recently, a top-down regulation style approach has been used (see Madiaga, 2019), but it has been criticised for potentially inhibiting innovation. The new approach allows AI systems to be developed ethically, whilst preserving AI developer autonomy.

In understanding this approach, it’s important to realise that it is not an impact-based model. There have been recent discussions about using impact assessments to determine how to develop an AI system (Kazim, 2021) (the ethical system being that which does not lead to negative consequences). Ultimately, however, it is not possible to know the full impact a system might have until it has been developed and put into place. We might reflect afterwards (based upon impact) that an AI system should not

be the way it is (i.e. the negative consequences of the Cambridge Analytica situation), but this approach would be *reactive* as opposed to *proactive*. We want to ensure that there are no negative consequences for the AI systems that are developed, and therefore, that they are ethical before they are integrated into the world.

As opposed to assessing an AI system based upon its (previously unknown) impacts, the AI system is assessed according to its (previously known) implications. The implications can be determined through case studies, interviews and workshops with impacted stakeholders. For example, I might assess how a new medical diagnostic device will implicate the end stakeholders (i.e. patients) by asking a series of questions to determine how the system (in principle) will affect their lives. A conversation such as the following might be used in a user workshop to ascertain this:

AI Analyst: We are developing an AI tool to assess how likely somebody is to develop diabetes based upon their current diet and lifestyle.

Patient: That sounds good, but some days I am healthy, other days I eat unhealthy because it depends upon how busy I am...

This is a very short interview, but it highlights the types of conversations that need to be held to truly ascertain the human implications an AI system might have. As this example demonstrates, after discussion with a patient we see that such a tool would need to take account of variable diets and lifestyles (since some patients eat mixed diets), therefore, we are able to determine that there is a requirement to ensure that any AI system used to predict diabetes risk, considers diets that change. After significant further analysis, we begin to develop a requirements-based model for Ethical AI Development of this specific system.

There are many ways to satisfy the requirements specified during the analysis phase, but what is important to note is that doing so should be an exercise in *requirementsbased design*, rather than harm/bias etc. mitigation. In this instance, the new system has the need to accommodate for changing diets. It then becomes the AI developer's creative agency to design a system which satisfies this.

4 EKIP: A practical methodology for Ethical AI Software Development

The requirements-based approach to Ethical AI software development is one that has been proposed previously (Guizzardi et al, 2020), however there has not yet been a thorough articulation for why this approach is needed, how it significantly differs to current methodologies, and – most importantly - what an implementation of the solution might look like.

EKIP (shorthand for “Ethische KI Plattform” = “Ethical AI Platform”) - an Austrian, FFG-sponsored project - was established in an attempt to provide a very practical way to integrate ethics into software development by proposing that ethics be integrated into an already-established AI development platform created by the company Gradient Zero, called ‘DQ0’.

DQ0 offers a privacy-preserving platform for data scientists to develop Machine Learning algorithms to run queries against sets of (sensible) data. The aim of EKIP then is to extend the platform so that when AI developers create their systems, it facilitates creation of them in an ethical way.

Referring to the Ethical Requirements Framework approach detailed in the previous section, therefore, DQ0 best enables this facilitation (and in turn, preserves developer creativity and agency) by providing a platform to integrate the elicited requirements. Therefore, the most effective way to practically integrate ethics that resolves the ‘methodological gap’ is to create a new software development process that allows for *Ethical Requirements Elicitation*. Figure 3 shows how this should be done in respect to the DQ0 platform.

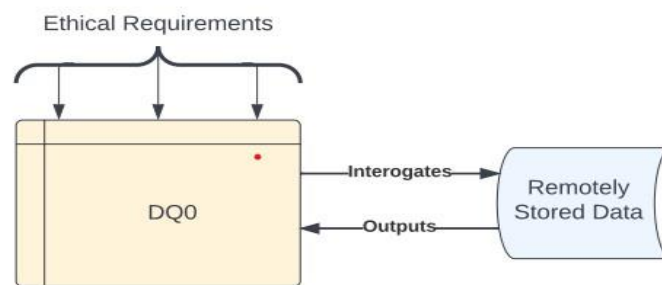


Fig. 3. EKIP - A practical approach to embedding ethics in AI software development

As mentioned previously, the ethical requirements would be ascertained through case studies, interviews, or general analysis techniques. For instance, in the (brief) conversation scenario outlined above, we determine that the ethical requirement is that: ‘The AI system must accommodate for variable diets’, with the development *problem to be solved* being left to the freedom of the AI developer. It may be that a new role is required within a project team to satisfy this new type of requirement – an individual with the capacity to facilitate such workshops and map implications for affected stakeholders. The key to this phase is in ensuring that the analysis assesses stakeholder implications rather than just looking toward impacts.

5 Integrating ethics into pre-existing software development methodologies (i.e. Waterfall and Kanban)

It is important in understanding how to execute this framework, that it integrates well into the pre-existing software development processes, and that it can also be distinguished from these. Where we are able to know the outcomes of a traditional piece of software, when it comes to AI software development, because AI acts autonomously, the outcomes are not always known.

Previously, before AI software development became more commonplace, methodologies such as ‘The Waterfall Method’ (see Figure 4) were traditionally used to ensure that software systems were designed according to the needs of the customer (or business). Traditionally, this included several phases, the first being ‘project scoping’ and ‘requirements elicitation’, where it was determined (1) how useful the new system might be and then (2) what design requirements the system should have to satisfy the overarching business objectives. Though The Waterfall Method has evolved since its initial inception (i.e. more agile software development approaches such as Kanban (Huang, 1996) are now frequently used), fundamentally, the same procedural steps remain (albeit in a more dynamic way). However, though still useful for broad software development, these are insufficient for AI software development. Because an AI system acts on its own, it is not possible to design it to meet specific conditions (i.e. have a blue screen) - since these will be variable. Instead, mechanisms are required to ensure that the AI system acts in an *ethical way* - hence it must meet ethical design requirements.

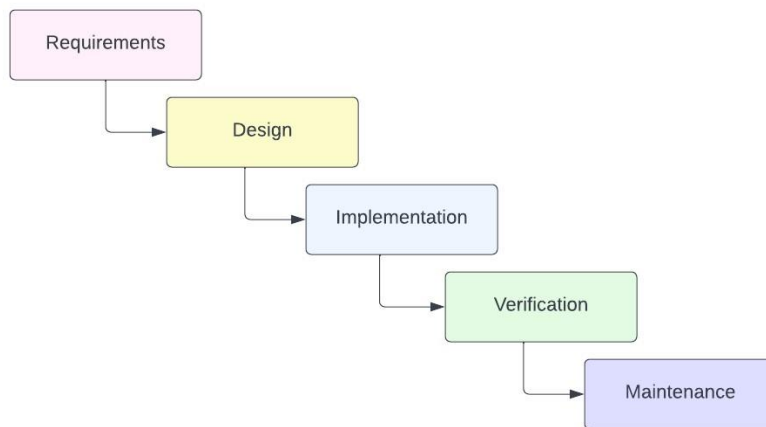


Fig.4. The ‘Waterfall’ software development methodology

Considering ethical requirements elicitation alongside traditional software development approaches, therefore the new process for Ethical AI Software Development should look as per Figure 5.

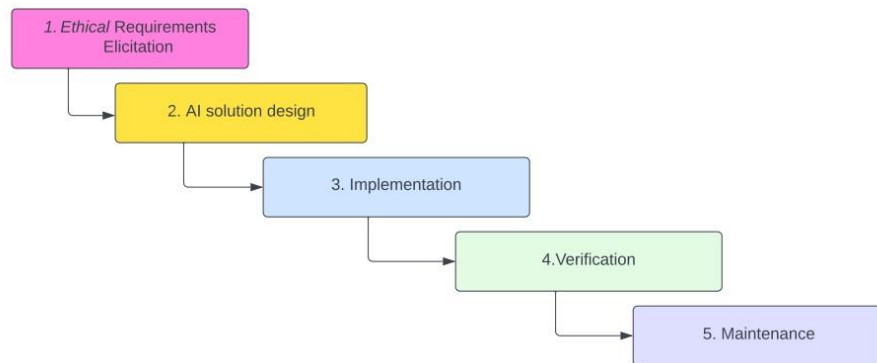


Fig.5. The Ethical AI Software development cycle.

Though the models are very similar, as already mentioned, the type (and methodology) for the requirements elicitation (and thereby solutions design) are very different. In Ethical AI case, workshops are used to find out not just what needs to be done to meet business needs, but what needs to be done to ensure that human beings have a fulfilling and flourishing life. As already mentioned, this might require a new role within the project team – a psychologist or social scientist? – but it is a twist on the original requirements elicitation framework.

As with traditional software development, it is important that there is still verification of the AI system in the end (i.e. it should match the earlier prescribed requirements) along with constant re-evaluation and maintenance of the AI system to ensure issues such as ‘AI drift’ (the susceptibility to later become biased) don’t creep in. In the case of AI systems this means periodic checking to ensure that the system still meets its initial ethical requirements (or requirements for the current stakeholders) and an iterative design process to constantly improve and evaluate the AI system.

As a final point, it is worth mentioning processes that have already been established to try to address the differences between traditional software and AI software development. For example, Microsoft’s ‘ML-Ops’ (Alla and Adari, 2021) was designed to accommodate for these very differences. However, though offering an agile environment for AI development, they neglect the need for ‘human-centricity’ in AI software development. This can only be achieved through thorough a proper analysis of humanity and human lives.

6 Conclusion

Much current academic literature and industry conversations talk about integrating ethics into software, but few practical methodologies are available to do this in specific software development and business contexts in a way that is useful to developers and respects the way they work and solve problems. In this paper we have described a motivation for, and given a broad description of the EKIP project, which

enables us to work on precisely this problem through exploring how ethics can be integrated into software development methodologies in the context of an AI platform.

We propose that the best approach to achieving the Ethical AI aims set out by bodies such as the IEEE and The Ethical AI Institute, is by integration of an *ethical requirements gathering* technique into pre-existing software development processes.

In terms of next steps, to progress this research further work is required to develop the finer details of the ethical requirements elicitation process. There need to be tools to facilitate this methodology as well as specific roles and responsibilities. There is also space for further discussion about what constitutes a fulfilling, flourishing life, in the context of Technology Ethics and Ethical AI more broadly, and a plea to work more towards this *practical* type of approach.

Acknowledgements

As this is an FFG-sponsored project, we would like to thank our funders for giving us the opportunity to develop this programme of work. We would also like to thank the team at Gradient Zero who have been keen and enthusiastic partners in this pursuit, in particular Jona Boeddinghaus who gave guidance and insightful comments on a late version of this paper.

References

- Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. and Chatila, R., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges to ward responsible AI. *Information fusion*, 58, pp.82-115.
- Alla, S. and Adari, S.K., 2021. What is mlops?. In *Beginning MLOps with MLFlow* (pp. 79124). Apress, Berkeley, CA.
- BBC News, 2020, 'Your data and how it is used to gain your vote': <https://www.bbc.co.uk/news/technology-54915779> (Last Accessed: 27/04/2022)
- Bostrom, N., 1998. 'How long before superintelligence?'. *International Journal of Futures Studies*.
- Coeckelbergh, M., 2020. *AI ethics*. MIT Press.
- Floridi, L. and Cowls, J., 2021. A unified framework of five principles for AI in society. In *Ethics, Governance, and Policies in Artificial Intelligence* (pp. 5-17). Springer, Cham.
- Friedman B, 2004. Value sensitive design. In: Bainbridge WS (ed) *Berkshire encyclopedia of human-computer interaction*. Berkshire Publishing Group, Great Barrington
- Guizzardi, R., Amaral, G., Guizzardi, G. and Mylopoulos, J., 2020, May. Ethical requirements for ai systems. In *Canadian Conference on Artificial Intelligence* (pp. 251-256). Springer, Cham.

- How, J.P., 2018. Ethically aligned design [From the Editor]. IEEE Control Systems Magazine, 38(3), pp.3-4.
- Huang, C.C. and Kusiak, A., 1996. Overview of Kanban systems.
- Kazim, E., Denny, D.M.T. and Koshiyama, A., 2021. AI auditing and impact assessment: according to the UK information commissioner's office. AI and Ethics, 1(3), pp.301-310.
- Madiega, T.A., 2019. EU guidelines on ethics in artificial intelligence: Context and implementation.
- Mökander, J. and Floridi, L., 2021. Ethics-based auditing to develop trustworthy AI. Minds and Machines, 31(2), pp.323-327.
- Mulvenna, M., Boger, J. and Bond, R., 2017, September. Ethical by design: A manifesto. In Proceedings of the European Conference on Cognitive Ergonomics 2017 (pp. 51-54).
- Tasioloas, J. (2021) The role of the arts and humanities in thinking about artificial intelligence (AI) | Ada Lovelace Institute
- The Independent, 2021, 'How racist robots are being used in recruitment': <https://www.independent.co.uk/news/world/americas/robots-racism-algorithms-jobhiring-b1860835.html> (Last Accessed: 27/04/2022)
- The Guardian, 2020, 'The Guardian view on A-Level algorithms: failing the test of fairness.': <https://www.theguardian.com/commentisfree/2020/aug/11/the-guardian-view-on-a-level-algorithms-failing-the-test-of-fairness> (Last Accessed: 27/04/2022)
- Raper, R., Boeddinghaus, J., Coeckelbergh, M., Gross, W., Campigotto, P. and Lincoln, C.N., 2022. Sustainability Budgets: A Practical Management and Governance Method for Achieving Goal 13 of the Sustainable Development Goals for AI Development. Sustainability, 14(7), p.4019.
- Shahriari, K. and Shahriari, M., 2017, July. IEEE standard review—Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In 2017 IEEE Canada International Humanitarian Technology Conference (IHTC) (pp. 197-201). IEEE.
- Umbrello, S. and De Bellis, A.F., 2018. A value-sensitive design approach to intelligent agents. Artificial Intelligence Safety and Security (2018) CRC Press (.ed) Roman Yampolskiy.
- Van den Hoven MJ, 2007. ICT and value sensitive design. In: Goujon P et al (eds) The informationsociety: innovation, legitimacy, ethics and democracy. Springer, Dordrecht, pp 67–73
- Wired, 2020, 'Police built an AI to predict violent crime: it was seriously flawed': <https://www.wired.co.uk/article/police-violence-prediction-ndas> (Last Accessed: 27/04/2022)

Responsibility by Design: Actionable strategies and a tool for leveraging technology ethically and enabling innovation responsibly

Veikko Ikonen¹, Emad Yaghmaei², Janika Miettinen¹, Giovanna Sanchez Nieminen¹

¹Technical Research Centre of Finland VTT Ltd, Tampere, Finland

²Yaghma, Delft, Netherlands

Veikko.Ikonen@vtt.fi

Abstract. In this paper we present practical measures for implementing ethics and responsibility to the development of new technologies, applications and services. We propose modifying the current Responsible Research and Innovation (RRI) approach regarding terminology, the conceptual and procedural approach, and practical implications in order to make it even more usable and useful for innovation ecosystems. Responsibility by Design as a strategy relates (besides RRI approach) to the Human-Centered Design, Human Driven Design and Ethics by Design approaches. As a standardized tool, Responsibility by Design offers a practical model and actions (i.e. roadmaps) for implementing interventions based on ethical approaches and RRI dimensions. It serves as a product and service which provides a focused view to the innovation process: how to integrate responsibility in the long-term developmental work of emerging innovations.

Keywords: Responsible Research and Innovation, Responsibility by Design, Ethics by Design, Human Driven Design, Engagement, Actionable strategy, Materiality Matrix, Ethics Audit

1 Introduction

In the context of technological innovation, we have conducted both ethical assessment and review. At the beginning, we mainly considered the research ethics and research integrity aspects when performing scientific studies and disseminating from those, but gradually we directed our focus to the design of technologies as applications and services, their expected usage context and potential users. Thus, it became clear for us that the ethical approach should not be limited to identifying current or future problems but should actively design for and be inspired by achieving more ethically and sustainable technical solutions. When searching for potential methods and approaches which could answer our needs, the Ethics by Design (EbD) approach seemed to be the answer at first, as it is a positive, forward-looking and proactive ethical thinking approach. When EbD is applied in a collaborative research project

that includes industrial stakeholders, ethical issues are taken into the negotiations and collaborative activities of the project early, with the aim of creating a positive, ethical, target-oriented mindset among project partners. (Nieminen et al. 2014a; Ikonen et al. 2015)

However, the EbD approach falls short on providing practical procedures and tools for its implementation in innovation ecosystems outside the public-funded research context. Thus, we propose to combine and further develop both EbD and the Responsible Research and Innovation (RRI) approach, in order to enhance their practicality beyond the European Commission (EC) funding research context. Our proposal is the Responsibility by Design (RbD) framework, as a collaborative approach in terms of promoting participatory and co-design methods to empower users and other stakeholders in the innovation design process, but also in terms of ensuring successful outcomes by means of a deep understanding of the different stakeholder's needs, values and circumstances which are context specific. RbD is responsible in terms of being aware of the human, societal and ethical values related to a particular design and reflecting them in the design. We aim to modify the current RRI approach in regard to its terminology, the conceptual and procedural approach, and practical implications in order to make it even more usable and useful for industrial innovation ecosystems.

In the next chapters we describe the both sides of the coin. In following chapters 2-6 we describe more profoundly development of RbD: Strategy approach. These chapters we will give the background and the current status of the approach. In chapter 7 *RbD: tool* a more detailed description of the tool will be provided. In last chapter both limitations and future development of the RbD will be discussed.

2 Ethics by Design

Ethics by design (EbD) is a future-oriented approach that concentrates on preventing ethically unacceptable outcomes emerging from research and development (R&D) activities and on boosting sustainable, acceptable and desirable outcomes. (Nieminen et al. 2014a; Ikonen et al. 2015). EbD examines responsibility in a particular context from different perspectives: in contextual design, one needs to understand the opportunities and limitations of technology, processes and/or behaviour and needs to balance and optimise the benefits and potential drawbacks. This is done by increasing knowledge among researchers and other people involved in R&D activities that involve stakeholders in development processes, creating structures that support ethical decision-making and using ethical assessment that supports the ethical requirements and assessment of responsibility-related themes. EbD should not be seen as a limiting approach that only seeks to find ethical problems – it should be seen as a way of finding new, more favourable solutions and innovations that society embraces. In projects like SniffPhone (2022), VOGAS (2022), SPARCS (2022) A-Patch (2022) and FRANCIS (2022), the EbD approach was developed and implemented in line with the responsible research and innovation (RRI) framework defined by the EC (Von Schomberg, 2013). When ethical, commonly agreed and shared values are

guiding the development of new technology from the start and stakeholders are engaged in the process, socially acceptable outcomes that include real value for users and for society are more likely to emerge.

SniffPhone, VOGAS and A-PATCH represent new concepts that address major societal challenges in the health and well-being of the general population while taking into account essential ethical and privacy aspects. These projects aim to achieve easy-to-use, non-invasive and widespread health screening for gastric diseases, especially cancers (through testing exhaled breath), and aim to identify and monitor tuberculosis via a smart patch concept. On the other hand, SPARCS ‘supports European cities in transforming into Sustainable energy Positive & zero cARbon Communities by creating citizen-centric ecosystems that are equipped to bring about meaningful change.’ (SPARCS, 2022) It means building up completely new areas and districts in the cities empowered by innovations and technology not only for energy production and for consumption but also for envisioning a very new way of living with services in these new or rebuilt areas. Finally, FRANCIS aims at the development of frugal innovations in open innovation challenges that involve citizens.

EbD is used in above mentioned projects to identify potential ethical or responsibility-related issues that arise during R&D activities and during the future usage of new technology to consider these aspects in the design from the very beginning of the development process and implement more sustainable solutions. Activities such as producing ethical design guidelines, producing future scenarios for the use of the device and smart systems, evoking discussions about the ethical aspects of these concepts and collecting professionals’ and citizens’ views on the technology have been and will be performed during the projects.

The ethical issues in the projects are two-fold: ethical issues in the research activities within the project (e.g. the ethical treatment of research subjects, responsible conduct of research) and ethical issues in the technology and applications (e.g. data security, privacy, targeted usage context and users) (European Commission, 2010). During the projects, ethical issues with regard to the project’s research activities should have been identified. The EbD approach has been used to entrench an ethical and responsible research culture in the projects’ participants. In addition, ethical guidance that relates to user studies within the project in particular, but also to the early stages of the technical development, has been provided. The technologies targeted e.g. to the early detection of diseases cannot be separated from the health care service system that is used. Similarly, a holistic approach is necessary when designing smart cities. This is why EbD should be used in concept design that contemplates the broader service solutions around the technology and applications.

EbD concentrates on anticipating and preventing ethically unacceptable outcomes emerging from R&D activities and boosting sustainable, acceptable and desirable outcomes. This is done by integrating ethics into design and development instead of adding it to existing frameworks in order to guarantee that new solutions, technologies and innovations follow the norms and values of society. It is important to increase both knowledge about ethical issues and discussion among researchers, research subjects and other people involved in R&D activities; to involve stakeholders – especially end users – in the design and development processes; to create structures

that support ethical decision-making; and to use ethical assessment that supports the ethical requirements. EbD should not be understood as a limiting approach that only highlights problems – it should be seen as a source of solutions that society embraces.

Legislation and standardisation – at national, EU and international levels – are an important part of securing safe and trustworthy products, environments and services by providing common rules for actors and defining minimum demands for research and innovation (R&I) activities and outcomes. However, legislation alone cannot guarantee that the outcomes of R&I and the direct and indirect effects of the undertaken R&I activities are ethically acceptable and sustainable. EbD goes beyond the legislation by embedding ethical thinking into design and development processes in order to produce solutions that include real value for the users and for society. Instead of asking ‘Is it legal to do this?’, we should ask ‘Is this the direction in which we want to go? What will be the effects of our actions on individuals, communities and society, and how do we achieve the best possible results?’ According to EbD, both positive and potential negative effects and impacts should be considered, and the focus should be on finding the best possible solutions instead of only pointing out ethical issues. The approach also highlights that, in order to find the best solutions, a variety of stakeholders (those who are directly or indirectly affected by the actions) should be engaged in the R&D processes from the start. This viewpoint is also underlined in discourse ethics, where ethical understanding should be constructed in interaction with people who are involved in or affected by the activities (Habermas, 1990).

When it comes to emerging technology, laws and regulations are often behind raising the importance of ethical reflection. In many cases, legislation also fails to give straightforward answers to ethical questions as, in many cases, the ethical issues are context dependent and unique. Laws are meant to set down common rules instead of offering guidance for specific and complex questions; they are provided more for preventing both unethical behaviour and already identified risks rather than for offering ethical advice, and even if laws are clear, this does not mean they are necessarily ethically correct. (e.g. Hankin 1923) This is why ethical reflection in the context of technology R&I is crucial. Ethical consideration is needed, especially in cases where the line between right and wrong, good or bad, is unclear and the drawbacks and benefits are somehow intangible and contradictory. As Reijers et al. (2017) argued, the complex and erratic nature of ethical issues in R&I is a consequence of changes in ‘people’s behaviours, socio-economic relations, power relations between people and institutions, and changes in the environment’. This makes it important to anticipate the changes in the environment and in the norms and values of society and individuals as a part of R&I.

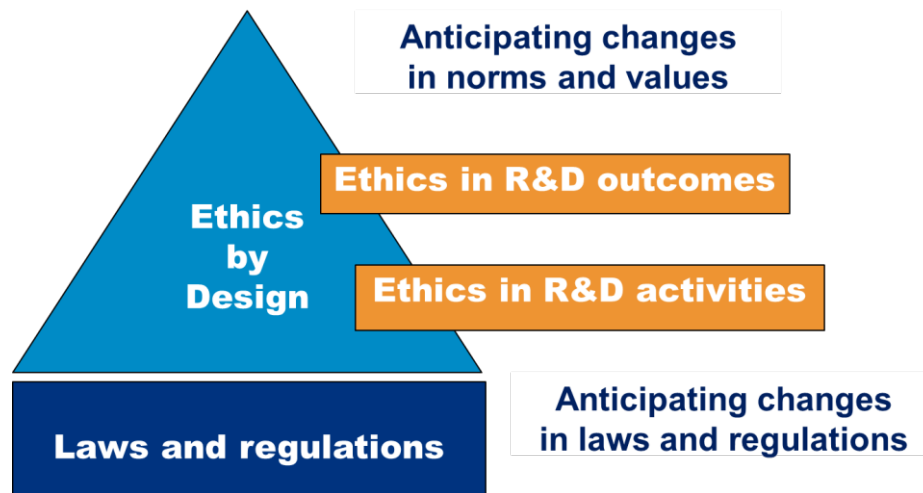


Fig. 1. Laws and regulations and the EbD approach as presented in relation to the SniffPhone project [2]

From the commercial perspective, it is essential to understand legislative demands when developing new technologies and innovations in order to avoid sanctions and to guarantee that the final products and services can be brought to market. Actors also seek competitive advantage by forecasting changes in laws and regulations in order to gain a position as a forerunner. However, fulfilling the legislative demands is not enough if the outcomes are not accepted and desired by society (e.g. genetically manipulated organisms). EbD is targeted to secure the success of new products and services by pursuing optimal solutions with minimum negative consequences for individuals, communities, the environment and society. This requires deep understanding of the social environment, people's norms and values (and even the conflicting values and interests of different groups and stakeholders in some cases) and the effects that one's activity can have on these. Forecasting changes in norms and values is also essential for innovation because an innovation takes time to deploy.

Ethical aspects can be transferred into design outcomes through ethical design processes. Therefore, ethical reflection needs to already be integrated into the very beginning of R&D activities. Borrett et al. (2017) underlined the importance of already embedding ethics in the design of a research proposal in order to gain an ethical research culture within the researchers. This research EbD can be seen as a part of the EbD approach in which the design of the project itself, included a strong emphasis on the 'ethical design' of the research proposal, will be actualised. In addition to the ethical project design, ethical consideration of the emerging technology and the possible services around the emerging technology should be considered. (Borret et al. 2017)

3 Ethics in development and innovation

The RRI framework [6] and the EbD approach have been integrated as a part of the above-mentioned projects, which means that projects' actions will go beyond the obligatory requirements (e.g. legislative requirements). The European Code of Conduct for Research Integrity will guide the research activities. According to this conduct, the fundamental principles of research integrity are (Allea 2017):

- Reliability in ensuring the quality of research, reflected in the design, the methodology, the analysis and the use of resources
- Honesty in developing, undertaking, reviewing, reporting and communicating research in a transparent, fair, full and unbiased way
- Respect for colleagues, research participants, society, ecosystems, cultural heritage and the environment
- Accountability for the research (from idea to publication) and its management and organisation, for training, supervision and mentoring, and for its wider impacts.

According to the EbD approach, all our projects base activities on the high involvement of stakeholders (laymen and citizens, healthcare professionals, technology developers, city planners, policymakers and researchers) in order to understand and implement the expectations and demands as a part of the R&D activities and technology design. There is a strong industry involvement in these projects as companies are project partners and see the benefit of the collaboration in an open innovation ecosystem. Deploying a more systemic approach calls for collaboration where stakeholders work and share together. Without taking into account the limitations and challenges, as well as the opportunities and possibilities, of the different viewpoints, it is hard to bring together theoretical, practical and methodological approaches that envision an improved or better solution to the identified challenge or issue.

Apart from the effects that research and innovation can have on individuals, wider societal effects also needed consideration. These included asking questions such as Who has access to these technologies? Does the new technology support equality? and Is the technology socially acceptable? In order to answer the wide range of ethical questions, it is important that an ethical way of thinking is implemented as a part of the project; ethical and RRI training is offered to the re-searchers and other R&D parties in order to identify and comprehend ethical issues and in order that stakeholders are engaged in the R&D activities in an authentic manner.

According to Reijers et al. (2017), the methods for practicing ethics in R&I can be categorised as ex ante, intra and ex post methods. Ex ante methods refer to situations where practising ethics is started from the very beginning of the R&I process when the actual concept and design are not yet formulated. These methods, including foresight methods and scenario building, are therefore targeted to foresee the ethical impacts of emerging technologies and to identify the possible ethical issues beforehand. Intra methods focus on the stage of R&I where technology design is realised, including prototype building and testing. Intra methods are targeted to integrate the ethical values into the design of a new technology. Differing from the ex

ante and intra methods, ex post methods aim at identifying and analysing the ethical impacts of existing technologies with a retrospective approach. These methods are not used in the design phase of emerging technologies but later on, in order to identify and evaluate the ethical impacts of technologies (e.g. developing checklists based on the gained knowledge on ethical issues). (Reijers et al., 2017),

The methods used in EbD can be described as mostly a combination of ex ante and intra methods. It is important to integrate ethical design thinking in the project from the start and to secure that not only the technical aspects but also the ethical and societal aspects are taken widely into account. The ethical aspects of technology design (e.g. the need for the built-in protection of user data and privacy) should already be recognised at the beginning of the project. In addition to ethical technology design, the EbD approach can be used for ethical service design. Ethical service design is especially crucial with health and well-being technologies which cannot meet their targets without being connected to other health care services. However, fundamentally, all our R&I activities should take into account health and well-being aspects, regardless of the actual technology or application. Thus, ethical thinking and EbD should be an integral part of any kind of innovation and design activity.

According to the future-oriented EbD approach, ethics will already be included in the projects in the conceptualisation phases and sustained throughout the whole process (Niemelä et al. 2014a). An ethical governance model will be used to define a system of decision-making procedures for the whole project and insert the ethical aspects and factors into this decision-making system. Project partners with ethical expertise should be included in the project consortium to provide ethical support for all stages of R&D. The EbD and human-driven design (HDD) approaches should be included to the projects' tasks to make sure that the ethical and societal considerations flow through the project. An ethics advisory group, including external ethics and responsibility partners, will help to formulate ethical recommendations, guidelines and dialogue between stakeholders.

4 Human-driven design (HDD)

An important part of any technological project is to understand the needs and requirements of the target user groups and implement these as part of the technology development. This HDD approach includes identifying the target users of the technology, application or service and involving them in the development process. The scope is to understand human and social values and needs as an integral part of the design process. HDD is an important part of the EbD approach wherein ethically acceptable results are pursued through collaboration. The intention of HDD is to foster user satisfaction, accessibility, acceptance and, of course, to highlight the possible negative effects of use on human health, safety, well-being and performance. Instead of focused product development, HDD broadens the design perspective in a more holistic direction, enabling multiple perspectives as a part of the design.

As a term, *human-driven design* was introduced by Braund and Schwittay (2006) and Ikonen (2009). Brand and Schwittay brought out four dimensions that should be

taken into account in developing information technologies for developing regions: local practices, participatory design processes, socio-cultural contexts and political conditions. Furthermore, they emphasised that so-called rapid ethnography is not the right approach in these research contexts – what is needed is long-term participant observation. Ikonen presented HDD as an approach to the design of future smart environments and emerging ICT and describes HDD as an ‘approach which broadens the perspective from a focused product or service development process model to the more holistic design perspective’. Ikonen also described stakeholder-based design that ‘furthermore broadens the scope and role of involved participant groups in the actual design process’ (Ikonen 2009). In addition to this, Ikonen called for ethical assessment to continue throughout the design process. Ikonen called the resulting combination empowering design (Ikonen 2009). Based on further development of the concept, it was defined that HDD should integrate the following three endorsed perspectives: it should be holistic, it should strive for collaboration with different stakeholder groups and it should include an anticipatory approach and be ethically reflective. (Niemelä et al., 2014b) In the midst of many large-scale societal and technological transformations, there is a need for design approaches that are able to integrate multiple perspectives in design in order to work towards outcomes that are interesting, feasible and sustainable in all senses of the term. For this purpose, an approach is proposed for designing technology that is driven by human and social values; collaborative in nature; strongly future oriented with a special view of emerging, enabling technologies; and reflective in terms of responsibility and ethics in design. (Niemelä et al. 2014a) This HDD approach focuses on various phases of design and aims to produce technologies, applications and services for a more sustainable future society and a better quality of life for all citizens. (Niemelä et al. 2014b) A recently launched approach called Extreme Citizen Science (ECS) (Haklay 2012) connects similarly to empowering design practice while bringing out an authentic, bottom-up approach: in this, the highest level of citizen science research would be initiated and carried out by laypeople themselves or in equal collaboration with scientists.

5 Recommendations for implementing EbD

EbD is a proactive approach that relies on the involvement of and dialogue with a wide range of stakeholders who are directly or indirectly affected by R&D actions and outcomes in order to provide ethical, innovative and acceptable out-comes. The approach needs to be integrated into the design process from the very beginning, starting from the point where tentative ideas are being translated into actual design outcomes, leaving room for the changes identified during the ethical design process. EbD should be tailor-made and flexible, however, it should include the possibility of applying already known methods, tools and practices (e.g. ethical guidelines, foresight methods, ethical impact assessment) (Yaghmaei and van de Poel, 2021)

It is crucial that all project participants are aware of the ethical design approach and in this way foster the processes of mutual learning and co-development wherein

stakeholders' values, norms and expectations – especially citizens' and end users' values, norms and expectations – are used as the basis for technology design. Project partners or external experts specialised in the ethical and societal aspects of emerging technologies are to guide the ethical design process. Their job is to raise ethical awareness within the project partners and stakeholders, govern the ethical work and support the ethical design process during the project. This is enabled by integrating ethical design as a part of the project structures. The researchers should also have an important role in entrenching ethics in their field of excellence. Therefore, ethical reflection should be made with them and not for them.

Figure 2 represents the EbD approach as seen in the context of our experiences in Horizon 2020 projects. Even though EbD can be understood as the work done in the design phases to make the actual design of the target technology, product or service ethically sound, it can also be seen to include the activities that support and enable the ethical design process itself. This broader conception of EbD meets the projects' approach is based on the idea that the ethical design process will not just happen – it needs to be actively pursued throughout the project. Hence, the ethical design approach covers the project design, the technology and product design, the concept and service design, and the ethical anticipation of wider social and societal effects.

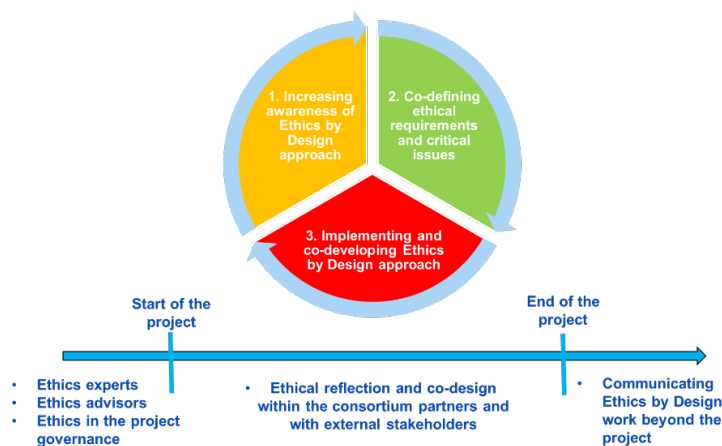


Fig. 2. EbD within a project timeline

Based on the project experiences, certain actions are especially important in supporting EbD, especially in the area of emerging health technology R&D. These include involving ethical experts (either project partners or external experts) in the work of the project; designing a project structure that supports ethical design; raising the ethical awareness and supporting the ethical reflection of project consortium members and project stakeholders; and evaluating and communicating the ethical work and its outcomes inside and outside the project consortium during and beyond the project. The ethical reflection of project consortium members and project stakeholders helps to identify real ethical challenges as well as the norms and values

that are to be followed. The increased ethical understanding also makes it possible for every project member to apply ethical practices in his or her field of expertise, which also strengthens the ethical re-search culture. The communication of ethical work and its outcomes increases the knowledge of case-specific ethical challenges and the benefits of the EbD approach, which can be used for the benefit of future R&D work.

At the core of the EbD of technology and service R&D are: (1) stakeholder engagement activities that help to understand ethical issues and demands, (2) the definition of ethical requirements based on the stakeholder engagement activities and (3) the implementation of ethical requirements in the actual technology, product and service design (see Figure 2). These stages follow each other cyclically until the ethical aspects are comprehensively recognised and genuinely implemented in the final technology, product and service design. The needed stakeholder engagement activities should be defined separately for different parts of the ethical design process, starting from the very beginning where the actual concept and design are not yet formulated. The activities should also be flexible to allow for changes based on the results of earlier activities.

6 How to customize and enhance ethics and responsibility approaches beyond EC -funded R&D projects

In EC-funded R&D projects, the funder may incentivise applicants to follow procedures, guidance and recommendations. This may be done in the description of the call for proposals or it may be one of the cross-cutting criteria that are to be evaluated in a funding review. RRI keys thus fit quite well with this kind of industrial ecosystem's innovation structure and process, which calls for large stakeholder engagement and quadruple helix approaches. Ethics, engagement, gender, open science and science education can usually be customised to the project's governance structure and activities quite easily on some level. An ethical approach that mostly follows laws and regulations is naturally well appreciated, but an ethical approach that goes beyond a legal approach may seem artificial to them in this context. Thus, the question is how one can dig up more meaning and reasonable concrete actions in relation to RRI keys which quite often puzzle the partners of the consortium (i.e. the industrial ecosystem). The modified Responsibility by Design approach should then incentivise ecosystems however they are funded (e.g. EC, private foundation or self-funded). The benefits of taking ethics and responsibility into account in an industrial ecosystem's innovation should be presented in such a way that is seen as bright, logical and adjustable (as not all ecosystems or innovation projects are the same). A more systemic approach should picture 'using an integrated framework for convergent product development and a model will be developed that links the industrial ecosystem stakeholder and customer "value perspectives" via a business model to the value network and required capabilities' (Phillips 2015). Naturally, this value network and business model should thus include responsibility-related perspectives and criteria in order to build up the most sustainable industrial innovation ecosystem.

We propose a slightly different terminology, conceptual and procedural approach, and practical implications with the RbD approach. We base our approach and novel terminology on our experiences when interacting with various stakeholders in R&I projects. Our six E's concept includes embedding the whole RbD (including elements from HDD and EbD) approach while also broadening the perspective to be even more holistic— which should support the easier customisation of the responsibility aspects to specific projects and should offer practical procedures and tools for implementing RbD approach in industrial innovation ecosystems. The six E's are: enlightenment, engagement, ethics, empowerment, evaluation and execution.

The starting point is enlightenment. This means that the project, innovation action and ecosystem have a true vision of their goals and targets. A shared vision of a more sustainable solution needs to be built up together. That is why engagement has a vital role. Balanced stakeholder engagement is challenging, but without taking into account the diversity (not only in terms of gender) of various stakeholders and target groups, all the other actions might just be consultation actions, trying to justify participation, which nowadays, in many cases, is also required by regulation. True engagement enables commitment to the planned action and promotes the plan's acceptance and desirability. It should happen in a co-creative manner, taking into account all the aspects in both the short-term and the long-run perspectives.

Ethics and legal aspects need to be attached to the innovation process from the beginning. Laws and regulations have to be followed, or at least one needs to be sure if something can be done as an exceptional pilot activity beyond current legislation. As mentioned earlier, while law and ethics are not the same, it is important to use an ethical approach as a tool for understanding the challenges that might be relevant when developing new solutions for specific purposes. What might seem to be a very reasonable solution at first may, in the end, violate some basic human rights or people's understanding of autonomy or dignity.

Related to the previous E's (enlightenment, engagement and ethics), we consider empowerment to be equally important. In order to be able to be engaged, committed, share the vision and have power in a participatory space, participants need to feel empowered to do so. This is related to the science education, open science and capacity building of RRI keys, which means that, in many cases, the process cannot be ad hoc and short-term action but must be well planned, continuous collaboration – long-term 'partnership' – in order to achieve the trust and common vision among ecosystem participants about targets and goals.

Evaluation is a key element in the monitoring of the plans, actions and process at a general level. It should have clear milestones and criteria for decision-making as, in many cases, the ecosystem needs to adjust the plans during the projects or even make the decision to cancel the project for reasons that were not seen in the beginning. That is why the sixth E is also very important. Execution means that the ecosystem has structure and governance which cover all the procedures for activities and decisionmaking of the previous Es and can be flexible in cases where unpredicted consequences happen, or new factors appear on the developmental horizon of an innovation ecosystem.

Our aim is not to replace the RRI keys defined by the EC or to suggest rival conceptualisations for some other model on a responsibility or sustainability theme. If anything, we would like to incorporate elements from RRI, sustainable development goals or some other approaches in the RbD approach and in the six E's. We believe that a suitable mix of all these can be modified into a practice-oriented procedure that takes into account specific, context-dependent issues. The six E's will be further elaborated during these on-going and future R&D activities.

7 RbD: tool

The Responsibility by Design toolkit ensures that ethical values are taken into account at the design and development level of innovation and innovation activities. Based on a method called the "materiality matrix" the RbD Tool helps to identify and define ethical and responsible values for an organization and its stakeholders. Through the materiality matrix, an organization can come to understand the ethical and social priorities most important (and urgent) for their stakeholders. However, to obtain the ethical values most pertinent to the organization, an ethics audit is performed. This audit refers to an organization undergoing a review of potential or existing ethical issues residing within their organization or in a project. The purpose of the review is to encourage the organization to consider impacts they or their project may have on their stakeholders and to assure their (best) interests are being met. Such a commitment to an ethics review reveals to stakeholders that the organization is concerned with their values and are in line with their expectations.

Following the process of the ethics audit, stakeholders should be, in principle, engaged to reflect on the ethical values that arose from the audit. Here, the stakeholder can help to identify the most important and urgent ethical values. Through a rating system, stakeholders can directly quantify the importance of each value and issue topic. It is suggested that these stakeholder findings should be concretely and systematically taken into account when designing and developing technology, solutions or innovation.

In short, the materiality matrix method has three main benefits. The first allows for the integration of stakeholder values and priorities at the center of early decisionmaking in innovation processes. Secondly, by identifying and defining these priorities, organizations and project leaders are able to pinpoint and refine their focuses in accordance with the importance and urgency of the identified values. Thirdly, it allows for the possibility to reflect and externalize, either in a report or through actions, how ethical and responsible values are incorporated into innovative actions.

8 Conclusions

The EbD approach and the further-developed RbD approach strongly emphasise that new technologies and innovations should not be designed apart from the total system and environment in which they will operate. The EbD approach is especially crucial

with health technologies – such as those used in SniffPhone, VOGAS and A-Patch – that cannot meet their targets (including the better chance of early treatment and survival through the early recognition of disease) without being connected to other health care services in a particular service system. However, projects like SPARCS and FRANCIS – a projects which aims to create ‘an information society that places citizens at the centre of the decision-making process and increases general public awareness of the drive towards sustainability in the city’ or aiming to implement extreme citizen science approach – have the same challenges. This means that at the beginning of the technology development process, broader ethical issues already need to be acknowledged and discussed in addition to ethical technology design and research ethics in order to avoid developing technologies that may result in issues within the system they are planned to be used within. In light of this, the EbD approach offers ways of ethical engagement that make it possible to consider both the R&I activities and outcomes, as well as their wider consequences, from an ethical and societal perspective. Our future work aims to broaden the EbD approach to be even more practical beyond EC-funded projects. In the RbD approach, the six E’s provide a methodology and tools for the more holistic planning and execution of the development of technologies, applications or services in innovation ecosystems that include all the relevant stakeholders in an empowered manner. We propose six E’s: enlightenment, engagement, ethics, empowerment, evaluation and execution. Our six E’s concept includes embedding the RbD approach – which should support the easier customisation of the responsibility aspects to specific projects and should offer practical procedures and tools for implementing RbD – in any innovation ecosystems.

At the core of the EbD of technology and service R&D are: (1) stakeholder engagement activities that help to understand ethical issues and demands, (2) the definition of ethical requirements based on the stakeholder engagement activities and (3) the implementation of ethical requirements in the actual technology, product and service design. These stages follow each other cyclically until the ethical aspects are comprehensively recognised and genuinely implemented in the final technology, product and service design. The needed stakeholder engagement activities should be defined separately for different parts of the ethical design process, starting from the very beginning where the actual concept and design are not yet formulated. EbD needs to be well integrated to the more general design approach. The ultimate goal would be authentic, democratic, continuous interaction between various stakeholder sharing the same design vision. Thus, revised RbD approach (terminology, tools, procedure, model) may support this attempt. In this perspective, RbD should be considered as a holistic, and more systematic, approach in the research and innovation of future solutions.

Acknowledgements

SniffPhone has received funding from the European Union’s Horizon 2020 R&I programme under grant agreement No 644031, VOGAS grant agreement No 824986,

A-Patch grant agreement No 824270, SPARCS grant agreement No 864242 and FRANCIS grant agreement No 101006220. We would like to thank all the members of the projects, the external ethics advisors and other participants for their contributions to the RRI and SDG exercises and other participatory actions of the projects.

References

- ALLEA - All European Academies. (2017). The European Code of Conduct for Research Integrity.
- A-Patch (2022), Retrieved 15.5.2022 from <https://apatch.technion.ac.il>
- Borrett, D. S., Sampson, H., and Cavoukian, A. (2017). Research ethics by design: A collaborative research design proposal. *Research Ethics*. vol. 13(2). pp. 84–91.
- Brand, P. and Schwittay, A. (2006) The Missing Piece: Human Driven Design and Research in ICT and Development. In *Proceedings of the International Conference on Information and Communications Technologies and Development*, May 2006
- European Commission. (2010). *European Textbook on Ethics in Research*. Publications Office of the European Union. Luxembourg.
- Francis (2022), Retrieved 15.5.2022 from <https://www.francis-project.eu/>
- Habermas, J. (1990). *Discourse ethics: Notes on a program of philosophical justification*. In *Moral consciousness and communicative action* (pp. 43–115). Cambridge, MA: MIT Press
- Haklay, M., (2012): Citizen Science and Volunteered Geographical Information. In: Sui, D.Z., Elwood, S., Goddchild, M.F. (eds.): *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*. Berlin: Springer, 105-122
- Hankin, G. (1923). *Ethics and Law*. *International Journal of Ethics*, 33(4), 416–435. <http://www.jstor.org/stable/2377596>
- Ikonen, Veikko. (2009) *Towards Empowering Design Practises*. *Ambient Intelligence and Smart Environments*, Vol. 1, *Ambient Intelligence Perspectives, Selected Papers from the First International Ambient Intelligence Forum 2008*. Mikulecky, Peter et.al. (ed.). IOS Press. Amsterdam, 129 -136.
- Ikonen, V., Kaasinen, E., Heikkilä, P., & Niemelä, M. (2015). Human-driven design of micro and nanotechnology based future sensor systems. *Journal of Information, Communication and Ethics in Society*, 13(2), 110–129. <https://doi.org/10.1108/JICES10-2013-0039>.
- Niemelä, M., Kaasinen, E., & Ikonen, V. (2014). *Ethics by design - an experience-based proposal for introducing ethics to R&D of emerging ICTs*. Paper presented at ETHICOMP 2014 - Liberty and Security in an Age of ICTs, Paris, France.

- Phillips, M.A. (2015). Understanding emergent innovation ecosystems in health care. In proceedings of POMS 2015. Washington, May 2015
- Reijers, W., Wright, D., Brey, P., et al. (2017) Methods for Practising Ethics in Research and Innovation: A Literature Review, Critical Analysis and Recommendations, Science and Engineering Ethics, pp. 1–45.
- SniffPhone (2022), Retrieved 15.5.2022 from <https://www.sniffphone.eu/>
- SPARCS (2022), Retrieved 15.5.2022 from <https://www.sparcs.info/>
- VOGAS (2022), Retrieved 15.5.2022 from <https://www.vogas.eu/>
- Von Schomberg, R. (2013). A Vision of Responsible Research and Innovation. In R. Owen, & J. e. Bessant, Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society (pp. 34-48). Chichester: Wiley
- Yaghmaei, E., & Van de Poel, I. (Eds.). (2021). *Assessment of Responsible Innovation: methods and practices* (1st ed.). Abingdon-on-Thames: Routledge.

Factors contributing to the intention to use exoskeletons in the workplace – exploring the role of ethics

Stéphanie Gauttier ^[0000-0003-3285-9189]

Grenoble Ecole de Management, Grenoble, France

stephanie.gauttier@grenoble-em.com

Abstract. Given the hard conditions of workers in railway maintenance, initiatives to develop robotic and wearables solutions to alleviate workers' from muscular strain and fatigue have developed. In particular, exoskeletons are being introduced into the workplace. This research investigates the ethical issues in this regard. A review of the applied ethics literature on exoskeletons is proposed, which highlights issues related to access, performance assessment, wellbeing at work, identity, long-term consequences, and safety. This work is the first step of a wider research project, in which workers' perspective on exoskeletons is surveyed quantitatively and qualitatively, as well as their first use of the exoskeleton.

Keywords: Exoskeletons, Physical Enhancement, Wearables, Ethics, Technology Acceptance

1 Introduction

Although many industries have seized the potential of robotic technologies, the railway sector is lagging behind, with most of the work still performed manually by workers and human-operated machines (Kefalidou et al., 2018). Performing tasks in this way, railway maintenance operators have to make significant physical effort to carry around the weights involved in their daily tasks, such as when replacing sleepers. Over time, this can have a negative impact on their health and bodies and lead to sick leave. For instance, the physical demands of the jobs mean that Musculo-Skeletal Disorder (MSD) affects up to 38.1% of workers during their careers (Schneider et al., 2010). Besides, many tasks have to be performed at night, to allow for daily traffic, which increases the difficulty of the tasks to be performed. Manual railway workers and machine operators are constantly subject to a varying degree of cognitive burden, which undermines the quality of their work, repeatability, and productivity. This can potentially lead to a significant risk of accidents caused by human errors (Fan et al., 2018; Karakhan et al., 2019). In turn, this leads to productivity losses.

Railway maintenance companies also face difficulties in terms of managing the workforce: young people are not necessarily attracted to such jobs, but in southern

Europe, over 50% of the workers in the railway industry were over 50 years old (European Commission, 2019).

Exoskeletons can assist workers and counterbalance the issues described above. They are to assist the workers in their daily tasks, reduce the strain on the body, sensation of efforts required, or fatigue (Bances et al., 2020). Ultimately, the weight that can be lifted should go from x_1 to x_2 for a stable amount of effort y . This narrative presents the use of exoskeletons as a help to ensure the safety and health of workers, their overall well-being, and the productivity of companies. However, the literature in ethics highlights ethical issues linked to exoskeletons due to their enhancing potential. To date, there is little research about the use of physical enhancement devices in the Management Information Systems literature, even though prior research has shown that ethics plays a role in the acceptance of wearables and insideables (de Andres-Sanchez, 2021). This research aims to understand the role of ethical considerations in shaping acceptance of exoskeletons at work. A first step of this research project involves a review of the literature on the topic, analyzing its conclusions in the perspective of using the devices for work purposes. This analysis is then used to create a survey to gather workers' perception of exoskeletons, of which the results are pending.

2 Exoskeletons in the applied ethics literature

While exoskeletons are presented by industry as assistive devices, they can be conceptualized as devices for physical enhancement insofar as they allow users to do better or more with their natural physical abilities. Exoskeletons support movement, substitute part of the work done by the body, to reduce muscular strain, fatigue, and effort.

A traditional criticism towards enhancement technologies is the risk of dependence (Greenbaum, 2015), leading to using devices for things one could do naturally. A second common concern is access (Greenbaum, 2015). Enhancement should be offered to all workers. Yet, workers are diverse in terms of age, weight, body shape, history of injuries, and other forms of disability. Exoskeletons may not be designed to accommodate so many different profiles, they could be perceived as enabling care to a specific population group. Exoskeletons might reinforce issues of accessibility for disabled workers if they are designed taking into account only non-disabled populations (Davis, 2012). In case of exoskeleton shortage, criteria to decide who can use one are needed.

Going further, there is a risk that workers might be tasked with lifting heavier objects than they normally would, since the exoskeleton absorbs part of the weight of the objects. This transforms the human into a machine that can be equipped to do more, treating it as a means more than as an end. The risk of being asked to carry heavier loads than without the exoskeletons is also minimal insofar as overworking of workers is prohibited (Greenbaum, 2015; Vrousalis, 2013). However, there are questions related to the fair wage of workers as their productivity is meant to change with the use of the device (Gauttier, 2017), even though it is unclear how this relates

to the ability of the worker or of the machine, and how it differs from the work of non-users. Moreover, there is a risk that individuals who do not want to use exoskeletons might be deprived from some work opportunities, a concern shared by laymen (Jensen et al., 2020). The question of the classification of the equipment is important to solve these issues: should the exoskeleton be seen as a tool to protect the worker and its use made compulsory, no discrimination can be made between users and non-users. However, this removes individual consent to use something that can affect one's identity and lived experience of the body (Kapeller et al., 2020). The literature highlights that exoskeletons are related to the embodiment of norms considering what a healthy and performant body is (Pedersen and Mirrlees, 2017), which can then have an impact on one's quality of life.

The introduction of smart and enhancing wearables can affect wellbeing at work: it may reduce collaboration and direct communication between individuals. Users can be perceived by their fellow workers differently from those who cannot or choose not to use the exoskeleton (Pedersen and Mirrlees, 2017; Breen, 2015). The ergonomics and look of the exoskeleton can reinforce a relation of alterity to users of the exoskeletons, or prevent such a relation to emerge, endangering one's sense of humanity and togetherness. There are issues related to a perceived diminution of agency and responsibility (Fischer and Ravizza, 2000). The movement is shared by the individual and or exoskeleton (Cornwall, 2015) and potentially a lack of congruency between the two could occur, making difficult to attribute responsibility to one of these agents only. Workers need to understand how they share and take responsibility, allowing them to be in control. Given the duty to ensure workers' safety, this symbiotic relation should be breakable only for safety purposes and with the worker's awareness.

The exoskeletons should be protected against misuse by design, even though misuse can also occur through the use of the device in a new context, for instance out of the work site (Greenbaum, 2015). Transparency on data processing, to monitor productivity, use, or other purposes need to be transparent, to minimize the risks linked to negative feelings of control linked to smart wearables.

Going further, exoskeletons can be seen as changing the relationship between employer and employee. Indeed, while employees traditionally lend their working abilities, including physical ones, and time to their employer in exchange for a salary, exoskeletons suggest that employees have to surrender their body to a technology to be able to work. Such actions imply that the employees are not sufficient on their own and need to be completed through technology. Exoskeletons can be passive and store energy and release it during works. They can also be active, using motors, hydraulic or pneumatic systems to enhance human strength and actively augment the power of the body. Active control modes provides the necessary movement to the human limb by the exoskeleton. Such exoskeletons illustrate this notion that the human body is a support for the technology. The role of humans at work is at stake. Despite the noble intentions behind the use of the exoskeletons described in the literature in terms of healthcare and increasing security, the use of exoskeletons goes against the humanity principle, which posits that humans should be seen as an end in themselves, and not as means. Such concerns are expressed in the literature on industry 4.0, which discusses

the complementarity between robots and humans (Sherwani et al, 2020; Nardo et al., 2020). According to Lyytinen et al. (2020), metahuman systems are arising, which are machines able to join human learning and create new systemic capabilities. Such systems require new management styles and task divisions. They imply that human workers are hybrid by nature. Therefore, investigating workers' perception of exoskeletons requires considering their identity at several levels: as individuals, as workers, as humans.

Besides, while the literature focuses on assessing the usefulness of exoskeletons in terms of effort and fatigue reduction, it does not provide knowledge as to what the workers have to change in the way they do their work and the way they move. Exoskeletons being literally on the body, they may be changing the way one feels strength and feelings while in movement. The tasks to be performed by workers in railway maintenance occur on the tracks, which are narrow spaces. Having to learn to move in new ways in such a context could bring about new security issues. This leads to potential ethical trade-offs to be settled between health and safety on the one hand, and between human dignity and security on the other.

For organizations, workers, and society to be able to decide whether exoskeletons are acceptable devices, it is necessary to move from a theoretical literature and assess the perception and experience of exoskeletons of workers in context.

3 Next steps

Analyzing the acceptable character of exoskeletons requires a multi-level investigation. First, a quantitative survey of railway workers is currently conducted online to gauge the relationship towards these different ethical issues (in green on the figure) and intention to use the technology. It is administered in organizations that have taken an active role in the project itself. Firstly, data was collected in organizations based in Spain and Italy and which are set to participate in real-life testing of the exoskeleton. Given the small number of individuals concerned with the exoskeletons in these highly specialized companies and a difficulty to reach a number of answer allowing for statistically significant results, an additional round of data collection is taking place in similar organizations in the United Kingdom, Germany. The survey is translated and conducted in the national language of workers, to enable maximum participation.

The survey accounts for subjective considerations as well as traditional constructs of technology acceptance (Davis, 1989; Venkatesh et al., 2012). A summary of these dimensions is proposed in Figure 1, with traditional elements in blue and those stemming from the ethics literature shown in green in Figure 1.

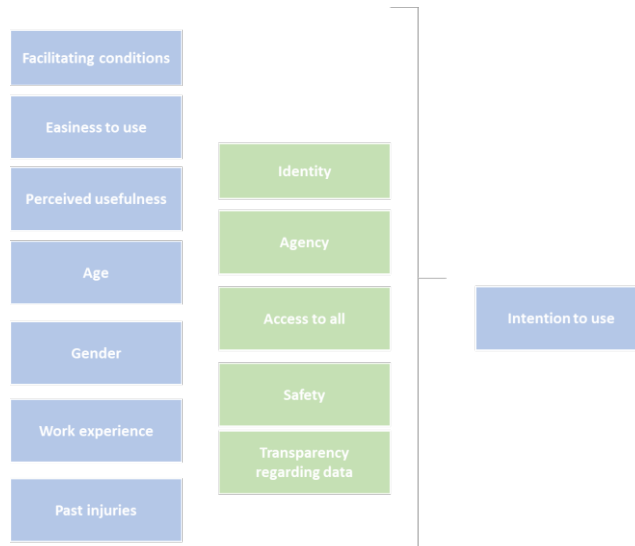


Figure 1. Framework for the empirical investigation

Second, a qualitative study following Q-methodology (Stephenson, 1935; 1953) is performed with workers who experienced active and passive exoskeletons while participating in lab experiments. The workers are presented with 25 statements about the way they experience their work and the tools given to them. The statements focus on the impact that tools have on the body from the perspective of health, of felt strength, of embodied knowledge and gestures, of feeling sufficiently equipped, etc. Workers have to classify the statements according to the degree with which they agree or disagree with them. First, they do this exercise thinking about the tools they traditionally use. Then, they repeat the exercise thinking about their experience with the passive, and then active exoskeleton. This data is analyzed at the individual level. The aim is to understand whether each individual perceives traditional tools differently from the exoskeletons. Then, the data is analyzed at the aggregated level to describe the different points of view held by workers when it comes to working with and without exoskeletons. This step enables collecting in-depth data about embodiment, which is a dimension of technology experience that is not key to acceptance model that is critical to understand how acceptable exoskeletons are.

Finally, this research project includes tests to be conducted in real-life settings. This will enable collecting data to assess the impact of the exoskeleton onto the workers as well as the way in which they perform their tasks looking at indicators such as time spent, accuracy, or errors. Together with the quantitative and qualitative steps described above, this information will be used to deliver a cost-benefit analysis, considerate of economic and financial arguments. This cost-benefit analysis will also consider a precautionary approach including risk assessment, uncertainties, and value changes linked to the introduction of the devices (Stirling and Coburn, 2018). It will enable

organizations as well as other stakeholders to reflect on the acceptability of exoskeletons beyond productivity parameters.

4 Conclusion

The ethical considerations on the use of exoskeletons found in the literature echo traditional debates on human enhancement. These ex-ante analyses do not consider the perspectives of users. To understand the (in)-acceptable character of exoskeletons for railway workers, it is necessary to study the traditional factors of technology acceptance as well as dimensions of experience linked to the characteristics of exoskeletons as embodied devices. Based on qualitative and quantitative approaches, we can identify key drivers and barriers. Looking at actual use, we assess the performance of workers using exoskeletons. Taken together, these data enable discussing the costs and benefits of exoskeletons in an applied manner, providing information for society to decide whether such devices are acceptable or even desirable.

References

- de Andres-Sanchez, J., Arias-Oliva, M., & Pelegrín-Borondo, J. (2021). The influence of ethical judgements on acceptance and non-acceptance of wearables and insideables: Fuzzy set qualitative comparative analysis. *Technology in Society*, 67, 101689.
- Bances, E., Schneider, U., Siegert, J., & Bauernhansl, T. (2020). Exoskeletons towards industrie 4.0: benefits and challenges of the IoT communication architecture. *Procedia manufacturing*, 42, 49-56.
- Breen, J. (2015). The exoskeleton generation—Disability redux. *Disability & Society*, 30 (10),1568-1572.
- Cornwall, W. (2015). In pursuit of the perfect power suit. *Science*, 350 (6258), 270-273.
- Davis, J. (2012). Progress versus ableism: The case of ekso—Cyborgology. *The Society Pages: Cyborgology*, 1-6.
- Davis, D. F. (1989) “Perceived usefulness, perceived ease of use, and user acceptance of information technology.” *MIS Quarterly* (13:3), pp. 319-339.
- European Commission. (2019). Report From The Commission To The European Parliament And The Council: Sixth Report On Monitoring Development Of The Rail Market.
- Fang, Y., Cho, Y. K., Durso, F., & Seo, J. (2018). Assessment of operator's situation awareness for smart operation of mobile cranes. *Automation in Construction*, 85, 65-75.
- Fischer, J., & Ravizza, M. (2000). *Responsibility and control: A theory of moral responsibility*. Cambridge: Cambridge University Press.
- Gauttier, S. (2017). Vers une “entreprise augmentée”: De nouveaux challenges pour la recherche en management et systèmes d'information. *Terminal. Technologie de l'information, culture & société*, (120).

- Greenbaum, D. (2015). Ethical, legal and social concerns relating to exoskeletons. *ACM SIGCAS Computers and Society*, 45(3), (pp. 234-239).
- Jensen, S. R., Nagel, S., Brey, P., Kudlek, K., T, D., Oluoch, I., . . . HET, S. (2020). SIENNA D3.4: Ethical Analysis of Human Enhancement.
- Kapeller, A., Felzmann, H., Fosch-Villaronga, E., & A.M, H. (2020). A Taxonomy of Ethical, Legal and Social Implications of Wearable Robots: An Expert Perspective. *Science and Engineering Ethics*, 1-19.
- Karakhan, A., Xu, Y., Nnaji, C., & Alsaffar, O. (2019). Technology alternatives for workplace safety risk mitigation in construction: Exploratory study. In *Advances in informatics and computing in civil and construction engineering* (pp. 823-829). Springer, Cham.
- Kefalidou, G., Golightly, D., & Sharples, S. (2018). Identifying rail asset maintenance processes: a human-centric and sensemaking approach. *Cognition, Technology & Work*, 20(1), 73-92.
- Lyytinen, K., Nickerson, J. V., & King, J. L. (2021). Metahuman systems= humans+ machines that learn. *Journal of Information Technology*, 36(4), 427-445.
- Nardo, M., Forino, D., & Murino, T. (2020). The evolution of man-machine interaction: The role of human in Industry 4.0 paradigm. *Production & manufacturing research*, 8(1), 2034.
- Pedersen, I., & Mirrlees, T. (2017). Exoskeletons, transhumanism, and culture: performing superhuman feats. *IEEE Technology and Society Magazine*, 36(1), 37-45.
- Schneider, E., Irastorza, X., Bakhuys Roozeboom, M. M. C., & Houtman, I. L. D. (2010). Osh in figures: occupational safety and health in the transport sector-an overview.
- Sherwani, F., Asad, M. M., & Ibrahim, B. S. K. K. (2020, March). Collaborative robots and industrial revolution 4.0 (ir 4.0). In *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)* (pp. 1-5). IEEE.
- Stephenson, W. (1935), "Correlating Persons instead of Tests", *Character and Personality*, vol. 4, n°1, p. 17-24.
- Stephenson, W. (1953), *The study of behavior: Q-technique and its methodology*, Chicago, University of Chicago Press.
- Stirling, A., & Coburn, J. (2018). From CBA to precautionary appraisal: practical responses to intractable problems. *Hastings Center Report*, 48, S78-S87.
- Venkatesh, V., Thong, J. Y., & Xu, X. (2012). Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. *MIS quarterly*, 157-178.
- Vrousalis, N. (2013). Exploitation, vulnerability, and social domination. *Philosophy & Public Affairs*, 41(2), 131-157.

What does privacy mean to users of voice assistants in their homes?

Edith Maier¹, Michael Doerk², Michelle Muri², Ulrich Reimer¹ and Uwe Riss¹

¹ Eastern Switzerland University of Applied Sciences, St. Gallen, Switzerland ²
 Lucerne University of Applied Sciences, Lucerne, Switzerland
 edith.maier@ost.ch

Abstract. Voice assistants (VAs) have been spreading rapidly and encroaching into our private space, which raises serious privacy concerns. The paper presents the findings from an ongoing research project which examines the impact VAs have on everyday practices, but also investigates people's attitudes and actual and intended behaviour. We present the results from two groups: 1) people who volunteered for a four-week self-monitoring study in their homes, 2) students who engage with VA technology as part of their course work.

In the data analysis privacy emerged as a key issue, with many responses pointing to the privacy paradox, i.e. the contradiction between generic attitudes to privacy and actual behaviour. Participants of both the in-home study and the student group agree that privacy is important and should be protected, the latter group even more so. Still, most are pragmatists who weigh the potential pros and cons of sharing personal information. Factors such as type of data, setting, location and the purpose of data use also influence people's risk assessments. Few take steps to self-manage privacy e.g. by adjusting settings or deactivating cookies. Instead most people appear resigned to (potential) violations to their privacy. We therefore conclude that privacy-protective measures have to be taken at the regulatory and political levels. Although the GDPR has largely implemented the principles of privacy by design and default, users, however, are still confronted with incomprehensible privacy policies full of legalese and often misleading cookie choices.

Keywords: voice assistants, privacy, privacy paradox, data protection

1 Introduction

No other technology has been spreading as rapidly in the home as voice assistants (VAs) such as Siri or Alexa. They have been encroaching into our private space by becoming accessible via smart speakers such as Amazon Echo, Apple HomePod and Google Home (see e.g. Ammari et al., 2019). They promise to make our daily lives

more efficient by letting us control smart home devices, search for information or listen to our favourite music without having to get up from the couch.

In a 4-year project (2020–2023) funded by the Swiss Science Foundation, a team of researchers with expertise in human-centred design and interaction, home automation, digital services, data science, and behavioural economics has been investigating how the presence of VAs affects people’s domestic lives, norms and values. The study is expected to generate novel insights into the emerging issues associated with VA use in Switzerland and beyond.

Among the main goals of the project are:

- to examine the impact VAs have on everyday practices and routines as well as on the interaction between the members of a household
- to determine users’ concerns about privacy and data security associated with the use of VAs and find out how they deal with them
- to propose innovative design responses to the challenges identified and to increase general awareness about the social, societal and ethical implications of VAs.

The focus of this paper is on attitudes to privacy and intended as well as actual behaviour in privacy-sensitive contexts. With their VAs users can perform various tasks, control other devices and enjoy third party services. VAs therefore gather enormous amounts of personal information which is why they evoke serious privacy concerns regarding the collection, use and storage of personal user data. The results so far also highlight the contradictions between generic attitudes to privacy and actual behaviour. This phenomenon is often referred to as the ‘privacy paradox’.

The paper discusses the findings from two different Swiss study groups. Participants in the first group (Group1) volunteered to take part in a four-week self-monitoring study in their homes. The second study group (Groups2a & b) consists of students who, rather than volunteering, engage with VA technology as part of their study requirements for course work. Whilst Group2a has already completed the course work, Group2b is still in progress. In Section 2 we describe our methodological approach and how we adapted it as a result of preliminary findings. In Section 3 we discuss related work on the privacy paradox before presenting the findings of our own study in Section 4. This is followed by a discussion of their implications (Section 5) and finally, in Section 6 we wrap up with some conclusions and an outlook on future research.

2 Methodological Approach

To achieve these goals, we have gathered empirical data from two different study groups. The 31 participants in Group1 volunteered to take part in a four-week selfobservational study in their homes. They were reimbursed for the purchase of a basic VA device of their choice if they did not already have a VA, had to install it independently and report over four weeks how they wanted to and were able to use the device. They used a diary app to record their experiences. In this group, participants solved tasks sent to them by the research team on a weekly basis through

the remote app. Besides, further interviews were conducted 3 to 4 months after the in-home study so as to capture the longer-term impacts of VA presence. Because of Covid-19 all interviews had to be carried out via Zoom or MS Teams. Participants in the in-home study were selected to include a wide range of lifestyles, social and living situations. Participants ranged in age from 17 to over 70 and included technical novices as well as technical geeks.

In contrast to Group1, Group2 can be described as relatively homogeneous: students mostly in their early twenties, living on their own or with fellow students and they were all digital natives. Rather than volunteering, they have to engage with VA technology as part of their study requirements for course work over a period of 13 weeks. Their observations and responses are gathered via the web-based training platform (WBT) «relax-concentrate-create» (rcc platform) which is accessible to students and staff in Switzerland. Group2a with 33 participants could choose to explore VAs from one of three perspectives:

- already using a VA
- curious to find out more about VA
- not interested in or opposed to using a VA.

They completed the course in January 2022. Group2b with 62 participants started their course in February 2022.

In the VA-PEPR project, the rcc platform is used to collect students' attitudes as well as practices related to the use of voice assistants in their private lives. The aim is to examine whether their resource management could be improved by using voice assistants and to record their self-reflections and comments. They also answer questions about data protection and privacy. The data thus generated has already produced important insights into students' perceptions of what is private and how these have changed as a result of their own research into and use of voice-controlled gadgets. Therefore the data collected via the rcc platform complements and enriches the findings gathered via other methods such as interviews or data collected via ethnographic apps.

The ethnographic material from the in-home study, the text entries from the student journals as well as the transcriptions of the audio files from the semi-structured interviews have been analysed in a uniform procedure using qualitative coding. To guide the analysis at the beginning we defined categories in line with the research questions such as:

- What are people's motives/expectations with regard to installing/using a VA in their homes?
- How do people use the VA in their homes? For which practices/purposes?
- What does privacy mean to VA users in their homes?
- What kinds of changes in attitude towards privacy occur?

Apart from privacy, high-level categories included concepts such as Practices, Perceptions, Exploration & Coping Strategies, Technological Affinity, Desiderata. These categories were used as codes in the MAXQDA software. Researchers from different disciplines then came together to compare the allocation of text fragments to codes and emergent themes with the aim to arrive at a common understanding.

In the course of data analysis privacy and data protection emerged as key issues, which is why we decided to explore the topic in more depth. The analysis also showed that generic attitudes to privacy were poor predictors of context-specific behaviours.

Many responses, esp. to questions asked in the interviews, pointed to the so-called ‘privacy paradox’, i.e. the phenomenon where people say that they value privacy highly, yet in their behaviour relinquish their personal data for very little in exchange or fail to use measures to protect their privacy (see e.g. Solove, 2021; Winegar & Sunstein, 2019; Barth & De Jong, 2017).

As a result of these findings we adapted the course requirements related to the rcc platform. In the questionnaire at the beginning of term, i.e. February 2022 (Group2b, 62 students), we included questions to elicit answers that should cast more light on the attitude-behaviour dichotomy. For the questionnaire to be filled in at the end of term we designed privacy-sensitive scenarios to explore individuals’ perceptions of privacy in specific contexts or situations. So far only the results of the initial questionnaire are available.

3 Related Work on the “Privacy Paradox”

The Privacy Segmentation Index introduced by Westin (2003) has been widely used to measure privacy attitudes and make longitudinal comparisons (Krane, Light & Gravitsch, 2002; Kumaraguru & Cranor, 2005). Westin’s Index categorizes individuals into three privacy groups:

1. *Unconcerned* – those who give privacy little thought
2. *Pragmatists* – those who worry about threats to privacy but believe that reasonable safeguards are in place or can be created
3. *Fundamentalists* – those with high privacy concerns and high distrust in government, business, and technology

One might expect that people’s attitudes would have a significant impact on their behaviour. Previous research, however, has failed to establish a robust correlation between the Westin categories and actual or intended behaviour. Numerous studies have documented an attitude-behaviour dichotomy (also referred to as the Privacy Paradox), in which participants’ privacy-related attitudes do not coincide with their behaviour (see e.g. Barth et al., 2019; Rittenberg & Tregarthen, 2013).

Solove (2021) distinguishes between two types of arguments about the privacy paradox: the “behaviour valuation argument” and the “behaviour distortion argument”. The first contends behaviour is the best metric to evaluate how people actually value privacy. Since people are willing to disclose personal data for ‘free’ goods or services, it is concluded that they ascribe a low value to privacy. The latter argues that people’s behaviour is not an accurate metric of preferences because behaviour is distorted by biases and heuristics, manipulation and framing (see e.g. Barth & de Jong, 2017; Weinberger, Bouhnik & Zhitomirsky-Geffet, 2017; Acquisti & Grossklags, 2007).

Other reasons given for the dichotomy include: instruments such as the Westin Privacy Segmentation Index measure general attitudes, while behaviour is contextspecific; individuals may perform privacy risk assessments but choose the most viable or convenient options, even if they are not in accordance with their privacy preferences (for an overview, see Gerber, Gerber & Volkamer, 2018).

Woodruff et al. (2014) explored the connection between the Westin categories and individuals' responses to the consequences of privacy behaviour. For this purpose they conducted a survey of 884 Amazon Mechanical Turk participants to investigate the relationship between the Westin Privacy Segmentation Index and attitudes and behavioural intentions for both privacy-sensitive scenarios and privacy-sensitive consequences. The results showed a lack of correlation between any of the categories.

Solove (2021) argues that the privacy paradox is a myth created by faulty logic. First of all, he points out that the privacy paradox is often based on misunderstandings, e.g. by equating it with secrecy.

“When people want privacy, they don’t want to hide away their information from everyone; instead, they want to share it selectively and make sure that it isn’t used in harmful ways. Privacy isn’t all-or-nothing – it’s about modulating boundaries and controlling data flow.” (Solove, 2021, 23)

The value of privacy must therefore not be conflated with individual valuations. His argument is supported by behavioural economists such as Winegar and Sunstein (2019) who consider it extremely difficult to place any kind of monetary value on data privacy given the divergence between statements of value and actual behaviour.

Whilst it makes sense to assess the value of a product by asking people how much they would pay for it, privacy has a value beyond individual assessments. According to Solove (2021) privacy is valuable for the following reasons:

- It puts limitations on power.
- It demonstrates respect for the wishes of individuals.
- It enables people to manage their reputations.
- It helps maintain appropriate social boundaries.
- It is a prerequisite for establishing and maintaining trust and candour in relationships.
- It is essential for having control over one’s life.

Moreover, Solove considers privacy to be essential for freedom of thought and speech as well as social and political activities. It also allows people to change and have second chances, protects their intimacy and it matters because it means not having to explain and justify oneself (idem, 32-33). He views privacy as a constituent element of a free and democratic society, an opinion which is shared by others such as Schwartz (1999) or Véliz (2020).

The fact that people share personal data does not mean that they do not care about privacy. Actually, we live in an age where it is nearly impossible not to disclose personal data if one wants to participate in social life and engage in economic activities. Actually, people constantly make risk assessments when they trade their personal data in exchange for gaining access to information or services important to

them. In a comprehensive empirical study by Kesan et al. (2015) on data privacy, trust and consumer autonomy, more than 80% of respondents said that they had at least once provided information online when they wished that they did not have to do so. Quite often we are not afforded much choice or rather a take-it-or-leave-it choice, which is connected with the business model that dominates the Internet: Provide free online content and monetize it collecting, using or selling personal data.

Such a state of affairs often triggers a feeling of helplessness and powerlessness on the part of users. Participants in our study have echoed these sentiments. In the following section we look at the findings in more detail.

4 Results

In total, 391 quotations have been assigned to the code "Privacy". These have been gathered both from the in-home studies and the text entries from students of different disciplines at a Swiss University of Applied Sciences and Arts on the rcc platform. The quotes have been translated by the authors from the original German and assigned to Group1 or Group2.

4.1 Attitudes to Privacy

Among the participants, we can distinguish between three different attitudes regarding privacy, namely:

1. *Unconcerned* about privacy and/or unrestricted use of VA
2. *Privacy concerns*, but willing to engage with VA if they see an added value in its use
3. *Critical*, very restricted use of VA

Table 1: Frequency of use correlated with importance attributed to privacy when using VA

How important do you consider privacy and data protection when using a voice assistant (VA)?		<i>Unconcerned</i> ("not important", "I don't really care where data is stored or processed provided the VA works")	<i>Privacy concerns</i> ("I do care but when the benefits outweigh the risks, I use the VA")	<i>Critical</i> ("I consider privacy very important therefore I deliberately restrict the use of the VA")	Total
How often do you use the VA?	Several times a day	5	28	5	38
	Several times a week	1	6	0	7
	Rarely	4	6	7	17
Total		10	40	12	62

This finding very much coincides with the categories identified by Westin (2003). As explained in Section 2 we decided to conduct a more systematic study of the attitudebehaviour dichotomy by introducing relevant questions in the course work related to the rcc platform for Group2b.

As can be seen in Table 1, the majority, i.e. 40 out of 62 respondents, belong to the pragmatists who care about privacy but are ready to make trade-offs if the benefits are big enough. Ten do not consider privacy very important and just want the VA to function properly, whereas 12 harbour serious concerns with regard to their privacy.

In the following we present the results in more detail and illustrate them with quotes from both Group1 and Group2a.

Unconcerned

Participants do not care or do not mind when VA collects personal data because they feel they have nothing to hide.

"What can they do with all that information? Most of the things I say are not that interesting." (Group1)

"Other devices listen in too, why should it make any difference with the VA. Personal data are collected by one's smartphone and other devices anyway." (Group1)

“I’d rather Siri listen to me than my neighbour.” (Group1)

Some participants trust particular VA manufacturers, especially Apple. They believe that Apple respects data protection and only transmits data to third parties if users opt in or explicitly allow this to happen. But even with Apple, so they admit, there’s no 100% guarantee that personal data are not misused. A few participants think that VAs from other manufacturers such as Google do not transmit or misuse data either. For example, one of the participants reported the following interaction:

“‘What data do you send to Google?’ I asked my Google Assistant, to which it replied ‘Only things that are directly addressed to me and nothing else.’ That reassured me.” (Group1)

A few participants said that they trusted the Swiss government to prevent any misuse of their data. However, if they lived in an authoritarian country such as the former GDR or in China, they would hesitate to use a VA.

Pragmatists

Pragmatists refer to those who have privacy concerns, but still adopt and use a VA even though they might have

- a sense of discomfort in knowing that the VA is ‘always listening’,
- concerns about the companies’ use and/or storage of personal data,
- a feeling of uneasiness about personal profiling and personalised targeting

Some participants found it quite ‘spooky’ and ‘scary’ when they looked at the voice logs and saw all the verbatim recordings of their conversations. On the other hand, they were also aware of the fact that other devices, the smartphone in particular, was recording their locations, conversations and other activities as well. However, they would not want to do without it.

*“...one is quite shocked when one looks at the voice recordings:
It makes you ask: ‘That’s really how I talk?’”(Group1)*

“In the digital world, we are all open books anyway.” (Group1)

Therefore even those who are concerned about their privacy, and they do account for the majority, lack the time or do not care enough to deal with the issue in-depth or look at the potential implications. Or they feel that the benefits outweigh the risks and/or do not want to opt out of the digital world altogether.

Fundamentalists

Representatives of the last group can be found mostly in the student group, which may be due to the fact that they had not volunteered but were free to decline the use of a VA.

Quite a few said they did not want a device in the home that constantly listened and overheard everything they said. According to them the potential benefits did not

warrant the disclosure of personal data. Some would be willing to use a VA, provided data protection were guaranteed and their privacy concerns allayed.

Further reasons given included that they did not trust the tech companies to protect their data. Others were opposed to the use of such devices in general because of

- fear of surveillance and becoming a completely transparent citizen,
- fear of cyber attacks,
- fear of public exposure or threat to one's reputation.

Some explicitly stated that they were opposed to the dominant business model on the Internet, i.e. 'free' service in exchange for personal data.

"It is not transparent which data are collected and how they are used. It is unclear where the data goes and how it is processed. Are the data encrypted or can they be associated with me?" (Group2b)

"Even when it is [highly unlikely] that some embarrassing personal details get out, it may ruin your reputation." (Group1)

"According to [an expert] it is possible to activate VAs by means of ultrasound signals which are imperceptible by humans...This makes it possible to control the devices remotely." (Group2a)

4.2 Mediating factors

Type of data

Most users distinguish between data that are perceived as sensitive/confidential vs. nonconfidential. Generally, financial data (e.g. account or mortgage information), data related to health/disease or intimate relations tend to be considered private. This confirms that perceived risks and protection requirements differ by data type as has been shown by Fukuta, Murata & Orito (2020).

"There are pages where you can check if a person is creditworthy. If information about your private finances gets out, all of a sudden you might not get a loan, or a grant." (Group1)

Purpose of use

When assessing the risks of disclosing personal data, many participants also consider the purposes for which the data is used. Whilst using the VA for listening to music, switching lights on and off or searching for public transport connections is considered largely innocuous and low-risk, using it for placing orders in online shops, transmitting health data or making financial inquiries is seen as more of a risk.

As for purposes, most interviewees are not opposed to having their (anonymized) recordings used for improving the service quality or speech recognition of a VA but would be against transferring them to third parties. The majority of users also do not

welcome personalised recommendations or advertisements, especially from third parties, that try to influence their purchases.

“In the end the device and thus the system or manufacturer behind it might know more about oneself than me.” (Group2a)

Still, there were a few participants who felt that if the VA were more familiar with their personal preferences or everyday needs, it might be helpful if the VA e.g. could remind them to take medication, switch on the heating before they get home or made them welcome with their favourite music.

Contextual factors

Privacy concerns are furthermore mediated by contextual factors such as setting and location. For example, most participants explicitly stated that they would not want to have the VA in their bedroom, the children's room or in the bathroom. However, a few participants who use the VA for controlling smart home devices do have a VA installed in the bathroom (e.g. for listening to music) and/or bedroom (e.g. for use as an alarm clock).

When using VA on one's smartphone, there is concern about communicating with a VA in public, e.g. when on a train or on a bus since it could reveal confidential information by accident. Some switch off the VA when they are expecting visitors because they are not sure if a VA can distinguish between different voices. However, in the course of time, some users tend to forget to switch it off and/or no longer care.

Quite a few exercise some form of self-censorship because they do not trust the VA not to listen in on conversations even when they have not used the appropriate wakeword.

“It's like having a spy in my home.” (Group2b)

“The presence of a device that is constantly listening bothers me deeply.” (Group1)

Risk assessments and trade-offs

In the second round of interviews with participants of the in-home studies we tried to determine what levels of risk were considered acceptable in specific contexts and for specific activities. Most respondents would find it acceptable if a company used their personal data to improve their services (e.g. in terms of speech recognition) or to develop new products. Some would be in favour for trading their smart home data for a reduction of insurance premiums. Most users, however, were against personalised ads that may influence their purchase decisions.

“I would feel ‘cheated’ because I wouldn't get the whole range of what's on offer.” (Group1)

This shows that people tend to make trade-offs between privacy risks and benefits associated with the VA. For example, they may be willing to hand over their data in

exchange for a useful and reliable service or for convenience (e. g. hands-free control when driving) or to be able to use all functions of a device or system. Whether someone continues to use a VA is also related to the actual or expected effort or time one might have to spend in familiarising oneself with the device. Many felt that – at least for the time being – the VAs were not providing enough added value and did not warrant any (more) investment of their time.

Although participants in both groups show a diffuse awareness of risks, hardly anyone bothers to read the privacy policies which, after all, might help clarify some of their concerns. The Terms and Conditions that come with the purchase of a VA are considered too complex, cumbersome, too time-consuming to read or incomprehensible.

Privacy-protective measures and desiderata

Few users take steps to self-manage privacy e.g. by adjusting settings or deactivating cookies. Instead, most people appear resigned to or cynical about (potential) violations to their privacy. The coping strategies mentioned to prevent or mitigate violations of one's privacy include:

- Adjust the default settings related to privacy so as to control which data are stored and/or transmitted.
- Change the wake word so the VA cannot be activated by others.
- Switch off the microphone when VA is not needed or when visitors are expected.
- Adjust the setting in a way that the VA has to be activated manually first (especially on the smartphone).

Apart from these individual strategies users expressed a wish for VAs without internet connection, i.e. a local system, so that their data could not be sent to the manufacturers, or an open-source VA independent of the big tech companies. Also, many would like to have a simple opt-in/opt-out option with regard to data transfer. Furthermore, most would welcome clear rules or regulations – including sanctions if these are violated –, which define what is allowed, which data have to be treated confidentially, what can be passed on, and what they can be used for.

5 Discussion

As with the Internet, users of VAs often tend to imagine that they are protected by hightech companies or data protection laws. It is only when newspaper articles or headlines report leakages of personal information, unintended purchases or VAs actively listening in to or recording private conversations that people realize that much goes on 'behind the scene' when using smart speakers. Trust or lack of trust in technology and the service provider plays an important role whether someone decides to install and use a VA. Furthermore, the findings also reveal the moderating role of experience, need for human interaction, and the perceived private nature of both data and setting.

Overall, most people care about privacy as shown in the responses to the adapted questionnaire of the rcc module (see Table 1). Only ten respondents are unconcerned about data protection, whereas 52 out of 62 respondents are concerned to different degrees. Most are pragmatists who weigh the potential pros and cons of sharing information, evaluate the protection in place as well as their trust in the company or organisation and then decide whether they are prepared to divulge personal information.

Participants of both the in-home study and the student group agree that privacy is important and should be protected. Students seem to be even more concerned about protecting their privacy than participants of the in-home studies. Moreover, they had access to articles or videos which focused on the potential threats to people's privacy and thus made them aware of the potential harms that breaches of privacy and security might entail.

Some users are quite aware of the contradictions between their privacy-related attitudes and their actual behaviour in other realms of their lives. For example, by accepting loyalty cards from shops they are willing to disclose the content of their shopping baskets in return for discounts. And many do not adjust cookie settings when they want to use a particular web-based service, or because they are in a hurry and cannot be bothered.

Users do not only distinguish between different types of data when it comes to assessing risk, they also make distinctions with regard to the purpose to which their data or actions would be used. For example, they would be prepared to disclose personal data for the purpose of improving voice recognition or service quality of the VA, in return for lower insurance premiums. Whilst the majority would be agreeable to having their data used for improving service quality – provided they were anonymized – very few would disclose personal information in return for personalized online advertisements.

Still, no matter how much people know about or are aware of the potential harms associated with privacy violations, hardly anyone decides to opt out of the information society altogether. This may apply to young people, in particular, as shown in a study by Hargittai and Marwick (2016) on social media use. Even the fundamentalists in Group2b do use their VAs despite their serious concerns (see Table 1).

According to Solove (2021), privacy self-management just does not scale. When each individual choice or action is viewed in isolation, he argues, the privacy-protective steps or measures appear simple and not too onerous. When people fail to take these small steps, one must not conclude that people do not care about privacy but look at the larger context. Even those who are quite aware of potential threats to their privacy and know how to reduce them tend to be resigned to the limited control over their data. Quite a few participants therefore express feelings of resignation, apathy or even cynicism because they believe that privacy violations are inevitable.

6 Conclusions and Outlook

People are very diverse in how they feel about risk. Whilst some feel that we should ask consent for all uses of personal data, others feel that consent is important only when there is a risk of harm. The latter would argue that in today's world we cannot avoid using or being exposed to digital technologies and therefore have to resign ourselves to being observed as happens anyway given the omnipresence of video cameras.

In the course of our project we have come to realize that privacy self-management is not a solution to the problem. At the beginning we assumed that we could encourage people to engage in more privacy-protective behaviour by increasing their knowledge and know-how. However, as other researchers before we have come to conclude that one cannot protect one's privacy without radically disconnecting from the modern world.

The matter is further complicated by the so-called "aggregation effect". It involves understanding how personal data can be analysed when combined into an extensive digital dossier about a person. People give out bits of data here and there, and each individual disclosure to one particular entity might be relatively innocuous. Modern data analytics works via algorithms examining patterns in large quantities of personal data. It is nearly impossible for people to understand the full implications of providing certain pieces of personal information to certain entities. When combined, these can reveal facts that people might not want to share.

Therefore privacy-protective measures have to be taken at the regulatory and political levels. Privacy regulation often seeks to give people more privacy selfmanagement. Instead, Solove (2021) argues, regulation should employ a different strategy, namely focus on regulating the architecture that structures the way information is used, maintained, and transferred. This may include outlawing certain types of personal data transfers or making them more difficult. Regulations, for example, can control downstream transfers and uses. Privacy regulation could also address the design of products or services and ensure for effective data security and restrict design that is insecure or creates unwarranted privacy risks.

To a large extent, the General Data Protection Regulation (GDPR) fulfils Solove's recommendations by implementing the principles of privacy by design and privacy by default (see esp. Article 25, www.privacy-regulation.eu/en/article-25-data-protectionby-design-and-by-default-GDPR.htm). The GDPR became applicable in May 2018 and offers standardised contractual clauses, providing an easy-to-implement tool to comply with data protection requirements.

But most Internet users are still confronted with privacy policies full of legal jargon and often misleading cookie choices. Perhaps it is more a question of enforcing the GDPR on a larger scale and in a more user-friendly manner, and punishing violations of the right to privacy rather than or in addition to raising people's awareness and knowledge of privacy-protective measures. With our continuing research we expect to shed more light on these issues and suggest ways of achieving a digital world in which the values of privacy and data protection are embedded in the technology.

Acknowledgements

This study was funded by Swiss National Science Foundation (SNF) project VA-PEPR, ref. no. CRSII5_189955.

References

- Acquisti, A., & Grossklags, J. (2007). What can behavioral economics teach us about privacy. *Digital privacy: theory, technologies and practices*, 18, 363-377.
- Ammari, T., Kaye, J., Tsai, J. Y., & Bentley, F. (2019). Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Trans. Comput. Hum. Interact.*, 26(3), 17-1.
- Barth, S., de Jong, M. D., Junger, M., Hartel, P. H., & Roppelt, J. C. (2019). Putting the privacy paradox to the test: Online privacy and security behaviors among users with technical knowledge, privacy awareness, and financial resources. *Telematics and informatics*, 41, 55-69.
- Barth, S., & De Jong, M. D. (2017). The privacy paradox—Investigating discrepancies between expressed privacy concerns and actual online behaviour—A systematic literature review. *Telematics and informatics*, 34(7), 1038-105.
- Fukuta, Y., Murata, K., & Orito, Y. (2020). Perceived Risk and Desired Protection: Towards A Comprehensive Understanding of Data Sensitivity. In *Societal Challenges in the Smart Society* (pp. 573-586). Universidad de La Rioja.
- Gerber, N., Gerber, P., & Volkamer, M. (2018). Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. *Computers & security*, 77, 226-261.
- Hargittai, E., & Marwick, A. (2016). “What can I really do?” Explaining the privacy paradox with online apathy. *International journal of communication*, 10, 21.
- Kesan, J. P., Hayes, C. M., & Bashir, M. N. (2015). A comprehensive empirical study of data privacy, trust, and consumer autonomy. *Ind. LJ*, 91, 267.
- Krane, D., Light, L., & Gravitch, D. (2002). Privacy on and off the Internet: What consumers want. *Harris Interactive*, 10003, 15229.
- Kumaraguru, P., & Cranor, L. F. (2005). *Privacy indexes: a survey of Westin's studies* (pp. 368394). Carnegie Mellon University, School of Computer Science, Institute for Software Research International.
- Rittenberg, L., & Tregarthen, T. D. (2013). *Principles of Microeconomics: Version 2.0*. Flat World Knowledge.
- Schwartz, P. M. (1999). Privacy and democracy in cyberspace. *Vand. L. Rev.*, 52, 1607.
- Solove, D. J. (2021). The myth of the privacy paradox. *Geo. Wash. L. Rev.*, 89, 1.
- Turow, J., Feldman, L., & Meltzer, K. (2005). Open to exploitation: America's shoppers online and offline. *Departmental Papers (ASC)*, 35.
- Véliz, C. (2021). *Privacy is power*. Melville House.

- Weinberger, M., Bouhnik, D., & Zhitomirsky-Geffet, M. (2017). Factors affecting students' privacy paradox and privacy protection behavior. *Open Information Science*, 1(1), 320.
- Westin, A. F. (2003). Social and political dimensions of privacy. *Journal of social issues*, 59(2), 431-453.
- Winegar, A. G., & Sunstein, C. R. (2019). How much is data privacy worth? a preliminary investigation. *Journal of Consumer Policy*, 42(3), 425-440.
- Woodruff, A., Pihur, V., Consolvo, S., Brandimarte, L., & Acquisti, A. (2014). Would a Privacy Fundamentalist Sell Their DNA for \$1000... If Nothing Bad Happened as a Result? The Westin Categories, Behavioral Intentions, and Consequences. In 10th Symposium on Usable Privacy and Security (SOUPS 2014) (pp. 1-18).

From Liberal Democracies to Blockchain Systems: The Instrumentalization of Decentralization

Sofia Cossar¹, Wessel Reijers² and Primavera De Filippi¹

¹ CNRS

² University of Vienna, Technion, EUI
BlockchainGov Project, Paris, France
sofia.cossar@gmail.com

Abstract. Blockchain technologies have been recognized as a new force in the regulation and governance of human societies. From their early origins in cypherpunk culture and libertarian communities of cryptographers and software engineers, blockchains have appealed to the ideal of ‘decentralization’, both as a technical ideal (i.e., as an alternative to ‘centralized’ communication networks) and as a political ideal (i.e., as an alternative to the ‘centralized’ management of current financial, legal, or political systems). While the focus often lies with technical decentralization, its political cousin is also frequently appealed to under the guise of ‘democratization’ of finance, trade, and even democracy itself. Behind this claim lies the assumption of ‘democratization’ as a greater decentralization of power to participate in decision-making. Yet, the extent to which blockchain systems promote greater ‘decentralization’ and ‘democratization’ remains contested. This paper draws from normative critiques of blockchain technology and the instrumentalization of decentralization within the broader debate in political philosophy on democracy. Extrapolating the analysis of political liberalism made by David Dyzenhaus and other critical theory scholars, we argue that blockchain systems experience a similar tension to liberal democracies between decentralization, the protection of value systems, and the reproduction of power structures. These tensions, which bring out a similar threat of instrumentalizing decentralization, lead to a diminishing of legitimacy of both decision-making systems.

Keywords: blockchain systems, liberal democracies, decentralization, instrumentalization

1 Introduction

Blockchain technologies have been recognized as a new force in the regulation and governance of human societies. In the wake of the argument that ‘code is law’, blockchain brings new possibilities that extend beyond the wielding of software code by human actors to regulate the behavior of others, to the idea of a ‘rule of code’, under which systems display a sense of regulatory autonomy beyond the human control (De Filippi & Wright, 2018). From their early origins in cypherpunk culture and libertarian

communities of cryptographers and software engineers, blockchains have appealed to the ideal of ‘decentralization’, both as a technical ideal (i.e., as an alternative to ‘centralized’ communication networks) and as a political ideal (i.e., as an alternative to the ‘centralized’ management of current financial, legal, or political systems) (Nakamoto, 2008; Magnuson, 2020). While at times the focus lies with technical decentralization, its political cousin is also frequently appealed to under the guise of ‘democratization’ of finance, trade, and even democracy itself (Atzori, 2015). We argue that, behind this claim, lies the assumption of ‘democratization’ as a greater decentralization of power to participate in decision-making. Indeed, from a system theory in political science perspective (Easton, 1965), we can equate blockchain systems and liberal democracies to delimited yet fluid collectives of interrelated parts with a structure for collective decision-making. Yet, the extent to which blockchain systems actually promote greater ‘decentralization’ and ‘democratization’ remains contested. At the core of the discussion, is the conceptualization of decentralization and the role that decentralization plays within liberal democracies and blockchains as decision-making systems.

Law, governance, and technology scholars have studied different aspects of the normative qualities of blockchain systems as regulatory and governance technologies. These include discussions concerning the extent to which blockchains are conducive to and the support of rule of law principles that (ideally) are fundamental to liberal democracies and global governance. The focus of these discussions lies in the law-like features of blockchain systems and the extent to which these may protect, promote or undermine individual rights in liberal democracies (Reijers et. al, 2018; Yeung, 2019; De Filippi & Hassan, 2018; De Filippi & Wright, 2018) as well as the its potential provide incentives and lead to better collective action, reinvigorating liberal principles in international cooperation and global governance (De Filippi, 2021; Reinsberg, 2021). Additionally, scholars have discussed the ideological origins and proclivities of blockchain communities (Brunton, 2019; Swartz, 2017) and ‘decentralization’ as a (contended) feature of blockchain systems (Srinivasan & Lee, 2017; Sai et al., 2021; Lemieux & Feng, 2021). Yet, the ‘democratizing’ aspect of blockchain technologies, i.e., the extent to which technical and political decentralization overlap, and how this relates to liberal democratic principles, has been largely neglected. This paper aims to address this gap in the literature, by drawing normative critiques of blockchain technology and the instrumentalization of decentralization within the broader debate in political philosophy on democracy. To do so, it primarily draws from the work of David Dyzenhaus, who has looked at a fundamental tension in liberal democracies between protecting core liberal values and promoting democratic decision-making. We also highlight critical theories’ approaches to liberal democracies and their argument on how legal and political systems reproduce existing and oppressive power structures. These views, we argue, are not mutually exclusive but share a critical view on the instrumentalization of decentralization in liberal democracy as a decision-making system. We contend that, contrary to claims of the ‘democratizing’ power of blockchain technology, blockchain systems reflect the same tensions between political decentralization (i.e., decentralization of ‘power’ to participate in decision-making), the

promotion of certain 'technological' values, and the accumulation of power by oppressive groups.

Our argument revolves around the comparison between decentralization in blockchain technologies and in liberal democracies, which urges us to have a notion of decentralization that could be applied to both contexts. Decentralization refers to either an existing or desirable feature of a system. It may be static (unchangeable, fixed) or dynamic (with the capacity to evolve). It can be a binary property — either 'decentralized' versus 'centralized' — or non-binary, with many degrees of decentralization (Fesler, 1965). Substantively, decentralization as a dynamic non-binary feature entails disseminating or transferring power from 'a few' to 'the many' (Schneider, 2003). Few and many refer to actors operating within the boundaries of the decision-making system or directly affected by its decisions or outputs. 'Power', as linked to decentralization, is seen as 'power to' (do *x*) or the *effective capacity* to participate in these decision-making systems (Avelino, 2021). Through 'power to', as opposed to 'power over,' decentralization and participation have a symbiotic relationship in which they are both a precondition and a result of each other (Litvack & Seddon, 2002). Participation, however, is the *effective actualization* of the disseminated 'power to' — it is 'power to' in action within the decision-making system (Dedding et al., 2020).

In what follows, we argue that (1) liberal democracies experience a fundamental tension between political decentralization, the protection of liberal principles, and the reproduction of power structures; (2) blockchain systems experience a similar tension between political decentralization, the protection of 'technological' principles, and the reproduction of power structures; (3) both these tensions bring out a similar threat of instrumentalizing decentralization, which could lead to a diminishing of legitimacy. In section two, we follow Dyzenhaus' argument about the tension in political liberalism between protecting core liberal values and respecting the democratic process. We follow Dyzenhaus in arguing that political liberals cannot do otherwise than sacrifice democracy for the sake of protecting liberal values and that this presents a threat of sacrificing legitimacy. We also explore critical theories and arguments on how liberal democracies allow for the continuation of pre-existing inequalities and structures of oppression. Both approaches share an implicit assumption on the value of decentralization as an end in itself, instead of a means to an end. In section three, we explore how decentralization works in blockchain systems and illustrate how it sometimes generates tensions in the governance of these systems, focusing on the case of the Bitcoin 2013 hard fork debate. In section four, as a conclusion, we bring the tensions of these systems into contact, arguing that although blockchain systems and liberal democracies are different decision-making systems, they both face the threat of an instrumentalization of decentralization.

2 Liberal Democracies and Decentralization

Most analyses of the normative (ethical, political) impacts of blockchain technologies argue from the vantage point of liberal democracies, i.e., whether blockchain technologies promote, diminish or reconfigure principles that underlie liberal democratic polities, such as the rule of law and the protection of ‘fundamental rights and freedoms’ (Hughes, 2017; Chow, 2018). A multifaceted decision-making system, some regard liberal democracy as a complex blend of liberal and democratic principles tailored to context-specificity (Parekh, 1992). The ideological origins of liberalism are commonly traced to the Age of Enlightenment. The intellectual and philosophical movement that prevailed in Europe during the 17th and 18th centuries promoted the separation between church and state. It anchored itself on the idea of the individual, the prevailing of reason, and the promotion of liberties. The foundations of democracy, on the other hand, are usually connected to Ancient Athens, where adult, male ‘citizens’ directly participated in the Ecclesia, a legislative and executive body. The historical catalysts of Western modern liberal democracies were the American and French revolutions, with the adoption of the United States Constitution in 1787 and the Declaration of the Rights of Man and of the Citizen in 1789, as well as the British Bill of Rights in 1689 (Shils, 1995). It was, however, during the late 20th century, at the end of the Cold War and the dissolution of the Soviet Union, that liberal democracies acquired more prominence, accompanied by a feeling of ‘triumphalism’ that presented them as the only viable form of ‘human government’ against ‘the remnants of absolutism’ (Fukuyama, 1989).

There are various perspectives on what these principles ought to be. Still, minimally it entails promotion and protection of individuals’ liberties, both positive and negative (the ‘liberal’ in liberal democracy) and of co-decision, both by limiting the concentration of power as well as facilitating direct or representative democratic decision-making (the ‘democratic’ in liberal democracy). To regulate the tensions between liberty and democracy, liberal democracies recognize the difference between the public and private spheres. In the public sphere, law and policymaking are discussed and enacted by elected representatives and in the private sphere, individuals exercise their autonomy without the state’s or church’s intervention (Parekh, 1992).

Are liberal democracies decentralized? To what extent? Next, we highlight some of the most popularized arguments. Liberal democracies disseminate ‘power to’ define laws and policies from a unique authority — such as the monarch — to the citizens through their right to vote and be elected. They also disseminate ‘power to’ decide how to conduct private affairs from a unique authority — such as the church — to individuals through their right to autonomy (Hart, 1972; Kaim, 2021). Depending on the case, liberal democracies may also disseminate ‘power to’ participate in law and policymaking from central to local governments (Crook & Manor, 2000; Dardanelli & Wright, 2021) and from governments to civil societies and grassroots organizations (Durbin, 2001). In all these cases, decentralization is seen as both a precondition and a result of democratic participation.

However, although liberalism and democracy are intertwined in liberal democratic states and synthesized through a unified system of law, a fundamental tension between

the two remains. This tension is targeted by David Dyzenhaus in his *Legality and Legitimacy* (1999), in which he examines a debate in legal philosophy between Hans Kelsen, Carl Schmitt, and Hermann Heller in the context of constitutional challenges that eventually led to the overthrow of the Weimar Republic in 1933. The core of this debate revolves around the capacity of a positive legal system to withstand the pressures of groups holding comprehensive doctrines to overthrow its constitution. Is the validity of decisions made within such a system derived from a 'basic norm', placed as an Archimedean point outside the system, or from a decision made on the exception by a sovereign agency? It turns out that democratic constitutions contain vulnerabilities that liberal legal positivists have to guard against, at the ultimate cost of making the legal order immutable to influence by democratic processes. In other words, core liberal principles that constitute a positive legal order ought to be protected against the threats to these principles by democracy.

Although this debate occurred in the 1930s, Dyzenhaus argues that it is still relevant to today's debates in legal and political philosophy. He contends that liberal democracies in Europe and North America have been undergoing a similar 'crisis of theory and practice.' An increasing number of dissatisfied and oppressed populations could not and were not allowed to 'satisfy their wants' within these legal and political orders. As such, he considered it paramount to debate the legitimacy of the legality of liberal democracies by dissecting its founding ideology, political liberalism. As the main proponent of political liberalism, John Rawls formulated a theory of justice to explain how free and equal individuals could peacefully coexist in democratic societies comprised of a plurality of incompatible comprehensive doctrines or conceptions of the good, including religious, philosophical, and moral traditions. His answer was a notion of justice in line with liberal political principles. His previous work, *Theory of Justice* (1992), presented the principles of his 'theory of the right.' First, each individual should be afforded the same equal rights and liberties. Second, social and economic inequalities could exist so long as everyone has fair and equal opportunities for offices and positions these are to the greatest benefit to the least-advantaged members of society. To this idea of 'justice as fairness,' Rawls adds a 'political conception' so that institutions situated within the boundaries of liberal democracies are not only just but foster unity and stability. The principles of such political order are 'freestanding' or distanced from any comprehensive doctrine. Yet, they can still be endorsed by individuals subscribing to competing and 'reasonable' comprehensive doctrines. These political principles express an 'overlapping consensus' and a reasonable pluralism within liberal democracies (Rawls, 1993/2020)

Dyzenhaus' (1999) argument on 'democracy' and the 'law' being instrumental to political liberalism refers, in fact, to the instrumentalization of decentralization as the dissemination of 'power to' participate in law- and policymaking within liberal democracies. To Dyzenhaus, Rawls' formulation of political liberalism is problematic on two fronts. Firstly, borrowing from Carl Schmitt (1932/1996), the political principles maintained by the overlapping consensus are not neutral but partisan. They represent the values of privatized moralities, the values of the liberal autonomous individual. As such, while the first and second principles legally allow for pluralism — for example, through the right to freedom of religion — they *de facto* exclude comprehensive

doctrines that do not abide by the partisan political liberal principles, such as ‘autonomy’ or ‘dignity.’ ‘Unreasonable comprehensive doctrines’ are, in fact, the doctrines that challenge Rawls’ overlapping consensus; which itself is the basis of critique while also being put outside of the range of critique. In other words, the political liberal is the only one adopting the law, enforcing it, reaffirming it in judicial interpretation, and deciding on the states of exception. Indeed, Dyzenhaus argues, one could claim that liberal democracies ensure citizenship participation. However, all actors within the confine of the state are supposed to agree on basic rules and freedoms they haven’t necessarily chosen. If any piece of legislation seeks to enact any particular view of ‘the good’ — that is, in contravention of the liberal view of ‘the good’ —, it is ruled out through judicial interpretation as ‘unconstitutional.’ If the political liberal deems a situation as an ‘attack or threat to public reason’ it can resort to a ‘state of exception’ to circumvent the guaranteed rights and freedoms and assert the prominence of the partisan and controversial political liberal values. In short: decentralization as dissemination of the ‘power to’ participate is upheld so long it poses no threat to a certain value system (Dyzenhaus, 1999).

Dyzenhaus himself points out that other critical scholars have questioned the legitimacy of liberal democracies, including communitarians, feminists, and democrats (Dyzenhaus, 1999). However, the framing is slightly different. For example, the prominent feminist Susan Okin questions core principles of political liberalism such as the distinction between the so-called ‘private’ and ‘public’ spheres as blind to the oppressive division of labour in the household, and the overlapping consensus as being impartial towards ‘reasonable’ comprehensive doctrines, including religions, that, in fact, propagate sexist oppression against women. (Okin, 1979). Much like other critical approaches, liberalism is perceived by critical feminists as enacting and promoting laws and policies that, while claiming to be neutral and universally accepted by ‘reasonably individuals,’ reproduce oppressive power structures. In other words, it paves the way for the white, bourgeois men to continue accumulating ‘power over’ (Avelino, 2021) groups which are oppressed based on their gender, sexual orientation, race or religion (Gannon & Davies, 2007). While the scope of their critique is wider than Dyzenhaus’ (1999) these approaches also imply an instrumentalization of decentralization, where dissemination of ‘power to’ participate in decision-making is only maintained insofar it doesn’t subvert pre-existing power structures.

So far, we have highlighted that the instrumentalization of decentralization in liberal democracies can be argued through two macro-lenses: to promote a specific value system, or to increase the power of oppressive groups in society. However, what is the value of decentralization as an end in itself? Having analyzed this unavoidable crisis of legitimacy of liberal democracies, Dyzenhaus contends that the citizen should be “first democrat, and then a liberal” (Dyzenhaus, 1999; 254). Examining the arguments of German philosophers Hermann Heller and Jürgen Habermas, he makes a case for deliberative, contemporary democracy. He adopts from Habermas (1981/1984) the idea that democracy is not instrumental but the only institution that can support a culture of justification through debate. Following Heller (1934/1971), he contends that the legitimacy of the law is contingent and that its content is grounded in morality. The law ought to serve the self-government of citizens. Deliberation and participation in

decision-making is the process of collective inquiry on ever-changing ideals of social equality. It is the process through which the ethical principles of the law and institutions, which are context-specific, are collectively held into account by the citizens (Dyzenhaus, 1999). What this means is that liberal democracies ought to be robust, but that democracy and decentralization ought not to be sacrificed for the sake of protecting partial political principles and values.

3 Blockchain Systems and Decentralization

Blockchain technology emerged amid the alleged ‘crisis of theory and practice’ in the West. Based on the historical context and ideological forces that accompanied their development, blockchain could be framed as a reaction to, as well as a continuation of, liberal democracies’ failure to distribute ‘power to.’ 2008 marked the year when the Global Financial Crisis, the most serious of its kind since the Great Depression of 1929, began to spread around the world's nations. The catalyst of this economic debacle was a housing bubble in the land pioneering political

liberalism: the United States. That same year, developer(s), under the pseudonym of Satoshi Nakamoto, shared a paper titled "Bitcoin: A Peer-to-Peer Electronic Cash System" (Nakamoto, 2008) with the Cryptography Mailing List. The mailing list had been created by Cypherpunks interested in exchanging information and ideas on privacy-preserving technologies. Influenced by previous attempts to develop a 'decentralized digital cash' system, Nakamoto's paper presented a solution to the double-spending dilemma. The answer was blockchain technology, a distributed ledger or database combining peer-to-peer networks, public-key cryptography, consensus algorithms, and hashing functions.

The raw data of Bitcoin's genesis block famously contained a reference to a headline in The (London) Times: ‘The British government was about to bail banks out for a second time.’ However, Bitcoin would not need to be bailed out: it brought about a payments system with a fixed supply of the cryptocurrency, where transactions did not require a middleman, corporate or government entity, to be validated. The cryptocurrency disseminated the ‘power to’ transact from central public and private authorities to individuals. Soon enough, Bitcoin caught the attention of anarchists, libertarians, and utopian technologists alike, interested in protecting individual privacy and promoting greater freedom away from the coercive hands of governments and States (Brunton, 2019). It did not take long before enthusiasts considered applying its underlying technology beyond digital payments alone. The Ethereum whitepaper, published by the programmer Vitalik Buterin (2014), devised a blockchain to create and deploy smart contracts or small snippets of code whose encoded rules are self-executing (Buterin, 2014). The *chain of blocks* expanded its reach as a general-purpose technology to regulate interactions in various fields. The application areas seemed endless: from finance to trade, governance, identity registries, health care, art, gaming, and novel organizational structures such as Decentralized Autonomous Organizations (DAOs).

For the purpose of this paper, ‘blockchain systems’ refer to decision-making systems that rely on or operate through public and permissionless blockchains. These complex decision-making systems are multi-layer and multi-stakeholder. Firstly, they encompass several interconnected layers of software and hardware, including a data, network, consensus, incentive, contract and application layers. On the other hand, there are multiple actors building, interacting, maintaining, or utilizing blockchain systems. These go from miners or transaction validators to nodes, token holders, third-party application providers, exchanges, hardware manufacturers, public institutions, end-users, and more. The multi-layered, multi-stakeholder nature of blockchain systems is essential for two reasons. On the one hand, it helps us understand the governance of these decision-making systems: how decisions are debated and implemented within them. By ‘governance,’ we refer to both *by* the infrastructure and *of* the infrastructure. Governance *by* the infrastructure includes rules directly embedded in the technological system, such as consensus algorithms defined by a blockchain protocol or provisions on token distribution hard-coded into a protocol or smart contract. Governance *of* the infrastructure involves social and institutional rules that operate outside the technology architecture and are not directly encoded. For example, social rules on how to submit proposals, debate, and vote on proposals to change features of the protocol, as well as laws and policies from a particular national jurisdiction trying to regulate how these systems should operate (De Filippi & McMullen, 2018). On the other hand, the multiple technological layers, stakeholders, and governance rules influence how decentralized blockchain systems are: how dispersed the ‘power to’ truly is.

Blockchain systems have been popularized as tools with the potential to revolutionize several aspects of our lives, including democratic governance, due to their ‘decentralized’ (Atzori, 2015). Indeed, public and permissionless blockchain systems, as opposed to private and permissioned or hybrid ones, require no previous authorization for participants to join the network freely, make transactions, or validate them (Pilkington, 2016). They are disintermediated, given that the network does not rely on nor cannot be controlled by a single, unique central authority (Swan, 2015). However, recent academic writings have warned that a common mistake of proponents of blockchain systems as revolutionary and ‘democratizing’ has been equating ‘disintermediation’ with ‘decentralization’ (Fard Bahreini et al., 2021). Most of these claims focus on technical decentralization as a binary property at the network layer, where blockchain disseminates power to validate transactions from a central actor to the participants of the network. More recent attempts have tried to measure decentralization more holistically, as a non-binary dynamic property of multi-layer, multi-stakeholder blockchain systems. The Nakamoto Coefficient was one of the first tries to calculate the level of decentralization of public and permissionless blockchains by establishing the minimum amount of ‘entities’ required to ‘monopolize the control’ of an ‘essential subsystem’ and compromise it (Srinivasan & Lee, 2017). Other researchers took on the idea of subsystems and developed a more comprehensive taxonomy of decentralization across six layers and thirteen factors. The value-added of their approach is that it incorporates aspects of governance *by* and *of* the infrastructure. Particularly, they highlight a governance layer, focusing on owner control and improvement protocol factors. Owner control refers to the fraction of the native

cryptocurrency or token kept by the founding developers of a blockchain system after early adoption. On the other hand, improvement protocol focuses on the number of participants authorized to moderate the debates (Sai et al., 2021).

Against popular claims, the findings of Srinivasan & Lee (2017) and Sai et al. (2021) argue that Bitcoin and Ethereum networks exhibit levels of centralization among different subsystems. However, neither writing is critical of decentralization as a means nor fully attempts to explain why these systems trade-off decentralization. Below, we will try to address these issues.

4 Instrumentalization of Decentralization in Blockchain Systems

What is driving this instrumentalization? After reviewing the most common reasons cited by leading actors in blockchain networks as well as researchers, we found that these fall into the same two categories that also apply to liberal democracies: either to promote a value system or as a result of a group trying to increase their 'power over.'

A widely cited explanation for lowering the level of decentralization is to 'scale' or 'secure' blockchain systems. Less dissemination of 'power to' is seen as an inevitability, under certain circumstances, posed by technological constraints. A foundation of this thesis is the well-known 'blockchain scalability trilemma,' a term coined by Vitalik Buterin. Most blockchain systems allegedly struggle to be decentralized, secure, and scalable simultaneously. Under this trilemma, scalability is their ability to handle increasing transactions at the same speed and cost. Security is their ability to resist attacks by a 'critical' number of nodes, depending on the consensus algorithm. Decentralization is their ability to build consensus without relying on any small group of actors and to allow entry of new nodes without any barriers (Buterin, 2021). Scholars have echoed the trilemma: decentralization has to be compromised to maintain the network's security or handle a higher volume of transactions at the same speed or costs (Hsieh et al., 2017). Interestingly enough, Buterin himself argues that the trilemma is not inescapable. There are, allegedly, technological ways to ensure systems keep all three properties (Buterin, 2021). We contend that the veil of neutrality under these 'technological justifications' to the instrumentalization of decentralization resembles, to some extent, Dyzenhaus' critique of political liberalism. Just as the political principles of justice were not neutral, technological justifications are not unequivocal universal truths. Albeit anchored on some observable challenges of distributed systems, these arguments carry presuppositions of scalability, decentralization, and security and how their interaction should be. Take the example of an update: what makes an update an 'improvement'? These 'provisions of the right' may also contain 'provisions of the good' - very much like normative values.

Higher levels of centralization resulting from interest group capture is another reason raised. Blockchain systems, multi-layered and multi-stakeholder, are not immune to the attempts of "conglomerates" to control aspects of the blockchain governance of the infrastructure and by the infrastructure and increase their power (Ferreira et al., 2022). These conglomerates may comprise diverse stakeholders, most notably miners or

transaction validators, nodes, developers, token holders, hardware manufacturers, or even public institutions. The shared presupposition is that these actors wish to increase their 'power over,' not 'power to' (Avelino, 2021), which means power over other system members to their benefit. Blockchains to date cannot prevent captures from happening or, at worst, may incentivize them. Objective incentives embedded in blockchain protocols, which usually offer some reward, combined with resourced-based consensus algorithms such as Proof-of-Work or Proof-of-Stake (Siddarth et al., 2020), encourage the emergence of plutocratic, centralized groups. Even when decisions in blockchain systems cannot be enforced by a centralized party and require consensus from the rest of the nodes, 'Schelling points' facilitate interest captures. A concept drawn from game theory, the Schelling point indicates the choice each actor thinks everyone will make. As such, conglomerates may have an advantage in stirring decision-making to their benefit by influencing the perception of others that their decision is the preferred one by the majority of the network (De Filippi et al., forthcoming). We argue that this approach to the instrumentalization of decentralization resembles critical theories of liberal democracies. They see the legal and political orders, as well as the alleged 'overlapping consensus,' as a mere crystallization of power structures of oppression.

Overshadowed by the famous SegWit soft fork, the 2013 Bitcoin hard fork remains vastly under-researched and under-documented. It is perhaps one of the most compelling examples of centralization of the Bitcoin network at the level of governance on the improvement of the protocol. This factor was not flagged as particularly centralized by Sai et al. (2021). On March 11, 2013, a severe incompatibility issue between Bitcoin client 0.7 and 0.8 versions caused the main chain to fork into two. Once the "incidental fork" was detected, a handful of Bitcoin Core lead developers quickly deliberated on the course of action in the #bitcoin-dev IRC channel. The options seemed clear: instruct miners and merchants to upgrade to the 0.8 version and stick to the newer chain or downgrade to the 0.7 version and the older chain. One of the Bitcoin mining pools, BTC Guild, jumps into the conversation. At the time, BTC Guild controlled around 20-30% of the Bitcoin network's hashing power.

23:43	BTC Guild	I can single handedly put 0.7 back to the majority hash power
		I just need confirmation that thats what should be done
23:44	Pieter Wuille	BTC Guild: imho, that is was you should do, but we should have consensus first

Fig. 1. An excerpt of the conversation on the #bitcoin-dev IRC channel (Narayanan, 2015)

After another mining pool expressed it was better to downgrade to the 0.7 version, the Bitcoin developers started reaching out to miners and merchants, instructing them to do the same (Narayanan, 2015). The crisis was resolved in a matter of six hours. In Narayanan's words: "So much for decentralization! The fact that BTC Guild can tip the scales here is crucial" (2015).

Those who praised the action taken by the miners and developers stressed that centralized and swift action was necessary and that the decision made prevented

malicious attackers from figuring out ways to create double-spend transactions. Decentralization had to be traded off to ensure the network's security (Narayanan, 20215). But was the security of the network indeed being threatened? Some didn't think an intervention was inevitable or necessary, including Vitalik Buterin. According to him, if the developers had done nothing, the Bitcoin network would have continued as it was, albeit with some monetary loss. In Buterin's words: "Bitcoin is clearly not at all the direct democracy that many of its early adherents imagine and, some worry, if a centralized core of the Bitcoin community is powerful enough to successfully undertake these emergency measures to set right the Bitcoin blockchain, what else is it powerful enough to do?" (Buterin, 2013). Buterin's argument seems to imply that decentralization decreases as a result of capture by a conglomerate of Bitcoin core developers and miners, with the power and incentives to steer the Schelling point in their preferred direction. One could assume that the motivation of BTC Guild was to minimize revenue losses if the fork was left unattended and the chain continued to split. The handful of Bitcoin core lead developers who intervened may have feared losing support from other developers to their perceived leadership if the 'crisis' was not solved swiftly. The preferred action, in this case, was intervening and instructing the network on what to do, centralizing decision-making power.

The 2013 Bitcoin hard fork crisis illustrates some crucial points on the role of decentralization blockchain systems. Some authors claim that decentralization in blockchain systems, inherently inefficient, ought to be justified by an external rationale (Magnuson, 2020). We contend that, even if decentralization was inefficient or costly for decision-making, it is a precondition to the system's legitimacy.

5 Conclusion

In this paper, we have observed that blockchains, as decision-making systems, enhance some aspects of decentralization compared to liberal democracies. Transactions do not require the validation or attestation of a central authority. Beyond this, the claim that blockchain systems lead to greater 'democratization' is an oversimplification of the understanding of 'decentralization,' both conceptually and empirically. To begin with, decentralization needs to be addressed holistically, as a non-binary and dynamic feature, reflecting dissemination of 'power to' in symbiotic relationship with participation. Furthermore, there is the question of the role that decentralization plays. A common feature of both systems is that decentralization becomes instrumental to an ulterior purpose, be it to promote a particular value system or to increase the 'power over' of some oppressive groups within them. There is, however, a real value in building affordances to protect decentralization, particularly the power to participate in decision-making on-chain and off-chain: it is a precondition to the legitimacy of these systems. Further research is encouraged to study how decentralization as an end can be operationalized across different layers to better serve the purpose of each blockchain system in the specificity of its own context of application.

References

- Atzori, M. (2015). *Blockchain technology and decentralized governance: Is the state still necessary?*. Available at SSRN 2709713.
- Brunton, F. (2019). *Digital Cash: The Unknown History of the Anarchists, Utopians, and Technologists Who Created Cryptocurrency*. Princeton University Press. <https://doi.org/10.2307/j.ctvc77f9r>
- Buterin, V. (2013). *Bitcoin Network Shaken by Blockchain Fork*. Bitcoin Magazine. <https://bitcoinmagazine.com/technical/bitcoin-network-shaken-by-blockchain-fork-1363144448>
- Buterin, V. (2014). *A next-generation smart contract and decentralized application platform*. white paper, 3(37), 2-1.
- Buterin, V. (2021). *Why sharding is great: demystifying the technical properties*. vitalik.ca. <https://vitalik.ca/general/2021/04/07/sharding.html>
- Chow, C. (2018). *Blockchain for Good? Improving supply chain transparency and human rights management*. Governance Directions, 70(1), 39-40.
- Crook, R., & Manor, J. (2000). *Democratic Decentralization*. OED Working paper Series, No, The World Bank, Washington DC.
- Dardanelli, P., & Wright, K. (2021). *Devometrics: How to Measure Decentralisation? A Review of the Literature*. Local Democracy Research Centre at LGIU.
- De Filippi, P., & Hassan, S. (2018). *Blockchain technology as a regulatory technology: From code is law to law is code*. arXiv preprint arXiv:1801.02507.
- De Filippi, P., & McMullen, G. (2018). *Governance of blockchain systems: Governance of and by Distributed Infrastructure*. Doctoral dissertation. Blockchain Research Institute and COALA.
- De Filippi, P., & Wright, A. (2018). *Blockchain and the Law: The Rule of Code*. Harvard University Press. <https://doi.org/10.2307/j.ctv2867sp>
- De Filippi, P., Reijners, W. & Mannan, M. (forthcoming). *Blockchain Technology and the Rule of Code: Regulation via Governance*.
- De Filippi, P. (2021). *Blockchain Technology as an Instrument for Global Governance*. Digital, Governance and Sovereignty Chair, 1-16.
- Durbin, P. (2001). *Building Democracy from the Grassroots*. Inter-American Foundation and the Unit for the Promotion of Democracy of the General Secretariat of the Organization of American States.
- Dyzenhaus, D. (1999). *Legality and Legitimacy: Carl Schmitt, Hans Kelsen, and Hermann Heller in Weimar*. Oxford University Press. DOI: 10.1093/acprof:oso/9780198298465.001.0001
- Easton, D. (1965). *A Framework for Political Analysis*. Englewood Cliffs: Prentice-Hall.

- Fard Bahreini, A., Collomosse, J., Seidel, M. D. L., Sotoudehnia, M., & Woo, C. C. (2021). *Distributing and Democratizing Institutional Power Through Decentralization*. Building Decentralized Trust, 95-109. Springer, Cham.
- Fukuyama, F. (1989). *The End of History?* The National Interest, 16, 3–18. <http://www.jstor.org/stable/24027184>.
- Habermas, Jurgen (1984) *Theory of Communicative Action* (Thomas McCarthy, trans). Boston: Beacon Press. (Originally Published in 1981).
- Hart, D. K. (1972). *Theories of government related to decentralization and citizen participation*. Public Administration Review, 32, 603-621.
- Heller, H. (1971). *Teoría del Estado* (Gerhart Niemeyer, trans.) México: Fondo de Cultura Económica. (Originally Published in 1934).
- Hsieh, Y. Y., Vergne, J. P. J., & Wang, S. (2017). *The internal and external governance of blockchain-based organizations: Evidence from cryptocurrencies*. Bitcoin and beyond, 48-68. Routledge.
- Hughes, K. (2017). *Blockchain, the greater good, and human and civil rights*. Metaphilosophy, 48(5), 654-665.
- Kaim, M. (2021). *Rethinking Modes of Political Participation: The Conventional, Unconventional, and Alternative*. Democratic Theory, 8(1), 50-70.
- Lemieux, V. L., & Feng, C. (2021). *Building Decentralized Trust*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-54414-0>.
- Magnuson, W. (2020). *Blockchain democracy: Technology, law and the rule of the crowd*. Cambridge University Press.
- Nakamoto, S. (2008). *Bitcoin: A peer-to-peer electronic cash system*. Decentralized Business Review, 21260.
- Narayanan, A. (2015). *Analyzing the 2013 Bitcoin fork: centralized decision-making saved the day*. Freedom To Tinker. <https://freedom-to-tinker.com/2015/07/28/analyzing-the-2013-bitcoin-fork-centralized-decision-making-saved-the-day/>.
- Okin, S. M. (1979) *Women in Western Political Thought*. Princeton University Press.
- Parekh, B. (1992). *The cultural particularity of liberal democracy*. Political Studies, 40, 160-175.
- Pilkington, M. (2016). *Blockchain technology: principles and applications*. Research handbook on digital transformations. Edward Elgar Publishing.
- Rawls, J. (2020). *Political liberalism*. The New Social Theory Reader, 123-128. Routledge. (Originally Published in 1993).
- Rawls, John. (1971) *A Theory of Justice*. Harvard University Press. <https://doi.org/10.4159/9780674042605>.
- Reijers, W., Wuisman, I., Mannan, M., De Filippi, P., Wray, C., Rae-Looi, V., Cubillos Vélx, A. & Orgad, L. (2018). *Now the Code Runs Itself: On-Chain and Off-Chain Governance*

- of Blockchain Technologies*. Topoi 40, 821–831. <https://doi.org/10.1007/s11245-018-9626-5>.
- Reinsberg, B. (2021). *Fully-automated liberalism? Blockchain technology and international cooperation in an anarchic world*. International Theory, 13(2), 287-313.
- Sai, A. R., Buckley, J., Fitzgerald, B., & Le Gear, A. (2021). *Taxonomy of centralization in public blockchain systems: A systematic literature review*. Information Processing & Management, 58(4), 102584.
- Schmitt, C. (2008). *The Concept of the Political* (George Schwab, trans.) Chicago and London: The University of Chicago Press (Originally Published in 1932).
- Shils, E. (1995). *Some of the modern roots of liberal democracy*. International Journal on World Peace, 3-37.
- Siddarth, D., Ivliev, S., Siri, S., & Berman, P. (2020). *Who watches the watchmen? a review of subjective approaches for sybil-resistance in proof of personhood protocols*. Frontiers in Blockchain, 46.
- Srinivasan, B. S., & Lee, L. (2017). *Quantifying decentralization*. news. earn. com. <https://news.earn.com/quantifying-decentralization-e39db233c28e>.
- Swartz, L. (2017). *Blockchain dreams: Imagining techno-economic alternatives after Bitcoin*. Another economy is possible: Culture and economy in a time of crisis, 1.
- Yeung, K. (2019). *Regulation by Blockchain: the Emerging Battle for Supremacy between the Code of Law and Code as Law*. The Modern Law Review, 82(2), 207-239.

Industrial Limitations on Academic Freedom in Computer Science

Reuben Kirkham¹[0000-0002-1902-549X]

¹ Monash University, Victoria, Australia
reuben.kirkham@monash.edu

Abstract. The field of computer science is perhaps uniquely connected with industry. For example, our main publication outlets (i.e. conferences) are regularly sponsored by large technology companies, and much of our research funding is either directly or indirectly provided by industry. In turn, this places potential limitations on academic freedom, which is a profound ethical concern, yet curiously is not directly addressed within existing ethical codes. A field that limits academic freedom presents the risk that the results of the work conducted within it cannot always be relied upon. In the context of a field that is perhaps unique in both its connection to industry and impact on society, special measures are needed to address this problem. This paper discusses the range of protections that could be provided.

Keywords: Academic Freedom; Computer Science; Industry.

1 Academic Freedom: What is it and why does it matter?

Academic Freedom (otherwise known as Intellectual Freedom) is a civic right of considerable importance. It is about ensuring that individual academics are in the position to question received wisdom and speak truth to power. As such, academic freedom is intimately connected to the role of a University in the ‘search for truth’ (Hudson & Williams, 2016) and is “*at the very core of the mission of the university*” (Altbach, 2001). This means that a culture where academic freedom is fully supported is necessary to ensure that the knowledge produced by an academic community can be fully relied upon. More than ever (perhaps especially given the COVID-19 pandemic), this is fundamentally important to wider society: academic freedom is a key part of ensuring that the general public can trust and rely upon research findings.

It is difficult to understate how fundamental academic freedom *should* be in academia. It is intended to play “*an important ethical role not just in the lives of the few people it protects, but in the life of the community more generally*” to ensure that academics do “*not [feel compelled] to profess what one believes to be false*” and amounts to “*a duty to speak out for what one believes to be true*” (Dworkin, 1996). Moreover, it has a fundamental societal importance beyond the pursuit of truth itself, with academic freedom also playing a “*valuable role ... in supporting democratic government*” (Evans & Stone, 2021).

In our field, existing ethical codes expect us to take an active role in the community and advance ‘social good’. For example, the ACM Code of Ethics expects computing

professionals (including academics) to "*give comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks.*" (2.5), "*foster public awareness and understanding of computing, related technologies, and their consequences*" (2.7) and "*ensure that the public good is the central concern during all professional computing work*" (3.1). Yet, this ethical framework does not directly address academic freedom, and thus does not deliver a culture where these ethical principles will likely be put into practice. Nor has there been a publicized case where the ACM or the IEEE have taken steps to enforce academic freedom, which suggests at the least, a lack of prominence in their organizational thinking in respect of this concern.

The concern of academic freedom is perhaps of particularly strong importance in computer science. It is difficult to think of a research field that has a greater permeation into our day-to-day lives, especially in light of the global pandemic. For instance, computational modelling has been used as a basis for locking down populations, whilst our field has also been essential for enabling people to remain connected whilst socially distancing in the physical world. This means that the work conducted in our field has life-altering consequences for entire populations, making our work carries a particular social importance.

If we as a field are to be worthy of societies trust, then it is of considerable importance to ensure there are high standards of academic freedom. In practice, academic freedom tends to suffer from improper limitations. This means that other people are able to restrict or chill speech, or the research that is conducted. The undermining of academic freedom can come from a range of sources, such as other academics, an academics own institution, public funding bodies, or industry (Evans & Stone, 2021; Hudson & Williams, 2016). This paper focusses on the particular arrangements between industry and academia within the field of computer science and explains how they undermine academic freedom. It also proposes how existing practices should change so the public can have more confidence in the research conducted within our field.

2 How might academic freedom be uniquely limited in computer science?

Unfortunately, academic freedom is particularly restricted within computer science, relative to most other fields. One limitation arises from our fields extensive connection to industry, where unlike other fields, employs a large number of the best scientists in quazi-academic roles (Ebell et al., 2021). Just to give an example of the scale of this, Microsoft Research has been described as the "*largest computer science department*" in the world¹, whilst being listed as a partner on over £340 million of EPSRC funding.²

¹ <https://www.computerworld.com/article/2547042/microsoft-research--at-15--looks-ahead-to-more-innovations.html>

² See <https://perma.cc/L3TP-CQQB>

The issue is there is a considerable interconnection between large technology companies and the academic community. For a computer science academic, this means:

1. Industry influences who is awarded research grant funding. This is either by direct provision (e.g. Google or Microsoft grants), or by the increasing expectation of industry ‘partners’ on research grant applications to public funders (e.g. UKRI, or the ARC).
2. Many other career opportunities, including future employment, are also controlled by industry. This is particularly so for PhD students, many of whom undertake industry-based internships (and are often provided through connections to PhD supervisors). PhD students also usually have a subsequent career in industry after they graduate (instead of an academic career).
3. Academic careers are intrinsically driven by publications – many of these publications are supported by industry, be it by way of collaborations, or the provision of access to datasets or other materials. These resources are regularly controlled by large companies (e.g. Google, Facebook, Microsoft), giving them a direct influence over the publication capacity of many academics.
4. Academic conferences, which have a co-equal status to journals (uniquely in Computer Science) are not only sponsored by large companies, but their employees are routinely involved in the peer-review process, including by acting in an editorial role (e.g. as program committee members or chairs, or as journal editors).

It might be said that industry has a Jekyll and Hyde type relationship with the academic community in computer science. In its benevolent form, this relationship offers a wide range of expertise and opportunities – these are perhaps especially important for PhD students, many of whom will graduate to a well-paid job in industry. Yet, the types of concerns raised above are likely to have a considerable chilling effect on the free speech of computer science researchers and by implication, academic freedom. This is because many of the (particularly powerful) technology companies have the ability to heavily influence academic careers. This is also compounded by the increasing practice of university’s to measure publication performance and research grant income, and even to set targets in this regard for academic staff members. At the same time, there is already a lack of a debate culture within certain fields (e.g. see the account in (Button et al., 2015) in respect of HCI or (Kirkham, 2021) for an example of attempts to discourage certain types of research). This means that an important public function is likely to be being chilled on the ground: whilst this is certainly not only by industry (other academics are also often part of the problem), it is one important part of the problem to be addressed.

3 What might be done about this?

The overriding issue is ensuring that academic decision making is not compromised by powerful industrial interests, or the ‘chilling effect’ arising from the power that industry can hold over individual academic careers. The way to address this is to ensure that

decisions about individuals' academic careers cannot be influenced by industry. For wider debate, I would propose a number of steps that might be undertaken:

3.1 Not measuring industrial funding as a performance measure.

Unfortunately, it is common for grant income to be used as a measure of 'research performance' (and thus to determine whether an academic will continue to be employed), as well as also serving as a means for advancing academic careers in and of itself (e.g. by enabling more research to be done, and thus increasing the volume of publications). This is a major ethical issue in and of itself, even if industry were not involved. Existing competitions for research grant income between researchers have been long known to be a highly flawed measure of research performance (see e.g. (Gillett, 1991)), as well as amounting to a considerable waste of public resources (Domingos, 2022). Furthermore, the existing practice in most grant allocation processes – especially the failure to adopt anonymous blind review – is well known to promote discrimination against certain minority researchers in respect of funding decisions (Wittman et al., 2019). Of course, the existence of discrimination also suggests an element of capriciousness, with the decision makers being vulnerable to making decisions on irrelevant factors, rather than purely the matters they ought to be considering: indeed, it is well recognized that many funding processes have a "huge amount of randomness" (see e.g. (Grove, 2021)). Furthermore, the idea of a research grant system is itself inconsistent with academic freedom: the process involves 'senior' academics deciding what research other 'junior' researchers will be funded to do and thus amounts to an effective form of censorship of other researchers.

The provision of funds by industry further compounds this problem, making a bad situation worse. The effect of these financial provisions is threefold:

- Individual companies, or individuals within them, have the ability to advance the careers of specific researchers, both by providing funding directly and the prestige and/or status that comes with this funding (e.g. it is currently seen as prestigious to get funding from Google, Microsoft, Facebook etc.). This means they have the power to advantage the careers of researchers who support their commercial interests and agenda over those who do not.
- Individual companies can shape the research agenda of a field, by funding academics to work on specific research directions. This potentially distorts the research that is conducted. This is perhaps especially problematic if there is match funding provided by an academic department (e.g. a small Facebook grant ends up being topped up into being a full PhD scholarship).
- The amount of overall grant funding increases, thus increasing the grant targets imposed on academics overall, and penalizing those who do not wish (or are less able) to obtain industry funding.

In some respects, this is something that can be relatively easily resolved. First, academic institutions should not count industry supported research funding (including from other sources where industry 'partners' were involved in supporting the funding application) as an indication of performance, given the potentially corrosive effects on academic

freedom. Second, there should be an effort to eliminate ‘Matthew’ effects, so that if something is funded by industry, then other resources (e.g. discretionary funds) should be diverted into areas that industry is less likely to fund (for example areas that are of social or public importance). This would be different from the present situation, where a ‘profit’ normally gets sent to central funds (e.g. to fund buildings or administration) or alternatively to the research funds of the researcher who obtained the industry funding to begin with (thus rewarding them further), rather than being spent on alternative research. Third, there should be a prohibition on ‘match’ funding: instead, industry funded research should in effect be heavily ‘taxed’ (e.g. by a heavy ‘overhead’ charge) that gets redirected into non-industry related research by other academics.

Industry can take some positive steps too. I would argue that industrial organizations who act ethically should not be giving funding to *named* individual researchers, but instead funding organizations, thus dampening the individual prestige effect. Furthermore, if industry wishes to sponsor research, then a better position would be for this to be done following a *double-blind review* process where the reviewers are independent of the relevant corporate interests. There is nothing wrong with industry funding an area of research, provided that this funding is done in a manner that is conservative of academic freedom. The problem is that industry funding is presently being used in an inappropriate manner, and academic institutions have not put in place measures to deal with this problem.

A more radical perspective could be to take a more restrictive approach. Many academic institutions would not, for example, accept research funding from a ‘tobacco’ company (Thomson & Signal, 2005): this is due to the harm such research can impose on public health (Turcotte, 2003). Might there not be room for a similar approach towards certain technology companies that seek to fund research in University’s, whilst not fully respecting academic freedom? At the least, bodies like the IEEE and the ACM could make it a requirement that they will only accept papers funded by an appropriate source, i.e. where the underlying funding has been provided in an ethical manner which fully protects academic freedom.

3.2 Removing the role of industry in publicly funded grant competitions.

There is a related issue of concern. It is common for public bodies that fund research to favour applications that have already had industrial support (e.g. someone has had awards, collaborations or funding from Microsoft, Google or Amazon or other companies), or by the provision of future in-kind industry support with respect to a grant application. The purpose is to ‘translate’ academic research into tangible outcomes and benefit the economy. The difficulty is that the present approach of doing so risks undermining academic freedom, by providing a mechanism by which industry can favor certain academics, or their institutions. Beyond the direct industry influence, such a process is also unfair more generally, in that it fails to follow a *double-blind review* process, something which is well known to discriminate against minority groups, as well as itself risking academic freedom. This risk is because an anonymous reviewer – which might themselves have some connection to industry, or dislike another exercise of academic freedom – can reject a proposal without accountability

based on the identity or activities of the applicant, rather than the merits of their research.

Addressing this issue requires a focus on the requirement that is being imposed on a researcher, namely that the researcher themselves has a pre-existing *relationship* with a given sponsor, and this relationship helps their chances (or is a hard requirement to succeed) in these competitions. It is the *a-priori* quality of the relationship that is the problem: if there were a structured means for involving industry *after* the grant has been awarded, then the difficulty would likely disappear. There are various *post-hoc models* that can bring this about, whilst conserving academic freedom. For example:

- Having a panel of industrial organisations, who are allocated the most appropriate projects on a *post-hoc* basis, *after* an award/funding decision has been made.
- Placing more of an emphasis on start-ups and small-businesses (instead of large companies), which considerably reduces the ethical risk (these organizations could also be anonymized in the peer-review process).
- Supporting academics to start their own companies and startups, thus mostly sidestepping existing technology companies.

Adopting any of these models would be in the public interest. By doing it on a *post-hoc* basis, it is also arguable that collaborations would be better formed: rather than being based on who is already ‘connected’, the most appropriate industry organisation can later be assigned as the project partner. Industry and academic time would be saved, because only the (small proportion) of grant proposals which are eventually funded would need partnerships to be arranged. Furthermore, if configured appropriately, smaller and medium sized businesses would benefit, due to the reduction of unfair competition (e.g. researcher applicants don’t need a recognized ‘name’, but can focus on partnering with the most appropriate organizations when they have the funding down the track). The wider benefits, and the likelihood of this better serving industry (especially by mitigating the present bias in favour of large international tech companies), enhances the ethical case for reform in this regard. To put it another way, with carefully thought-out processes, the involvement of industry in supporting research can co-exist with academic freedom: the problem is the absence of appropriate policies.

3.3 Reducing the unfair influence of industry resources

For certain favoured academics, industry provides them with a range of resources, which helps advance their research. This does not only include funding, but also access to expertise, existing trade secrets, software tools, and data. In many cases, especially in respect of large technology companies, it is possible that certain studies can only be done with this level of access or assistance – for instance, a study might need the ability to control the platform in question (e.g. by tweaking the data presented to a proportion of end users), or require access to the enormous amount of data available to large corporations in respect of their own software platforms. To put it another way, many

experiments on Facebook (and other platforms) require the consent and co-operation of Facebook itself.

One issue is that the provision of these resources means that some academics will have an advantage over those who are less favored by these companies. There is a real risk that the power that these companies have can influence both the research, and the researcher. This is because of the possibility of this support being withdrawn, which amounts to a chilling effect. In turn, this means it is difficult to have confidence in the research, or indeed researchers who are associated with such matters. Yet at the same time, industry is substantially contributing to research by:

- Providing additional research resources that would otherwise not exist, thus increasing the volume and extent of the research that is done.
- Enabling research that can only be uniquely supported by industry.
- Providing opportunities for researcher development, such as internships for PhD students.

The issue is the existing lack of a governance framework that protects against the underlying chilling effect that arises from the influence of industry. The starting point should be that academics are not rewarded in their careers due to having industry connections (after all, networking ability and ones ‘connections’ in general has nothing to do with intrinsic academic merit, so rewarding this type of thing is arguably inappropriate to begin with), and the incentive models should be properly configured to protect against this. For instance, although papers supported by large companies should be reviewed and published, I would argue that they should not count as full publications for the purposes of academic promotions or other performance measures. Furthermore, active steps should be taken to remove the perceived ‘prestige’ of engaging with industry – it should be a matter of free-choice for individual academic staff if they engage (and if they do so, to what extent). Treating industry engagement as particularly prestigious is an inappropriate derogation from the principle of academic freedom, as it undermines a researcher’s individual freedom to *not* engage with it (or to engage only on limited terms).

3.4 Providing effective information access rights

There is a more fundamental concern that follows on from the foregoing one. This issue concerns a particular type of resource: information. I argue that there should be generally free access to a suitably-qualified academic to the operation of systems by large-scale technology companies. Instead of being a privilege for the ‘chosen few’ selected by industry (i.e. the status quo), whether a given academic has access should not be based on the patronage of a company (or the networks of that academic), but instead by way of provision made through a fair, merit-based and independent process. This process would mean that the legitimate interests of the company are protected, such as trade secrets and data security, whilst bona-fide academics have the right to conduct appropriate investigations, even if they are investigations the technology companies might find inconvenient.

It is notable that tech companies are particularly active when claiming to be ‘good’ corporations, with Google’s original motto famously being ‘do no evil’ (Crofts & van Rijswijk, 2020). In a modern democratic society, transparency is a recognized virtue. It is also an ethical standard that many in computing adhere to, perhaps exemplified by the ‘open source’ movement. At the same time, there are legislative expectations – transparency is an expectation associated with the GDPR (Wachter, 2018). Furthermore, when applied to public institutions – freedom of information is said to be an important “*democratic right*” (Walby & Luscombe, 2019) and given the role of some large technology companies (with some even having a market capitalization that goes beyond most countries annual GDP), there seems to be little reason why such information access rights should not apply to their operations too. One might go as far as to argue that the enhanced scrutiny that fair information access provides should be welcomed by those tech companies who like to impliedly assert that they are paragons of ethics. After all, a tech company acting sincerely should not be able to (or wish to) constrain research access to only ‘friendly’ researchers.

The ideal solution would be the expansion of freedom of information law to apply to these organizations as if they are public sector organisations: the problem is this is something that requires legislation (as well as an effective enforcement structure that FOI law tends to lack (Worthy, 2017)). However, there is a step that can be taken in respect of the underlying academic freedom issue: if there was not fair access to the underlying resource, then academic conferences and journals should refuse the submission. There are two reasons why this policy would be appropriate. The first is that the academic conference or journal cannot be sure of the reliability of the results, due to the lack of replicability, independence from the prevailing corporate interests, and the lack of any realistic ability to verify the data (for instance, if there was ‘cherry picking’ it would be very hard to investigate this). There is an obvious ethical risk in imbuing such work with the imprimatur of well-recognized publication venues, especially where industry has a particular interest in certain results and the endorsement by the conference or journal thus supports a particular commercial goal, rather than the public interest. The second is that the industry access is unfair to account for in individual careers: in other words, the industry involvement prevents a fair and meritocratic competition between academics, and also undermines academic freedom (for the reasons given in respect of 3.3). In effect, sufficient openness should become the price of entry into the academic community.

3.5 Adopting special considerations for papers that are designed to uncover wrongdoing.

Whether by accident or design, some research investigations end up uncovering improper practices, be it by industry, or other actors, such as parts of the state (e.g. as in the Post Office case in the UK (Wallis, 2021)). Alternatively, this work might uncover inadvertently problematic practices, but those which when identified, would have serious consequences for individual tech companies (for example a security bug in a computer chip). It follows that particular publications may have considerable commercial implications for some organizations, and thus advantage or penalize

particular industry players (depending on what the research has discovered). Yet there is an absence of a specialized procedure for reviewing such papers, even though they should be given particular attention for the following reasons:

- The risk of conflict of interest, which could either be in favour of or against acceptance of a particular work, depending on whether the work's specific outcomes favour a particular industrial organization's interests. A conflict of interest risks both reputational damage (and thus undermines the ethical imperative of public confidence in science), as well as the substantive fairness of the peer-review process (i.e. a risk of an unfair outcome). The involvement of industry in the determination of paper acceptance further amplifies this risk.
- The increased negative consequences of a false-positive acceptance, given the potential harm an error can cause to wider society. In short, special care is needed, to ensure the accuracy of this work. Peer-review is not always effective in accurately or rigorously identifying errors, so there is an enhanced concern with such works.
- The increased negative consequences of a rejection, which in most cases amounts to a delay (as the authors resubmit the work). In some cases, delayed publication may as well as be denied publication – the delay may undermine the underlying social impact that the work would have otherwise had. This issue is a well-known consideration in the context of Freedom of Information (where delayed information is often tantamount to denied information), and has equal force in this particular context.
- The fact that in some cases, the authors may in effect be whistleblowers, yet the publication process operates on the basis that the authors are not (for example) anonymous. There is one journal designed to deal with this issue – namely the *Journal of Controversial Ideas*, which allows anonymous paper authorship (McMahan et al., 2021) – but this is not necessarily configured to deal with matters of computer science, nor can it deal with most such cases that arise in our field.

Perhaps surprisingly, there is no such process at present in our field. Yet it would not be particularly difficult to establish one – it merely requires having a separate track with appropriate provisions and for the resources to conduct this enhanced (and more rapid) peer-reviewing to be provided. The only challenge would be funding – there is a likely need to pay reviewers so that decisions can be made with alacrity and to the required standard, but there is also a likelihood that reviewers may be more willing to do this type of reviewing in any event. However, this simply means that some resources within bodies such as the ACM or the IEEE need to be diverted to support this scheme, compared to other discretionary activities (e.g. by reducing the expenditure on the more lavish aspects of existing conferences' entertainment packages).

3.6 Requiring industrial related research to undergo special ethical review

There has been a history of problematic studies being conducted by industry, perhaps most infamously the ‘Facebook contagion’ study, as analysed in (Grimmelmann, 2015). This history demonstrates that there are significant ethical risks in respect of experiments conducted by large technology companies to human subjects. However, even if these organisations were to be subject to an independent IRB or other ethical review in respect of their own research activity (which they mostly are not), wider ethical risks remain that go beyond the direct subjects of experiments. Indeed, these wider ethical risks – i.e. social impacts – are the main concern. This means that a special ethical review process is arguably required that addresses these wider risks. In particular, such a process should consider matters such as:

- Does the industrial relationship risk the independence of the research?
- Was there a fair process for obtaining access to data and industrial resources?
- Are the resources that have been provided skewed in some manner, thus allowing the sponsoring company to shape the findings that arise, or lead to ‘uncomfortable’ questions being avoided within the research?
- Is there a risk of a corporate advantage being gained that has a wider ethical risk (e.g. by preventing fair competition, undermining individual consumer interests, or deflecting regulatory oversight)?
- Is there a risk to academic freedom, be it directly or indirectly?

These types of questions arguably should also be considered by a body that is genuinely independent of both the University and Industry. This is another way that such a process may need special constitution – existing IRB’s and ethical bodies are often not independent of the University, and thus may be at risk of being improperly pressured. Ideally, such matters should be considered by an independent administrative tribunal, much in the way that ‘freedom of information’ cases are considered. However, there would be a need to avoid the particular risks that arise with some administrative tribunals (e.g. the risk of bias in favour of political interests (Ellis, 2013)), as well as a lack of the relevant expertise and understanding of information technology (e.g. as with the UK’s information rights system (Kirkham, 2018)).

3.7 Requiring the support of academic freedom as a pre-requisite for participation

This final proposal is a more holistic one. Industrial organizations benefit from and participate within our academic community in a manner that does not tend to happen in other disciplines. Our system currently provides a range of soft support to industry, including access to our PhD students and graduates, the review of work conducted by industry-based professionals, and access to expertise within academia.

If they engage with and benefit from our community, then I would argue that we should expect in return that they respect the ground rules. As a minimum, this means a proper respect for the principle of academic freedom. Unfortunately, the academic freedom of industry-based researchers has not always been respected. A recent

prominent example is the case of Timnit Gebru and others, which putting aside other events, involved Google requiring the withdrawal of a paper submission ostensibly for perceived quality grounds (Ebell et al., 2021), and therefore undermining a key function of the academic peer-review process, whose role is to decide the merits of individual work. If Google – by arrogating the peer review function to itself – acts as a censor of submissions, then it is arguably unclear how it would be appropriate to allow submissions from researchers based at that organization, or to allow its senior researchers sit on program committees. Perhaps predictably, one effect of the Gebru case was to damage Google’s relationship with the academic community, with some conferences refusing to accept funding from them (Johnson, 2021).

It is arguable that an industry body that does not respect academic freedom should not be allowed to make paper submissions, have its researchers sit on program committees, or be otherwise connected to the academic community. Furthermore, given the ACM and IEEE code’s focus on individual conduct, one presumably expects that an academic in a University should not be supporting industry organizations who are insufficiently respectful of academic freedom, as it is difficult to see how doing so would amount to “*highest standards of integrity, responsible behavior, and ethical conduct in professional activities*” (per the IEEE code), or be “*encourag[ing the] acceptance ... of social responsibilities by members of the organization*” (per the ACM code). There is a caveat on this, in that the IEEE or ACM has not issued explicit guidance about how academics should (or more accurately should not) engage with industrial organizations who do not respect academic freedom, but as soon as this is spelled out, one would expect that computing professionals in academia would be careful to limit their engagement with them.

4 Conclusion

As a field, computer science is perhaps uniquely enmeshed with industry, which poses certain societal risks. This paper has identified a range of vectors by which industry is likely having a negative impact on academic freedom. This means that the part of academia that addresses matters of computer science is not always fulfilling one of its important public functions. This issue needs to be addressed for the good of wider society.

There is one more concluding remark that I should make. This paper deals with only one aspect of concern in respect of academic freedom: there is a wide range of other threats to academic freedom. For example, higher education institutions and academic researchers within them are also problematic in certain ways: this article of course, does not primarily focus on resolving these issues. In particular, there has been a lack of respect of academic freedom by some higher education institutions, with the Peter Ridd and Gerd Schröder-Turk cases in Australia serving as prominent recent examples (Evans & Stone, 2021). There have also been occasions where academics have engaged in campaigns or petitions aimed at undermining the freedom of other academics, with examples including the case of the computational geophysicist Dorian Abbot, who was subject to a campaign for politely expressing his views in support of ‘Merit, Fairness

and Equality’. Within computing, even the existing peer review process is concerning even without the industrial issue – for example, there have been attempts to accept or reject papers based on perceived ideological viewpoints (e.g. as discussed in Kirkham, 2021), and to engage in what is asserted to be ‘citational justice’ (see e.g. (Collective et al., 2021) for an concerning example of this), where the identity of the author is relevant, rather than the substantive contents of their research. The issue of academic freedom is a matter of wider importance, and it is important not to lose sight of the fact that industry is just one of the existing problems, albeit still a serious one that needs to be addressed as part of a wider debate. This paper aims to help start that debate.

References

- Altbach, P. G. (2001). Academic freedom: International realities and challenges. *Higher Education*, 41(1), 205–219.
- Button, G., Crabtree, A., Rouncefield, M., & Tolmie, P. (2015). *Deconstructing Ethnography: Towards a Social Methodology for Ubiquitous Computing and Interactive Systems Design*. Springer.
- Collective, C. J., Molina León, G., Kirabo, L., Wong-Villacres, M., Karusala, N., Kumar, N., Bidwell, N., Reynolds-Cuéllar, P., Borah, P. P., Garg, R., & others. (2021). Following the Trail of Citational Justice: Critically Examining Knowledge Production in HCI. *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, 360–363.
- Crofts, P., & van Rijswijk, H. (2020). Negotiating ‘evil’: Google, project maven and the corporate form. *Law, Tech. & Hum.*, 2, 75.
- Domingos, P. (2022). Pay researchers for results, not plans (<https://www.timeshighereducation.com/opinion/pay-researchers-results-not-plans>). *Times Higher Education*.
- Dworkin, R. (1996). We Need a New Interpretation of Academic Freedom. Academic Freedom and the Future of the University Lecture Series. *Academe*, 82(3), 10–15.
- Ebell, C., Baeza-Yates, R., Benjamins, R., Cai, H., Coeckelbergh, M., Duarte, T., Hickok, M., Jacquet, A., Kim, A., Krijger, J., & others. (2021). Towards intellectual freedom in an AI Ethics Global Community. *AI and Ethics*, 1(2), 131–138.
- Ellis, R. (2013). *Unjust by design: Canada’s administrative justice system*. UBC Press.
- Evans, C., & Stone, A. (2021). *Open minds: Academic freedom and freedom of speech of Australia*. Black Inc.
- Gillett, R. (1991). Pitfalls in assessing research performance by grant income. *Scientometrics*, 22(2), 253–263.
- Grimmelmann, J. (2015). Law and Ethics of Experiments on Social Media Users, The. *Journal on Telecommunications and High Technology Law*, 13, 219.

- Grove, J. (2021). Reputation risks of lotteries for research grants inhibit funders [<https://www.timeshighereducation.com/news/reputation-risks-lotteries-research-grants-inhibit-funders>]. *Times Higher Education*.
- Hudson, C., & Williams, J. (2016). *Why academic freedom matters: A response to current challenges*. Civitas London, England.
- Johnson, K. (2021). AI ethics research conference suspends Google sponsorship (<https://venturebeat.com/2021/03/02/ai-ethics-research-conference-suspends-google-sponsorship/>). *Venture Beat*.
- Kirkham, R. (2018). How long is a piece of string? The appropriateness of search time as a measure of ‘burden’ in Access to Information regimes. *Government Information Quarterly*, 35(4), 657–668.
- Kirkham, R. (2021). Why Disability Identity Politics in Assistive Technologies Research Is Unethical. *Moving Technology Ethics at the Forefront of Society, Organisations and Governments*, 475–487.
- McMahan, J. M., Minerva, F. M., & Singer, P. S. (2021). Editorial. *Journal of Controversial Ideas*, 1(1), 0–0. <https://doi.org/10.35995/jci01010011>
- Thomson, G., & Signal, L. (2005). Associations between universities and the tobacco industry: What institutional policies limit these associations? *Social Policy Journal of New Zealand*, 26, 186.
- Turcotte, F. (2003). Why universities should stay away from the tobacco industry. *Drug and Alcohol Review*, 22(2), 107–108.
- Wachter, S. (2018). The GDPR and the Internet of Things: A three-step transparency model. *Law, Innovation and Technology*, 10(2), 266–294.
- Walby, K., & Luscombe, A. (2019). *Freedom of information and social science research design*. Routledge.
- Wallis, N. (2021). *The Great Post Office Scandal: The Fight to Expose A Multimillion Pound Scandal Which Put Innocent People in Jail*. Bath Publishing Limited.
- Wittman, H. O., Hendricks, M., Straus, S., & Tannenbaum, C. (2019). Are gender gaps due to evaluations of the applicant or the science? A natural experiment at a national funding agency. *The Lancet*, 393(10171), 531–540.
- Worthy, B. (2017). *The Politics of Freedom of Information*. Manchester University Press.

Guidelines to Develop Trustworthy Conversational Agents for Children

Escobar-Planas, Marina^{1,2} [0000-0002-4513-020X], Gómez Emilia² [0000-0003-4983-3989] and Martínez-Hinarejos, Carlos-D¹ [0000-0002-6139-2891]

¹ Universitat Politècnica de València, Camino de Vera, s/n, 46022, València Spain

² European Commission, Joint Research Centre, Seville, Spain

marescplajob@gmail.com

Abstract. Conversational agents (CAs) embodied in speakers or chatbots are becoming very popular in some countries, and despite their adult-centred design, they have become part of children's lives, generating a need for children-centric trustworthy systems. This paper presents a literature review to identify the main opportunities, challenges and risks brought by CAs when used by children. We then consider relevant ethical guidelines for AI and adapt them to this particular system and population, using a Delphi methodology with a set of experts from different disciplines. From this analysis, we propose specific guidelines to help CAs developers improve their design towards trustworthiness and children.

Keywords: Conversational Agents, Children, Ethical Guidelines

1 Introduction

1.1 Motivation

Dialogue Systems, Virtual Assistants, Chatbots ... Conversational agents (CAs) have many different names, but they all refer to a computer program that supports conversational interactions with humans (McTear, 2020). Nowadays, CAs have become very popular, and recent developments allow people to interact with small computers placed in handy gadgets through voice: Google assistant can guide you in a car, Siri can send your messages in a smartphone, or Alexa can play music on a smart speaker. CAs are gaining popularity as a greater number of people are using them in their daily life.

Traditional CAs are composed of five different modules as illustrated in Figure 1: **ASR**, an automatic speech recognition engine that transforms audio speech inputs into text; **NLU**, a natural language understanding system that semantically interprets an input text; **DM**, a dialogue manager that manages the CA actions at the communication level and at the action level; **NLG**, a natural language generator that

translates the computer intent into text; **TTS**, a text to speech system that transforms text into audio output. Nowadays many systems use machine learning techniques to fulfil the task of one or several modules.

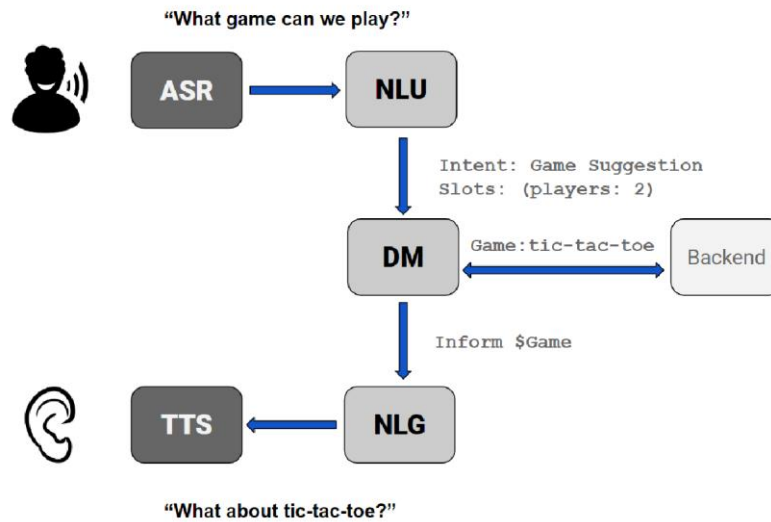


Fig. 1. Main modules of CAs.

Despite the adult centric design of CAs, accessibility and popularity among children should be carefully considered, since even the younger ones can interact with CAs through voice. For instance, Sciuto found out that children make a more extensive use of these devices and explore further capabilities than adults (Sciuto, 2018). This explains the huge impact CAs can have on the little ones. As an example, in a CNN article, Kelly quotes: “*The first four words my toddler understood were ‘mom’, ‘dad’, ‘cat’ and ‘Alexa’*” (Kelly, 2018). There is then a need to research the impact of CAs on children, being a vulnerable population widely exposed to these technologies. In addition, we need some ethical guidelines for the development of CAs that can be trusted by children.

1.2 Goals and structure of the paper

The goal of this paper is to provide some practical ethical guidance to CAs developers, considering children as a target audience. These guidelines are intended to enhance the opportunities of CAs while minimising the risks they may bring to this vulnerable population. Our study has then two main goals: (1) Identify opportunities vs risks of CAs for children and the main ethical considerations documented in the research literature; and (2) Adapt existing ethical guidelines for Artificial Intelligence (AI) to these particular systems (CAs) and target population (children).

These two goals are addressed as follows. Section 2 presents a literature review on the opportunities, risks and challenges of developing CAs for children, and an analysis of relevant ethical guidelines in the context of AI and children's rights. Section 3 presents our approach for adapting the two considered ethical guidelines to our particular context. Section 4 presents the obtained results, further discussed in Section 5. Section 6 summarises our main conclusions and steps for future research.

2 Literature Review

2.1 Opportunities, risks and learned lessons

The research literature has identified **opportunities** brought by CAs to children, summarised as follows:

- **Improvement of accessibility:** CAs can facilitate the interaction with computers to children too young to write, with dyslexia or physical disabilities (Catania et al., 2021; Pradhan et al., 2018).
- **Engagement of learning:** CAs can support information search (Landoni et al., 2020) language learning (Nasihati et al., 2018), or school material learning (Law et al., 2020; Xu and Warschauer, 2020).
- **Promotion of social behaviour:** CAs that requires the user to use persuasive strategies in games with them (Fraser et al., 2018), or help autistic children with their social skills (Zhang et al., 2020).
- **Support of health at homes:** CAs have been used to help recording treatments and track certain diseases (Sezgin et al., 2020).

Recent studies have also identified some **risks** brought by CAs and challenges that need to be considered, mainly bias due to adult-centric design of CAs, children's overtrust and potential unexpected impact. In fact, different issues have been identified depending on different modules (Table 1).

Table 1. Problems of CAs traditional modules with children.

Module	Children-specific characteristics
ASR	Speech acoustic characteristics, e.g. high pitch range, particular prosody.
NLU	Expressions, vocabulary and grammar.
DM	Information needs, protection, allowed functionalities according to age.
NLG	Need of simpler words or explanations according to age.

We have identified the following main **suggestions** to overcome the mentioned risks:

- **Communication abilities.** Children's speech and understanding should be considered in the interaction to improve children's inclusivity. Even with the current technology it is difficult to overcome these biases, however, some researchers try to improve CA's performance for children. Lavechin developed a speech identifier for babies (Lavechin et al., 2020), and other researchers identified good strategies to follow when a system does not understand a child (Cheng et al., 2018).
- **Dialogue management.** User's age should influence certain decisions of the system. Verbal and not verbal responses should also be appropriate.
- **Transparency.** Another relevant risk of CAs is to generate overtrust in children. Children tend to perceive CAs as friends, so CAs might influence them, e.g. in terms of data disclosure (Druga et al., 2020; Kahn et al., 2012). Straten taught us that transparent information helps to fight overtrust (Straten et al., 2020).
- **Continuous evaluation.** CAs are new in our lives, and the impact of their use on our society and children are still to be discovered. An example of an unexpected problem is parents asking Amazon to change the wake-up word "Alexa" in their CA product because their daughters, named Alexa, were suffering bullying at school due to this name coincidence (Johns, T., 2021). In consequence, evaluation and oversight of these devices have become highly relevant for the early detection of potential risks, and to implement the needed intervention practices.

2.2 Ethical guidelines for trustworthy AI and children

In recent years, several organisations have paid special attention to the ethical development of AI systems. Their aim is to generate awareness on AI systems, contribute to their understanding and evaluation, and study how to minimise the risks they can bring, while maximising their benefits. In this study, we focus on two main initiatives: the Ethical Guidelines for Trustworthy AI and UNICEF policy guidance on AI for children.

The High Level Expert Group (HLEG) of the European Commission developed the **Ethical Guidelines for Trustworthy AI** (AI HLEG, 2020), motivated by the need to protect people's fundamental rights in different contexts where AI systems are used. These ethical guidelines include seven requirements and are complemented by an assessment list for trustworthy AI (ALTAI), designed as a practical tool to help organisations self-assess the trustworthiness of their AI systems. ALTAI is a list of sixty-nine self-evaluation questions, grouped in the mentioned seven requirements as follows:

1. **Human agency and oversight** (11 questions). AI systems should respect human autonomy and decision-making, and should be supervised by humans.

2. **Technical robustness and safety** (21 questions). AI systems should be accurate, reliable and safe, having a preventative approach to risks.
3. **Privacy and data governance** (6 questions). AI systems should protect our privacy and have legitimate access to our data.
4. **Transparency** (5 questions). AI systems should have clear documentation and inform users about its decisions, capabilities and limitations.
5. **Diversity, non-discrimination and fairness** (10 questions). AI systems should ensure inclusion through all the AI system's life cycle.
6. **Societal and environmental well-being** (8 questions). AI systems should benefit the world and society.
7. **Accountability** (8 questions). AI systems should have mechanisms to ensure responsibility for development, deployment and use of AI systems.

The **UNICEF policy guidance on AI for children** is a guide intended to help policy makers and businesses by raising awareness of children's rights in the context of AI systems. It is proposed to complement existing work, guided by nine requirements that are presented in Table 2 (Digdum et al., 2021).

We performed a qualitative mapping between HLEG ALTAI and UNICEF AI for children requirements to understand the suitability of ALTAI having in account UNICEF consideration for AI and children's rights. From the detailed definitions, we related the 9 UNICEF requirements to the 7 requirements of HLEG. For instance, UNICEF requirement 9 refers to oversight, digital divide, and ethical development of AI from governments, that we connect to HLEG requirements on oversight, agency, fairness, and societal well/being. More details about the procedure to obtain this matrix is shared in <https://github.com/mescpla/CAs4Children-ETHICOMP22.git>.

We observe that ALTAI has a strong focus on the development and evaluation of specific AI devices. However, some UNICEF AI for children requirements have a strong focus on policies. Nevertheless, most requirements from UNICEF AI for children are connected to at least one major and some additional requirements of HLEG ALTAI, except for the educational aspect of requirement (8) which focuses on policies and is missing in ALTAI, which only refers to work and skills in HLEG requirement 6. In addition, all requirements present in HLEG ALTAI are covered by UNICEF AI for children guidance.

From this analysis we consider in the rest of our study the ALTAI list as a starting point, with a focus on CAs, complemented by the UNICEF guidelines on AI for children, as a complementary framework connected to children's rights, e.g. incorporating educational aspects.

Table 2 Mapping between HLEG ALTAI and UNICEF AI for children requirements. (*), () and (***) indicate low, mid or high correspondence between related requirements.**

UNICEF AI for children	HLEG ALTAI						
	Human Agency and Oversight	Technical Robustness and Safety	Privacy and Data Gov.	Transp.	Diversity, Non-discrim. and Fairness	Societal and Environm. Well-being	Account.
Support children's development and well-being <i>Let AI help me develop to my full potential</i>		**	**		**	***	
Ensure inclusion of and for children <i>Include me and those around me</i>					***		
Prioritize fairness and nondiscrimination for children <i>AI must be for all children</i>					***		
Protect children's data and privacy <i>Ensure my privacy in an AI world</i>	**		***				
Ensure safety for children <i>I need to be safe in the AI world</i>	**	***				**	
Provide transparency, explainability and accountability for children <i>I need to know how AI impacts me. You need to be accountable for that</i>	***	**		***	**		***
Empower governments and businesses with knowledge of AI and children's rights <i>You must know what my my rights are and uphold them</i>				*		***	
Prepare children for present and future developments in AI <i>If I am well prepared now, I can contribute to responsible AI for the future</i>	*			*		*	
Create an enabling environment <i>Make it possible for all to contribute to child centered AI</i>	**				**	**	

3 Proposed methodology

From the previous literature, we propose a methodology to adapt existing ethical guidelines to the use of CAs and children and incorporate identified considerations. For that purpose, we identified prioritisation and action points from ALTAI by performing a risk level analysis for every ALTAI item (question), following the metric below:

$$\text{Risk} = \text{Likelihood} \times \text{Impact} \quad (1)$$

In order to obtain these measures (*Likelihood* and *Impact*), we have followed the Delphi method (Linstone and Turoff, 1975), conducting a survey among four experts. Then, once the experts concluded the *Likelihood* and *Impact* of every ALTAI item, we assessed the risk by the matrix approach (Kovačević et al., 2019), identifying critical points.

3.1 Delphi method

To follow the Delphi method, we designed a questionnaire that asks experts to rate, for each ALTAI question, how relevant it is for children vs general population and for CAs vs AI systems in general. We use two main criteria: the likelihood or frequency of application (i.e. if the question would apply to all situations or only to certain) and the impact or relevance, using a 3 point likert scale for simplicity, given that the questionnaire has 69 questions, with 4 ratings per question (Table 3). In addition, we provided some space for experts to comment on their specificities and reasoning behind the rating when needed.

Table 3 Example of the questionnaire

ALTAI question	Dose it apply to children? (Likelihood)	Is it relevant for children? (Impact)	Does it apply to CAs? (Likelihood)	Is it relevant for CAs? (Impact)	Notes
“Do you communicate to users that they are interacting with an AI system instead of a human?”	Always Sometimes Never	High Medium Low	Always Sometimes Never	High Medium Low	

We explained the questionnaire to experts from different disciplines and areas (AI ethics, CAs, education) who independently filled it using their own criteria. Individual answers were then analysed in order to identify disagreements (the mean of similar answers was 74%) and critical points. An expert meeting was later organised to discuss the identified critical points and disagreements and arrive at a common criteria and consensus. After the meeting, experts had the chance to review and refine their individual responses and submit their final version (84% similar answers).

3.2 Risk assessment

For each of the ALTAI items (i.e. questions) we computed partial risk levels for children and CAs, as follows:

Risk value for ALTAI questions. For each item of ALTAI, we performed the arithmetic mean to combine the four reported *Likelihood* and *Impact* measures from the experts related to children and CAs in a separate way. Later on, we built the *Child Risk* and the *CA Risk* (called partial risks from now on) by using Formula (1) with their respective *Likelihood* and *Impact* means. In addition, a risk assessment matrix was built to measure the level of risk of our results (Fig.2.a).

Risk value for HLEG ALTAI requirements. We computed the arithmetic mean for every question inside a given requirement (e.g. *Human Agency and Oversight* requirement is composed of eleven questions), to calculate the *Likelihood* and *Impact* of the requirement. We did it separately to calculate the partial risks for children and CAs. Later on, the *Child Risk* and the *CA Risk* of a particular requirement was calculated following Formula (1).

From individual partial risks (*Child Risk* and *CA Risk*) per question and requirement, we calculate the *Total Risk* of every question and requirement by the following formula:

$$\text{Total Risk} = \text{Child Risk} \times \text{CA Risk} \quad (2)$$

Finally, we calculate a risk assessment matrix for the *Total Risk* (Figure 2.b) in order to understand the severity of the risk levels. Detailed results can be found in <https://github.com/mescpla/CAs4Children-ETHICOMP22.git>.

3.3 Risk assessment

In order to complement the quantitative assessments, we combined all notes provided by the experts in the questionnaire and the critical points discussed during the Delphi meeting. We carried out a thematic analysis (Braun & Clarke, 2006). First, we compiled all the annotated comments, identifying initial ideas. Secondly, we

grouped the ideas in possible themes that we discussed and refined until our final version. Finally, we selected the examples that we would use in the report.

4 Results

4.1 Ordered assessment list

Table 4 shows the ratings for likelihood, impact and risk for children and CAs dimensions as well as total values. We first observe that, in general, the obtained *Impact* values are higher for children than for CAs. However, the *Likelihood* is much larger for CAs than for children. This explains why the *CA Risk* is higher than *Children Risk* for every category but for: *Human oversight*, *Accuracy*, *Explainability*, and *Environmental well-being* (Table 4).

Table 4 Values for Likelihood, Impact for calculating the Child Risk, the CA Risk and the Total Risk, per each HLEG ALTAI requirement and sub-requirement. Colours indicate the risk level using the colour scheme presented in Fig.2.

	Children					CAs					Total Risk
	Likelihood		Impact		Child Risk	Likelihood		Impact		CA Risk	
	Mean	SD	Mean	SD		Mean	SD	Mean	SD		
HUMAN AGENCY AND OVERSIGHT	2.00	0.45	2.61	0.05	5.21	2.33	0.16	2.52	0.68	5.87	30.59
Human Agency	2.15	0.50	3.00	0.00	6.45	2.87	0.23	2.93	0.94	8.41	54.24
Human Oversight	1.80	0.40	2.13	0.12	3.84	1.80	0.00	2.00	0.40	3.60	13.82
TECHNICAL ROBUSTNESS AND SAFETY	1.54	0.46	2.19	0.16	3.38	2.03	0.33	2.02	0.45	4.10	13.85
Resilience to Attack and Security	1.38	0.58	2.06	0.19	2.83	2.11	0.19	2.06	0.48	4.34	12.27
General Safety	1.40	0.20	2.13	0.23	2.99	1.87	0.23	2.00	0.41	3.73	11.15
Accuracy	1.90	0.36	2.40	0.23	4.56	2.13	0.35	1.93	0.46	4.12	18.81
Reliability, Fall-back plans and Reproducibility	1.53	0.68	2.20	0.00	3.37	2.00	0.58	2.07	0.46	4.13	13.94
PRIVACY AND DATA GOVERNANCE	1.83	0.76	2.72	0.38	4.99	2.67	0.48	2.61	0.78	6.96	34.75
Privacy	1.88	0.89	2.67	0.58	5.00	2.83	0.29	2.67	0.84	7.56	37.78
Data Governance	1.81	0.69	2.75	0.29	4.98	2.58	0.58	2.58	0.74	6.67	33.26
TRANSPARENCY	1.90	0.40	2.40	0.23	4.56	2.27	0.35	2.27	0.59	5.14	23.43
Traceability	1.00	0.00	2.00	0.00	2.00	2.67	0.58	2.00	0.59	5.33	10.67
Explainability	2.00	0.50	2.50	0.58	5.00	2.00	0.00	2.17	0.48	4.33	21.67
Communication	2.25	0.50	2.50	0.00	5.63	2.33	0.58	2.50	0.69	5.83	32.81
DIVERSITY, NON-DISCRIMINATION AND FAIRNESS	1.75	0.44	2.37	0.40	4.14	2.63	0.46	2.17	0.64	5.71	23.63
Avoidance of Unfair Bias	2.00	0.53	2.53	0.46	5.07	2.80	0.35	2.33	0.73	6.53	33.10
Accessibility and Universal Design	1.31	0.33	2.17	0.29	2.84	2.50	0.58	1.92	0.53	4.79	13.63
Stakeholder Participation	2.25	0.50	2.33	0.58	5.25	2.33	0.58	2.33	0.60	5.44	28.58
SOCIETAL AND ENVIRONMENTAL WELL-BEING	1.27	0.53	2.04	0.89	2.59	1.50	0.94	2.00	0.33	3.00	7.78
Environmental Well-being	1.00	0.00	3.00	0.00	3.00	1.67	1.53	1.33	0.25	2.22	6.67
Impact on Work and Skills	1.43	0.84	1.47	1.42	2.10	1.33	0.69	2.20	0.33	2.93	6.17
Impact on Society at large or Democracy	1.00	0.00	3.00	0.00	3.00	2.00	1.00	2.33	0.52	4.67	14.00
ACCOUNTABILITY	1.22	0.34	2.54	0.51	3.10	2.42	0.83	2.04	0.55	4.93	15.28
Auditability	1.00	0.00	2.17	0.29	2.17	3.00	0.00	2.00	0.67	6.00	13.00
Risk Management	1.29	0.45	2.67	0.58	3.44	2.22	1.10	2.06	0.51	4.57	15.73

Considering the partial risk assessment matrix (Fig.2.a), our experts have identified *Human agency and oversight*, *Privacy and data governance* and *Transparency* as the main critical points for children, while *Privacy and data governance*, *Human agency and oversight*, and *Diversity, non-discrimination and fairness* are the main critical requirements for CAs. Regarding the *Total Risk*, and considering the combined risk

assessment matrix (Fig.2.b), we identify *Privacy and data governance* and *Human agency and oversight* (with a critical point on *Human agency*) as the main critical requirements to be considered when developing CAs for children. The only requirement which values are over the matrix diagonal is *Societal and environmental well-being*, with the lowest partial and combined risk levels scores.

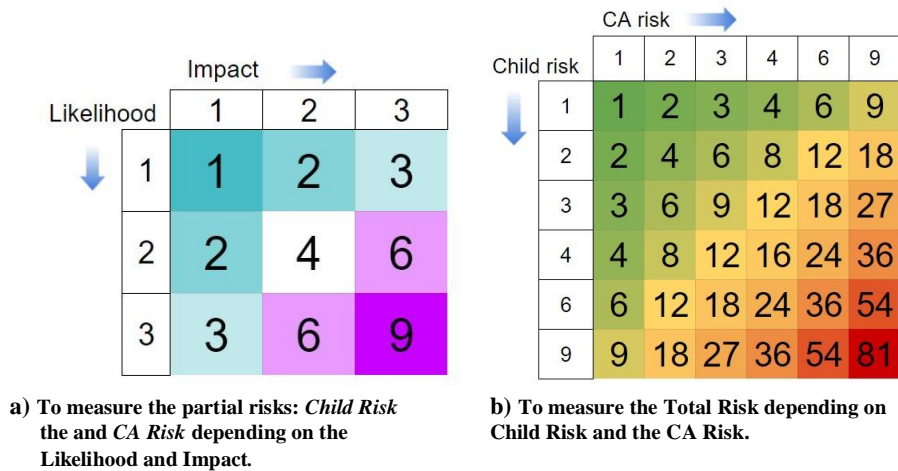


Figure 2 Risk assessment matrices.

4.2 Thematic Analysis

During the Delphi meeting, our experts (R1, R2, R3 and R4) discussed some relevant topics, and other critical considerations were pointed out as questionnaire notes (Table 3). As mentioned in the methodology (Section 3), we performed a thematic analysis and identified critical considerations for the ethical design of child-centric CAs. These considerations are visually presented in Figure 3 and Table 5, and summarised as follows:

Involve children stakeholders. Many of the experts commented on the relevance of involving children stakeholders (children, teachers, parents ...) in the design, use, and test of the system. R1 wrote: “*Include children, tutors and teachers as stakeholders*”, R2 mentioned: “*Multiple stakeholders need to be involved*”. Furthermore, it was mentioned the need to “*teach stakeholders*” (R2), so they can help to oversee the system. In addition the experts agreed that -as children must not work- their collaboration in the design, use, and test of the system needs to be done in a meaningful and entertaining way. As R4 said: “*We need to involve children in the*

design, but in a meaningful way as this participation should be far from job conditions”.

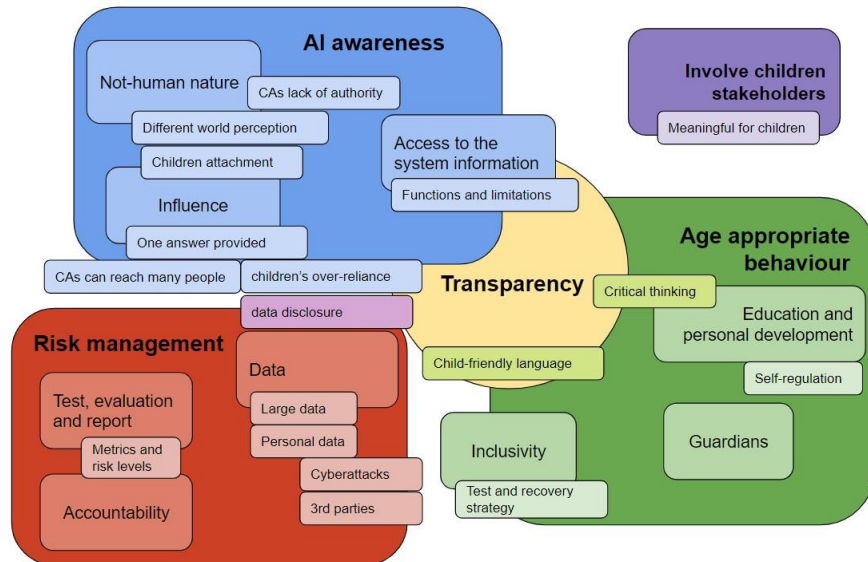


Figure 3 Annotation scheme of experts' comments.

AI awareness. The experts expressed many considerations regarding the fact that people -particularly children- should be aware of what a CA is, how it works and its limitations. In particular:

Not-human nature. The experts remarked that CAs communicate in a natural way, which can lead to confusion about their nature (*"Naturalness of CAs might create confusion"* (R2)). Therefore, there is a risk of developing attachment to the device, especially for children developing their social abilities and with a particular way to understand the world. R4 wrote *"Careful with human attachment as children are developing their cognitive and emotional abilities"*, R1 also shared: *"Children can think that something that is not alive has alive characteristics such as feelings"*. In addition, the social role of the device should be regarded, as mentioned by R4: *"Children might understand what the system is and that has no authority"*. The experts also discussed the positive points of being consistent with the information provided by the CA *"If we want a child to understand that a CA has no feelings, maybe it is better to avoid sentences such as 'I am happy' "*.

Influence. Another critical consideration is the influence a system might have on the user. Firstly, R3 remarked *"CAs can provide information critical to make decisions. People don't usually double check information"*. Additionally, R4 shared concerns about children's critical thinking *"Children*

need to learn to be critical, to look further from the provided information, to develop creativity". Secondly, regarding CA's influence, the experts expressed their concerns about children's vulnerability: *"Check overreliance. Consider children's vulnerability"* (R2); *"Special attention to kids over-reliance"* (R3). Children are highly impressionable -even more provided that we consider the previous point- and their judgement and awareness, on things such as data disclosure, are still in development. Following this, R4 pointed out: *"Kids don't yet have good judgement, careful with influence and over-reliance"*. Last but not least, R2 highlighted the relevance of checking a CA's influence, as their impact is magnified by the number of people able to use the device: *"CAs can influence a large number of people, so the social impact of CAs should be considered"*.

Access to the system information. From a general point of view, all our experts advised developers to be transparent about the nature, functioning and limitations of the system. This information promotes AI awareness, and can minimise some of the above mentioned risks. R3 also commented: *"Understanding CAs nature and interactions is important to avoid frustration"*. In addition, during the experts' meeting, they discussed that this information should be always accessible to the user, and provided in a proactive manner by the system in accordance with the duration and risks of the interaction. R1 mentioned: *"Regarding explainability it should also fit the purpose of the CA. It doesn't make sense to explain all the limitations of a system that will interact with you for a few seconds (e.g. to ask you on the phone what department you want to contact)." R3 added that this would be necessary in bank transactions or high risk interactions.*

Transparency (e.g. informing on how an AI system works, and what are its limitations) is a transversal topic mentioned in different requirements. It is considered as a tool to fight different risks brought by AI, by raising AI awareness and enhancing critical thinking. The access to information about the system (i.e. working principles and limitations), is then essential. Two clear examples of the use of information to fight data disclosure are the *"relevance to inform about recordings"* (R3); and being *"important to know if the system learns about them or not"* (R4). Moreover, all the experts highlighted the need to communicate with children in a child-friendly language (*"Questionnaires or explanations must be adapted to children"* (R1)).

Risk Management. Regarding the novelty of CAs in people's lives, it is paramount to detect potential risks and ensure the accountability of the developed systems. It is also essential to understand which personal data a CA is using and storing, as related to the privacy requirement. Some identified critical aspects of risk management:

Test, evaluation and reporting. Many CAs tend to have a low accuracy with small children and other vulnerable groups. In addition, risks for children require continuous vigilance. Defining metrics and levels to handle

risks can help with this problem; but, in addition, the experts suggested some strategies to minimise these risks. On the one hand, developers can evaluate the system in a controlled environment, involving children and other minorities (*“Relevance of testing and detect if a CA is having problems with children”* (R3); *“Test the system for different children considering vulnerable groups”* (R4)). On the other hand, developers can provide users including children- with mechanisms to report issues that appear after extensive use in the real world scenarios (*“Flag issues in a child-friendly way”* (R4); *“Children should be able to report issues”* (R2)).

Accountability. The experts shared some concerns about the vulnerability of children, and the relevance of people taking responsibility for the system. For instance, R2 wrote: *“Children are a vulnerable population in developmental stage”*, and R3 shared: *“As CAs use biometrical data, and children are a vulnerable population, a special attention to accountability is needed”*. Audits may help to keep track of this problem.

Data. Regarding data, the experts raised some concerns. First of all, they emphasise on the recording of biometric data by CAs (*“CAs data storage contain biometrics and personal data”* (R2); *“CAs use biometric data”* (R3); *“Voice is personal data”* (R4)) and children’s vulnerability (*“Children have the right to be protected”* (R1); *“Extra protection for vulnerability, careful with children overtrust”* (R4)). Therefore, not only should we make an extra effort for personal data protection, (compliance with European regulations such as *“GDPR (General Data Protection Regulation)”* (R4)), but we should also pay a special attention, if the data is shared or transferred to 3rd parties (*“Careful about selling data to 3rd parties”* (R4)), and protect the system from cyberattacks (*“CAs have critical data and must be protected from cyberattacks”* (R3)). Secondly, the consulted experts highlighted that some CAs use information from large data sources (e.g. CAs can search for information on the internet, or can use large datasets for training using deep learning). Among the risks of using these data sources, they mentioned the lack of control of possible risks that can later appear in the system (*“CAs with untrusted data sources may cause more damage”* (R3)). For instance, the experts mentioned in the meeting that, last summer, a game displayed on a CA that took information from the web, told a child to put a coin on a connected plug. The experts recommend a better risk management for those systems.

Age appropriate behaviour. During the study our experts also identified the need to recognise children, and act appropriately, i.e specific considerations/behaviours when interacting with children:

Inclusivity. CAs bring opportunities for inclusivity of illiterate people, or people with disabilities (*“CAs help with inclusivity and should take special*

care on this point, special attention to disabilities” (R3)). That is why developers should emphasise on the inclusivity of the device, setting an example to children (“Children can internalise bias, so it is important for them” (R3); “Children inclusivity for all culture, language, age, ..” (R2)). Nonetheless, these systems face some challenges understanding little children and underrepresented groups (“Biases linked to not available data, which is a challenge for all EU languages, dialects and children” (R4); “Consider limitations of CAs understanding different people” (R1)). Consequently, our experts remarked the importance of recognizing children as users, and to keep working against this bias (“Special attention to bias towards children” (R1); “Discrimination by age”, (R2)). In the meantime, a good strategy to fix conversations when the system cannot identify the user might help to mitigate this risk (“Important to use a good recovery strategy” (R1)).

Guardians. Children have a particular autonomy, as they need the supervision of their guardians. This should be taken into consideration when designing a device that can interact with a child. The experts highlighted their presence in different points. They advise to consider them in order to meet consent obligations (“Need of tutor consent” (R4); “Tutors and children must give their consent” (R1)), but also to take advantage of their presence during interactions (“Rely on adult supervision when low confidence” (R2); “Children are not aware, so an adult should supervise” (R3)). During the experts’ meeting, they discussed the supervision of guardians, identifying their presence, but bearing in mind that the system should be safe enough for the child to not require constant supervision. They agreed that the system can try to use them/call them in specific moments, although the security of the system cannot be just based on the guardians’ supervision.

Educational and personal development. All the experts missed a section on education and children development. For instance R1 commented: “We need to consider CAs in education”, R3 mentioned: “Need for educational consideration”, and R2 wrote: “We should consider adding to ALTAI education and development questions”, referring not just to school education, but also to personal development such as self-regulation (“Consider children addictive behaviour”, R3).

Table 5 Thematic analysis mapped to the requirements of HLEG ALTAI.

	Agency	Robust.	Data	Transp.	Diversity	Well-being	Account
Involve children stakeholders		x		x	x	x	
AI awareness	x	x	x	x		x	
Not-human nature	x			x		x	
Influence	x					x	
Access to the system information	x		x	x		x	
Transparency	x			x		x	
Risk management		x	x		x		x
Test and report		x			x		
Accountability							x
Data		x	x				
Age appropriate behaviour	x	x	x	x	x	x	
Education and personal development	x					x	
Guardians	x	x	x	x		x	
Inclusivity		x			x		

The experts' recommendations cover critical points for all the seven requirements from HLEG ALTAI (Table 5). Being *Societal and environmental well-being* the requirement with more critical themes pointed out by the experts, followed by *Human agency and oversight* and *Technical robustness and safety*. Experts also covered all the learned lessons from the literature review.

5 Discussion and recommendations for child-centric CA developers

From the results presented in Section 4, we recommend developers to consider the ALTAI assessment list, in the **following order of priority** (Fig.3): *Privacy and data governance*, *Human agency and oversight*, *Diversity, non-discrimination and fairness*, *Transparency*, *Accountability*, *Technical robustness and safety*, and *Societal and environmental well-being*. In addition, developers should pay special attention to the considerations outlined by the experts (Fig.6): *Involve children stakeholders*, *AI awareness*, *Transparency*, *Risk management* and *Age appropriate behaviour*. These recommendations will help developers to maximise CAs opportunities for children while minimising risks, creating more accessible CAs, supporting educational activities, social behaviour and safety.

Another interesting conclusion from our work is the identification of a subsection that could enrich current ALTAI guidelines for children: an **“education and self-development”** set of questions in the *Societal and environmental well-being* requirement.

Furthermore, in our risk assessment analysis, the **main critical point detected is *Privacy and data governance***. This point was also covered by our experts' critical considerations in the identified topics of *Risk management* and *Age appropriate behaviour* where they highlighted the presence of children's guardians. These considerations are aligned with previous studies (von Struensee, 2021). Nevertheless,

while the use of data protection regulations is well established, we found little research on the application of data privacy regulations considering AI, children autonomy, and guardians. Therefore, we recommend integrating research outcomes from existing medical studies that use biometric data from children (Hopf et al., 2014).

Besides, our results bring special attention to *Human agency and oversight*, with a special focus on the **not-human nature of the system**. This was also reflected on the thematic analysis in AI awareness using *Transparency* as a tool. These recommendations are in accordance with existing work (Straten et al., 2020).

We also identified some **limitations** of our study. Firstly, our study comes from an european-centric perspective, e.g. focus on Ethical guidelines for trustworthy AI and involving an EU expert group. Therefore, these results cannot then be generalised to different cultures such as Asian or African. Henceforth, we suggest complementary research in larger and more diverse groups with different cultural backgrounds.

Secondly, regarding our metrics, *Children Risk* was generally lower than *CAs Risk*, mainly because of a high rated CAs *Likelihood*. We recognize that our metrics put at the same level children and CAs considerations, when it might be more adequate to highlight children's considerations, e.g. through weights. We encourage complementary studies with alternative metrics.

Finally, HLEG ALTAI *Societal and environmental well-being* requirement had the lowest risk level - surprisingly low considering it is also a fundamental requirement on other studies on AI and children's rights (Charisi, et al., 2022). This might be due to the inclusion of a work impact section (with a low *Likelihood* and *Impact* on children), and the lack of an *Education and self-development* section. This new - and needed- section would change our results of the risk assessment. We encourage further research to build new items for ALTAI on this topic, and suggest the use of LifeComp competences for its development (Sala, et al., 2020): personal (selfregulation, flexibility, well-being), social (empathy, communication, collaboration), and learning to learn (growth mindset, critical thinking, managing learning).

6 Conclusions

We have performed a literature review on conversational agents (CAs), identifying opportunities, challenges and risks for their use with children. In addition, we have consulted a group of experts to measure the risk of all the items of the assessment list for trustworthy Artificial Intelligence (ALTAI) with a focus on children and CAs. With our results, we adapt ALTAI for this specific use, defining priorities on the requirements and adding additional considerations. We hope this research can help CA developers to build trustworthy child-centric systems that can respect fundamental and children rights, ensuring a future where children can also take the most from CAs. We may have safer and better-informed citizens with critical thinking in the future.

Acknowledgements

We thank our experts for their patience and dedication to our experiment, particularly Riina Vuorikari. We thank Isabelle Hupont-Torres and Marta Rivera for their guidance regarding the analysis and assessment of risk. Finally, we want to thank the HUMAINT team for their constant support and useful comments to the text.

References

- AI HLEG (2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment, *European Commission, Brussels*. <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2), 77-101. <https://doi.org/10.1191%2F1478088706qp063oa>
- Catania, F., Crovari, P., Beccaluva, E., De Luca, G., Colombo, E., Bombaci, N., & Garzotto, F. (2021). Boris: a Spoken *Conversational Agent* for Music Production for People with Motor Disabilities. *14th Biannual Conference of the Italian SIGCHI Chapter* (pp. 1-5). <https://doi.org/10.1145/3464385.3464713>
- Charisi, V., Chaudron, S., Di Gioia, R., Vuorikari, R., Escobar-Planas, M., Sanchez Martin, J.I. and Gomez Gutierrez, E. (2022). *Artificial Intelligence and the Rights of the Child: Towards an Integrated Agenda for Research and Policy* (No. JRC127564). Joint Research Centre (Seville site). <http://dx.doi.org/10.2760/012329>
- Cheng, Y., Yen, K., Chen, Y., Chen, S., & Hiniker, A. (2018). Why doesn't it work? voicedriven interfaces and young children's communication repair strategies. *17th ACM Conference on Interaction Design and Children* (pp. 337-348). <https://doi.org/10.1145/3202185.3202749>
- Dignum, V., Penagos, M., Pigmans, K. & Vosloo, S. (2021). Policy guidance on AI for children. Communications of UNICEF. <https://www.unicef.org/globalinsight/reports/policy-guidance-ai-children>
- Druga, S., Williams, R., Breazeal, C., & Resnick, M. (2017). "Hey Google is it ok if I eat you?" Initial explorations in child-agent interaction. *2017 conference on interaction design and children* (pp. 595-600). <https://doi.org/10.1145/3078072.3084330>
- Fraser, J., Papaioannou, I., & Lemon, O. (2018). Spoken conversational ai in video games: Emotional dialogue management increases user engagement. *18th International Conference on Intelligent Virtual Agents* (pp. 179-184).
- Hopf, Y. M., Bond, C. B., Francis, J. J., Haughney, J., & Helms, P. J. (2014). Linked health data for pharmacovigilance in children: perceived legal and ethical issues for stakeholders and data guardians. *BMJ open*, 4(2), e003875. <https://doi.org/10.1136/bmjopen-2013-003875>
- Johns, T. (2021, July 2). *Parents of children called Alexa challenge Amazon*. BBC. <https://www.bbc.com/news/technology-57680173>

- Kahn Jr, P. H., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., Ruckert, J. H., Shen, S. (2012). "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Developmental psychology*, 48(2), 303.
- Kelly, S.M. (2018, October 17). *Growing up with Alexa: A child's relationship with Amazon's voice assistant*. CNN Business. <https://edition.cnn.com/2018/10/16/tech/alexa-child-development/index.html>
- Kovačević, N., Stojiljković, A., & Kovač, M. (2019). Application of the matrix approach in risk assessment. *Operational Research in Engineering Sciences: Theory and Applications*, 2(3), 55-64. <https://doi.org/10.31181/oresta1903055k>
- Landoni, M., Murgia, E., Huibers, T., & Pera, M. S. (2020). You've Got a Friend in Me: Children and Search Agents. *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 89-94).
- Lavechin M, Bousbib R, Bredin H, Dupoux E, Cristia A (2020), An open-source voice type classifier for child-centered daylong recordings. *arXiv:2005.12656*.
- Law, E., Baghaei Ravari, P., Chhibber, N., Kulic, D., Lin, S., Pantasdo, K. D., Ceha, J., Suh, S., & Dillen, N. (2020). Curiosity Notebook: A Platform for Learning by Teaching Conversational Agents. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-9). <https://doi.org/10.1145/3334480.3382783>
- Linstone, H. A., & Turoff, M. (Eds.). (1975). The delphi method (pp. 3-12). *Reading, MA: Addison-Wesley*.
<https://www.researchgate.net/file.PostFileLoader.html?id=563b341d5cd9e375988b45bc&assetKey=AS%3A292381292285964%401446720541026>
- McTear, M. (2020). Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots. *Synthesis Lectures on Human Language Technologies*, 13(3), 1-251. <https://doi.org/10.2200/S01060ED1V01Y202010HLT048>
- Nasihati Gilani, S., Traum, D., Merla, A., Hee, E., Walker, Z., Manini, B., Gallagher, G., & Petitto, L. A. (2018). Multimodal dialogue management for multiparty interaction with infants. *20th ACM International Conference on Multimodal Interaction* (pp. 5-13).
- Pradhan, A., Mehta, K., & Findlater, L. (2018). "Accessibility Came by Accident" Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. *2018 CHI Conference on human factors in computing systems* (pp. 1-13). <https://doi.org/10.1145/3173574.3174033>
- Sala, A., Punie, Y., Garkov, V., & Cabrera, M. (2020). LifeComp: The European framework for personal, social and learning to learn key competence (No. JRC120911). *Joint Research Centre (Seville site)*. <http://dx.doi.org/10.2760/302967>
- Sciuto, A., Saini, A., Forlizzi, J., & Hong, J. I. (2018). "Hey Alexa, What's Up?" A MixedMethods Studies of In-Home Conversational Agent Usage. *2018 Designing Interactive Systems Conference* (pp. 857-868). <http://doi.org/10.1145/3196709.3196772>
- Sezgin, E., Noritz, G., Elek, A., Conkol, K., Rust, S., Bailey, M., Strouse, R., Chandawarkar,

- A., Sadovszky, V., Lin, S., & Huang, Y. (2020). Capturing at-home health and care information for children with medical complexity using voice interactive technologies: multi-stakeholder viewpoint. *Journal of medical Internet research*, 22(2), e14202. <https://doi.org/10.2196/14202>
- Straten, C. L. V., Peter, J., Kühne, R., & Barco, A. (2020). Transparency about a robot's lack of human psychological capacities: effects on child-robot perception and relationship formation. *ACM Transactions on Human-Robot Interaction (THRI)*, 9(2), 1-22. <https://doi.org/10.1145/3365668>
- von Struensee, S. (2021). Eye on Developments in Artificial Intelligence and Children's Rights: Artificial Intelligence in Education (AIEd), EdTech, Surveillance, and Harmful Content. *EdTech, Surveillance, and Harmful Content (June 4, 2021)*. SSRN: <https://ssrn.com/abstract=3882296> or <http://dx.doi.org/10.2139/ssrn.3882296>
- Xu, Y., & Warschauer, M. (2020). "Elinor Is Talking to Me on the Screen!" Integrating Conversational Agents into Children's Television Programming. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-8).
- Zhang, L., Weitlauf, A. S., Amat, A. Z., Swanson, A., Warren, Z. E., & Sarkar, N. (2020). Assessing social communication and collaboration in autism spectrum disorder using intelligent collaborative virtual environments. *Journal of autism and developmental disorders*, 50(1), 199211. <https://doi.org/10.1007/s10803-019-04246-z>

Emerging values in ICT during uncertain times: the case of COVID-19

Ana Saraiva¹[0000-0002-6989-5274], Luís Valadares Tavares²[0000-0001-5727-1264], Nuno Silva³[0000-0003-0157-0710]

^{1,2,3} Universidade Lusíada, COMEGI (Portugal)

11097316@lis.ulusiada.pt

Abstract. In this paper, the authors intend to study the change of values and ethics considering the five major dimensions of the relations between consumers and ICT within the general framework of Digital Society: philosophical dimension (1), technological dimension (2), educational dimension (3), consumption dimension (4), and health preservation (5). The main goal is to apply the study of the impacts of the COVID-19 pandemic on the values and ethics underlying the behavior of ICT consumers.

The authors were intended to identify major societal problems and propose general goals to overcome such situations (1), design plausible and alternative scenarios including the most critical exogenous variables (2), study the dynamics of technological knowledge in terms of production and import-export balances (3), assess the impact of technology areas to meet general goals within the framework of designed scenarios (4), propose strategic lines of development using the contribution of technological knowledge for societal development (5), to improve the governance by the use technical knowledge.

Three research questions intend to demonstrate how this pandemic is reshaping the values and ethics of society are: (1) which values have been changed and their relative importance in pandemic times for using ICTs? (2) how are the major patterns of ethics in ICT consumption due to the COVID-19 pandemic? and finally (3) how these changes can be perceived in the crucial area of e-learning. Concluding, a new matrix relating problems with priorities in values of ICT is discussed.

Keywords: ICT, Values, COVID-19, E-Learning, Distance Learning

1 Introduction

The recent ubiquitous development of ICT in human history is changing the structure of values and the ethics of consumers. The unprecedented challenges due to COVID-19 are a crucial cause of accelerating such a process of change involving the

psychosocial impacts of this pandemic (Lee, 2021). Big Data, machine learning, and artificial intelligence are becoming implemented (Zharinov, 2020) many times by invisible touch. But according to Silva (2022), it is essential to "mise en valeur" the educational technology, reshape training and business marketing actions, and problem-solving in value generation (level of adoption).

Technology may not be an impediment, but the interventions in public environments are affecting needs and values. The change of values and ethics will be studied using the basic model to problem-oriented and processes of change presented by Tavares (2003), and it will consider five major dimensions of the relations between consumers and ICT within the general framework of Digital Society:

- 1) Philosophical;
- 2) Technological;
- 3) Educational;
- 4) Consumption;
- 5) Health preservation.

The first dimension includes all the significant principles and values underlying the relations between consumer and ICT. The second dimension covers the broad spectrum of interactions due to the diversity and complexity of modern ICT. The third dimension focus on the changes in knowledge transfer about ICT. Furthermore, the fourth dimension concerns the principles and criteria governing the choice and acquisition of ICT. Finally, the last dimension is covering the urgent health problem or the form of a healthy lifestyle using ICT.

This framework will be applied to study the impacts of COVID-19 on the values and ethics underlying the behavior of ICT consumers under the challenges of this pandemic.

The methodology is focused on:

- 1) Identify major societal problems and propose general goals to overcome such problems;
- 2) Design plausible and alternative scenarios, including essential exogenous variables;
- 3) Study the dynamics of technological knowledge in terms of production and import-export balances;
- 4) Assess the impact of technology areas to meet general goals within the framework of designed scenarios;

Propose strategic lines of development using the contribution of technological knowledge for societal development.

The methodology assumes the practicality of collecting opinions and sponsoring debates within interdisciplinary groups concerned with society, development, economics, science, and technology.

The presented method has included the following steps:

- 1) Selection of cases to be used for benchmarking purposes by a systematic literature review;
- 2) Estimation of a set of indicators covering main societal issues;
- 3) Benchmarking and scoreboard of recent trends;

- 4) Identification of significant problems;
- 5) Construction of a matrix relating problems with priorities for public policies.

Moreover, the methodology (GOVSIGHT) can be applied to improving governance by using technological knowledge (Tavares, 2003).

Initially, the authors carry out a bibliographical review based on the best journals of the 2021 year in the field of "Operations and Technology Management," present in the Academic Journal Guide ranking using the keywords of the paper – ICT, Technology, Values, Ethics, COVID-19 Pandemic, Distance Learning. Then, the process of change of values and ethics is studied for the particular case of Distance Learning as this area of ICT applications was particularly important to cope with COVID-19. Such research was carried out using semi-structured interviews, and their results highlight the change process and the impacts of COVID-19 on ICT values and ethics.

The scope of this paper is to demonstrate how this pandemic is reshaping the values and ethics of society, and the obtained answer to the following three research questions:

- 1) Which values have been changed and their relative importance in pandemic times for using ICTs?
- 2) What are the major patterns of ethics in ICT consumption due to the COVID-19 pandemic?
- 3) How can these changes be perceived in the vital area of e-learning?

The answers to these questions can help to understand the change of values and ethics of our society in the crucial area of ICT consumption giving special attention to the critical case of e-learning.

2 Structure

2.1 A systematic literature review

To select the cases to be used for benchmarking purposes and to identify the significant problems, the authors conducted a systematic literature review charted by the PRISMA framework (Preferred Reporting Items for Systematic reviews and Meta-Analyses) and selected databases (e.g., B-ON Collections (<https://www.b-on.pt/en/collections/#Contents>)).

This research method intends to summarize this study's numerous researches to identify critical theoretical and methodological frameworks or highlight the gaps in the literature or questions to be addressed (Saraiva & Silva, 2021).

The authors started to state some research procedures:

- 1) Chosen Database: B-ON (the online knowledge library);
- 2) Papers with peer-reviewed and full-text only;
- 3) Academic journals only;

- 4) Selected subject areas and disciplines (the authors only choose two areas related to the paper: Education and Information Technology);
- 5) Selected timespan (only from 2021 to 2022);
- 6) Language (only papers written in English).

The authors started to focus on the paper Keywords:

- 1) ICT;
- 2) Values;
- 3) COVID-19;
- 4) Distance Learning;
- 5) E-learning.

The purpose was to eliminate identical results by doing some search equations with keywords, such as ICT and Values; ICT and Covid-19; Values and Covid-19; and so ever. The results are summarized in Table 1.

Table 1. Results of keyword equations on B-ON

Search equation	Number of results
ICT and Values	175
ICT and COVID-19	303
ICT and Distance Learning	75
ICT and E-learning	77
Values and COVID-19	1.400
Values and Distance Learning	352
Values and E-learning	159
COVID-19 and Distance Learning	2.158
COVID-19 and E-learning	714
Distance Learning and E-learning	410
TOTAL	2265

This systematic literature review results have identified a vast database of related papers (with a total of 2265), even with the imposed limitations. As we can see, numerous articles refer to the research, so it is crucial to select and distinguish the articles from approaches (Saraiva & Silva, 2021).

Following the previous selection, the authors selected the papers that had a focus on the five dimensions of this study: Philosophical (1), Technological (2), Educational (3), Consumption (4), and Health preservation (5). Within this review, the authors selected four papers for the first dimension (Philosophical), ten for the second dimension (Technological), three for the third dimension (Educational), four for the fourth dimension (Consumption), and one for the five dimensions (Health preservation).

2.2 Philosophical dimension

Pursuing the selected papers in the systematic literature review of this dimension, the authors noticed a gap in the literature. A few papers regarding the philosophical dimension of ICT had to be checked in references of the selected works to give a perspective of the dimension in this paper.

This dimension will (Sujoy & Rath, 2014):

- 1) Help people process information,
- 2) Improve decision making,
- 3) Reduce scarcity of resources,
- 4) Support relationships among people,
- 5) Help people understand each other.

In the philosophical dimension the of the relations between consumers and ICT within the general framework of Digital Society, the authors found four characterizations of this dimension (Nicholas & Mugeni, 2014) that can apply:

- 1) Philosophy of Computer Science
- 2) Philosophy of Information (PI)
- 3) Philosophy of Technology (PT)
- 4) Philosophy of ICT

In the first characterization, we can approach the philosophy of Computer Science to the Philosophy of science. It can be described as the “empirical study of the phenomena surrounding computers, not just the hardware, but the programmed, living machine” (Newell & Simon, 1976).

This perspective offers a variety of unique ways of explaining phenomena, such as computational models and algorithms (Nicholas & Mugeni, 2014). This perspective approaches the philosophy of science, that “(...) deals with issues such as the foundations of science, its assumptions and limitations, its implications, and what constitutes scientific progress” (Nicholas & Mugeni, 2014).

The second characterization (Philosophy of information (PI)) takes calculation as one of the management processes in which ICT can be engaged. This perspective shows how the ICT and the "technologies of computing" have changed it into a primary phenomenon (Nicholas & Mugeni, 2014). It is a philosophical field that focuses on the theoretical investigation of the conceptual nature and principles of information that involve its dynamics, uses, and sciences (Floridi, 2002).

The third perspective (Philosophy of Technology (PT)) can be defined as the search for technology, the understanding and evaluation of the consequences of this PT in our society and the human state, and how humans should react to this perspective of the technology (Brey, 2010a).

Relating to the previous perspectives (Philosophy of Computer Science and PI), even Heidegger states that technology depends on science, and science depends on technology.

In our last perspective (Philosophy of ICT), which is more related to this study's central theme, the authors describe and analyze the practices and products empirically

informed way and explain the philosophical theories appropriate to technology and engineering (Nicholas & Mugeni, 2014).

This characterization focuses exclusively on a moral assessment of technology and ethical requirements, which assesses the consequences of technology concerning different standards of goodness and badness, rather than focusing only on the morality of goodness or badness.

ICT in a philosophical dimension has emerged in new perspectives of human understanding and knowledge such as global ICT; ICT for development; and even the Internet, raising further philosophical questions, like “what is the difference between technology and communication, what is the relationship between technology, communication, and Internet, what is a knowledge economy and what is the global ethics (Nicholas & Mugeni, 2014).

Finally, Naaj and Nachouki (2021) emphasized the impact of COVID-19 on some factors influencing computer ethics:

- 1) Gender - mixed results have reported the relationship between gender and computer ethics;
- 2) Major – technology majors have better cyber ethics awareness statistically;
- 3) Age - students aged 24 years and above had statistically higher awareness rates of cybercrime;
- 4) Other Factors - religious work values related to computer use ethics attitudes, sustainability, moral degradation, weak law implementation, and lack of awareness of infringement of the law.

2.3 Technological dimension

The design and operation of computer systems have moral consequences and, therefore, should be subjected to ethical analysis (Brey, 2010b). These days, students spend considerable time online with personal technologies, emphasizing new forms of socialization and values. Several tools enable this kind of human relationship with free speech and intellectual freedom, but filtering content is a significant risk to privacy, security, identity, confidentiality, and anonymity. In addition, big data, machine learning, and artificial intelligence are also becoming implemented (Zharinov, 2020), often by invisible touch.

With the imposition of the COVID-19 pandemic time, external demands, and fast internal practice changes needed, the emergency was focused on upgrading existing technologies or the innovative development of technology. For instance, Kuo et al. (2021) reported essential considerations regarding the hardware when using instant communication technology, but significant ethical issues emerged in software: platform choices, security, ICT accounts, interview modes, video/voice recording, and time limitations.

Regardless, the progress of uncertain times looks in search for alternative design elements and makes relevant the importance of adaptability. A platform like Moodle is still used to carry content, while others like Zoom or Teams are moving from interactivity to more digitized repositories. While retaining added values related to standardization, access, privacy, and security, the ethical risk is overcoming political

decisions and values like equality, trust, language precision, or legal documents, which affect how we think and act.

Emerging values illuminate taken-for-granted aspects of technologies, but what means that value-added depends on values. Suppose ideally is a standard of importance that each individual, group, or society give to a determined action or subject (human or non-human). What is the basis of a shared understanding of some primary factors dependent on dominant market suppliers? Economic? Development has influenced a shift from absolute norms and values to increasingly rational, tolerant, trusting, and participatory values.

2.4 Educational dimension

The COVID-19 crisis increased the role of digitization in education, as teachers were pushed to apply, during the shutdown, an "emergency remote education," which differed from planned practices such as distance learning, e-learning, or b-learning, with very mixed outcomes and revealing weaknesses in the system such as the digital divide, inequality or social injustice (Valverde-Berrocso et al., 2021).

However, while technology may not be an impediment, it has significantly increased complexity and uncertainty. Without training, it is impossible to use new educational tools and technologies, but beliefs, norms, values, attitudes, and even politics are equally important. For example, to avoid poor learner perception of e-learning, it is essential the ability to separate e-learning into content, pedagogy, and the complex network of hard and soft technology (Mehta et al., 2019). In addition, ethical behavior is acting in ways consistent with one's values and the commonly held values of the organization and society. Gurr (2020) considers that the pandemic has been a time of crisis in terms of implications for leadership. There has been a need for teachers from all sectors of education to be more collaborative and to take on many leadership roles. Moreover, according to (Valverde-Berrocso et al., 2021), it is necessary to redefine teacher training for the integration of ICT into more contextualized, reflective, and participatory models.

The education sector is about politics and public values and embraces a shared responsibility. For example, when monopolistic tools are dominant from a purely commercial point of view, consultation culture and group work are significant to finding more variation and developments in educational technologies. Disruption always offers new opportunities and jobs that include orientation towards new values. To Li et al. (2020), digital ethics should be supplemented to restrain the moral level.

A study by Valverde-Berrocso et al. (2021), carried out on 251 teachers of public primary and secondary education centers in Extremadura (Spain), which have a Digital Education Plan (primary teachers - 40.6%, and secondary teachers -59.4%), said that the preceding sample uses the classroom as an educational space for the integration of digital technologies, with the highest average frequency of use. The second place is the physical family context (home) which is most frequently used for the educational benefit of ICT. The most commonly used spaces within the academic center are the IT or computer classrooms, followed by the library. In both cases, the average frequency of use is low. In other areas of the schools, such as laboratory

classrooms or specific ones for physical education activities, digital technologies are scarce.

While integrating e-learning in education, tools Zoom or Teams emerged as major solutions adopted for emergency remote teaching. According to Silva (2022), suppliers' news in the face of COVID-19 regularly mentions a generalized offer of these tools. Generosity, empathy, and the sharing of emotions grew and were fundamental for the whole school to function in the mandatory facts. From the point of view of values, one can start with the importance of personal values. Personal values are the things that are important to each of "us," the characteristics and behaviors that motivate us and guide our decisions. These unique values can affect emotional values that can cause anxiety, worry, or stress. Dynamic values have been weakened with e-learning, and it will be necessary to understand how this problem can be circumvented. For example, achievement, learning, reliability, safety, or simplicity are valued more.

The widespread use of digital technology in an epidemic situation not only improves efficiency and promotes innovation but also raises concerns about new issues, such as (Li, 2020):

- 1) Disclosure of personal privacy;
- 2) Information security;
- 3) Network violence;
- 4) Data monopoly.

While laws and regulations often lag behind the emergence of problems, digital ethics must be supplemented to restrict from the moral level; secondly, the risk of getting rid of the real to the virtual.

For instance, the experience of e-learning may have developed individualism and affected the value of tolerance. That is, generate moments of intolerance when returning to face-to-face format and should be necessary to return to group work. However, we have seen an increase in the value of generosity, private and social, which involved the loan of computer equipment between teachers, students, and families. All the deals, from the security of information and contents to the real benefits of using these resources, should be the purpose for which it is intended, to achieve students' capacity for self-study.

Finally, the sense of belongingness to a group should be a higher level of importance, and even personal values can influence the preferences about each available ICT tool. Interconnected through our shared humanity are emerged in the COVID-19 pandemic, as well as an ethics of care that values social and educational networks to enhance students' independent thinking.

2.5 Consumption dimension

An evaluation model of a website is not a neutral instrument because it may be constructed to assess the achieved level of ICT professional skills. In addition, the appropriate value function representing the satisfaction of the population of customers who are supposed to use a specific e-commerce website depends on the type of population to which the website is dedicated (Tavares et al., 2021).

In the consumption dimension, the authors try to assess the impact of technology areas to meet general goals within the framework of designed scenarios and propose strategic lines of development using the contribution of technological knowledge for societal development.

In ICT consumption, it is crucial to talk about e-commerce. It can be defined as using electronic systems in business activities, which means conducting financial transactions over the Internet, placing orders online, transmitting and disseminating data electronically, and using digital marketing (Burinskas & Burinskiene, 2019).

Also, Jędrzejczak et al. (2019) identified this term as a generic definition used to describe the buying and selling process supported by electronic devices and can be characterized by the following categories:

- C2C (consumer-to-consumer commerce, where the consumer participates in the transaction - is found on Marketplace platforms, such as Olx)
- C2B (commerce from consumer to business, where the consumer provides his product or service to the business - this happens, for example, on the Resn website, which is dedicated to doing advertising and multimedia work for Reebok and Adidas)
- B2C (a more traditional form of commerce, from company to consumer, such as Continente, Auchan, among others)
- B2B (business-to-business, e.g., CTT, which distributes its product and services to other organizations, i.e., other distribution and payment services (e.g., Banco CTT))

As we can see in a Portuguese study (Saraiva, 2022), to improve e-commerce, we must focus essentially on the user. It is essential to look at the digital, considering that online is not a competitor of offline, but rather a compliment. The Portuguese consumer is more conservative and age, with more difficulty adapting to new technologies (compared to the benchmarking analyzed) and needs a physical component to adjust to the digital purchase.

It is essential to increase the consumers' confidence, which involves innovation and creativity. It was proved that trust in the transaction (and the cost incurred if this trust is broken) plays a role in the Portuguese purchasing decision, so this feature should be widely improved. Below is a schematic illustration of options for strengthening commerce:

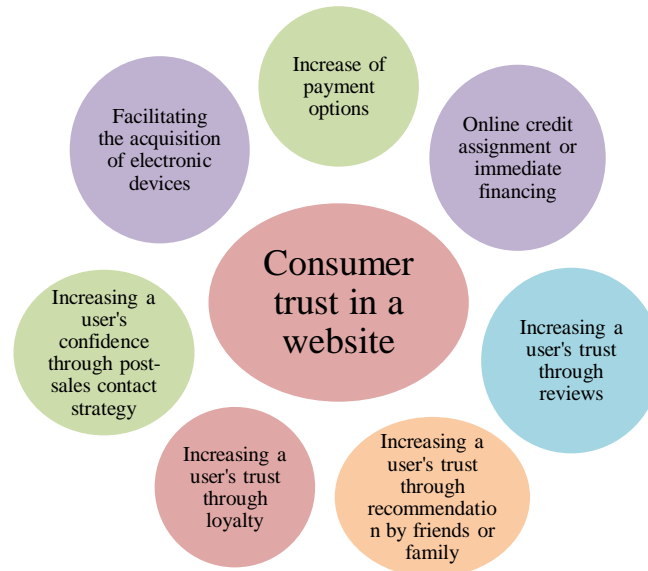


Fig. 2. E-commerce improvement features regarding the security criteria

Now we have a tool to assess the impact of technology areas to meet general goals within the framework of designed scenarios, so the authors can propose strategic lines of development using the contribution of technological knowledge for societal development.

2.6 Health Preservation

This dimension tried to understand the mental health of ICT users or how the confinement caused by this pandemic affected mental disorders that also altered users' and consumers' values.

In this sense, one can question the safety period during the lockdown caused by the covid-19 pandemic.

During this lockdown, people massively adhered to the technologies without much care regarding security and privacy because, in the first place, it was health preservation. These values quickly became extremely important, even with the easing of pandemic lockdown measures or in the face of attacks that have been verified and reported in the media.

Following Tarasenko et al. (2021), there is a need for moral and personal competence regarding the formation of knowledge of children of different ages, skills, skills for a healthy lifestyle, and health preservation.

3 Discussion and conclusion

This section try to discuss the obtained results to three research questions, meanwhile develop a new matrix for ICT values emerged.

- 1) Which values have been changed and their relative importance in pandemic times for using ICTs?

The five dimensions showed the change of the values of the importance in pandemic times for the use of ICTs. In the philosophical dimension, ICT has emerged in new perspectives of human understanding and knowledge, such as global ICT, ICT for development, and even the Internet equity, raising further philosophical questions. In the technological dimension, new platforms (like moodle and zoom, for example) have emerged during the pandemic. We can see moral consequences on the design and operation of computer systems and, therefore, should be subjected to ethical analysis.

In the consumption dimension, we have seen that the most important criteria for the consumers are security on a website; that is why consumer trust and its increment are so important. Security is also a value that has changed in health preservation, facing the importance of the pandemic times for using ICTs.

- 2) What are the major patterns of ethics in ICT consumption due to the COVID-19 pandemic?

Regarding this research question, it was essential to know that hackers take advantage of the available time (resulting from the confinement of the covid-19 pandemic) to exploit security vulnerabilities.

The significant patterns that influence ethics on ICT consumption in the COVID-19 pandemic were gender, education, age, sustainability and other factors like religion, ethics, lack of regulations and law, and morals facing the main issues as the exposure of people's privacy; security of information, network harassment, and data overcome.

Facing this pandemic, ICT seems to meet a change primordially in a way to restrict from the moral level; and facing the risk of getting rid of the real to the virtual. The human factor in the technological dimension is the principal and essential need in studying these patterns.

Several instruments enable this kind of human relationship with free speech and intellectual freedom, but content filtering is a significant risk to privacy, security, identity, confidentiality, and anonymity. In addition, big data, machine learning, and artificial intelligence are also often implemented by invisible touch.

- 3) How can these changes be perceived in the vital area of e-learning?

Education was one of the most significant changes in the digital world, as an emerging value in these uncertain times. The pandemic increased digitalization in education, and there was an "emergency remote education" through distance learning, e-learning, or b-learning. We have seen significant changes when integrating ICT into

education, such as Zoom or Teams tools, which have emerged as great adopted ways for emergency remote learning, and the increased values of generosity, private and social inclusion, which involved the loan of computer equipment between teachers, students, and families.

The capacity of a student is moved to self-study and autonomy, supported by ethics of care that values social and educational networks to enhance students' independent thinking. But in the other way, e-learning may have developed individualism and affected the value of tolerance, revealed weaknesses in the system such as the digital divide, inequality, or social injustice. We have also seen weakness in the dynamic values, as values like achievement, learning, reliability, safety, or simplicity started to value more.

Concluding, after the pandemic and facing the emergence of technological concepts, the values started to change, and suddenly the consumer began to give more priority to different things, such as security online, trust in online navigation, training and moral consequences on online interaction, beyond others. Figure 2 resumes the major conclusions.

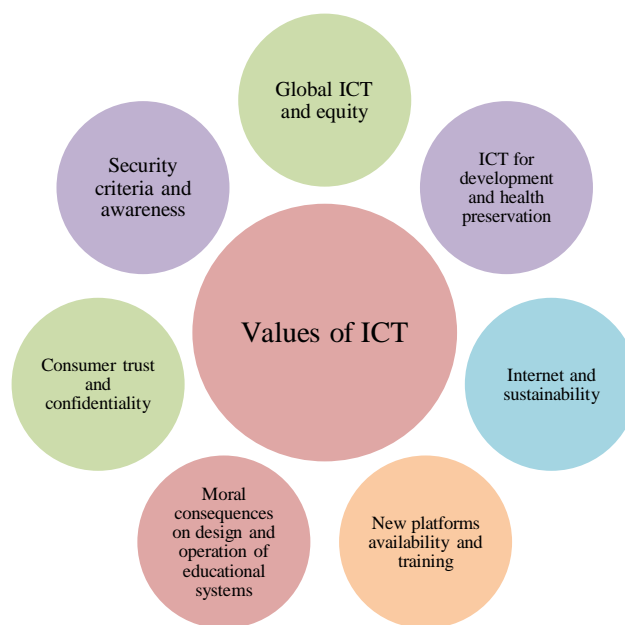


Fig. 2. New matrix relating problems with priorities in values of ICT

Acknowledgement

This work is supported by national funding's of FCT - Fundação para a Ciência e a Tecnologia, I.P., in the project «UIDP/04005/2020»

References

- Brey, P. (2010a). Philosophy of technology after the empirical turn. *Techné: Research in Philosophy and Technology*, 14(1), 36-48. <https://doi.org/10.5840/techné20101416>.
- Brey, P. (2010b). Values in Technology and Disclosive Computer Ethics. In L. Floridi (Ed.), *The Cambridge Handbook of Information and Computer Ethics* (pp. 41-58). Cambridge University Press. ISBN (Print) 97-805-21888-98.
- Burinskas, A., & Burinskiene, A. (2019). Cash-Flow Model for Efficiency Evaluation in Multinational Trade Enterprises. *Applying E-Commerce Engineering Economics*, 30(5), 515-529.
- Jędrzejczak, G. J., Barska, A. & Siničáková, M. (2019). Level of development of e-commerce in EU countries. *Management*, 23(1), 209-224.
- Kuo, Y. S., Lu, C. H., Chiu, P. W., Chang, H. C., Lin, Y. Y., Huang, P. S., Chen, C. J., Lin, C., Tang, S. J., Chang, H. Y., Chang, H. R., & Lin, C. H. (2021). Challenges of Using Instant Communication Technology in the Emergency Department during the COVID-19 Pandemic: A Focus Group Study. *International Journal of Environmental Research Public Health*, 18(12463). <https://doi.org/10.3390/ijerph182312463>.
- Floridi, L. (2002). What is the philosophy of information? *Metaphilosophy*, 33(1-2), 123-145.
- Gurr, D. (2020). Educational Leadership and the Pandemic. *Academia Letters*. Article 29. <https://doi.org/10.20935/AL29>.
- Mehta, A., Morris, N. P., Swinnerton, B., & Homer, M. (2019). The influence of values on e-learning adoption. *Computers & Education*, 141, Article 103617. <https://doi.org/10.1016/j.compedu.2019.103617>.
- Lee, Y.-C., Malcein, L. A., & Kim, S. C. (2021). Information and Communications Technology (ICT) Usage during COVID-19: Motivating Factors and Implications. *International Journal of Environmental. Research and Public Health*, 18(3571). <https://doi.org/10.3390/ijerph18073>.
- Li, Y. (2020). Association Between China's Digital Economy and Labor Education in Post-pandemic COVID-19 Based on Neural Network. *Journal of Intelligent & Fuzzy Systems*, 39(6), 8839-8845. <http://dx.doi.org/10.3233/JIFS-189281>.
- Naaj, M. A., & Nachouki, M. (2021). Evaluating Students' Cyber Ethics Awareness in a Gender-Segregated Environment Under the Impact of COVID-19 Pandemic. *TEM Journal*, 10(3), 1248-1256. <http://dx.doi.org/10.18421/TEM103-31>.
- Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry. *Communications of the ACM*, 19(3), 113-126.
- Nicholas, K., & Mugeni, G. B. (2014). Is it time for the philosophy of ICT? *International Journal of Information and Communication Technology Research*, 4(8), 299-303.

- Saraiva, A. C., & Silva, N. S. (2021). The COVID-19 Impact on Online Education – Opportunities and Challenges in a SWOT Analysis. *International Journal of Research in E-learning*, 7(2), 1–18. <https://doi.org/10.31261/IJREL.2021.7.2.03>.
- Saraiva, A. C. (2022). E-commerce in Portugal: Consumer Modelling and European Benchmarking (Unpublished doctoral dissertation). Lusíada University, Lisbon, Portugal.
- Silva, N. S. (2022). “Mise en Valeur” of E-learning in Basic and Secondary Education: the impact of COVID-19 (Unpublished doctoral dissertation). Lusíada University, Lisbon, Portugal.
- Sujoy, P., & Rath, S. K. (2014). ICT for peace: A philosophical perspective. *International Journal of Education for Peace and Development*, 2(1), 39-45.
- Tarasenko, N., Rasskazova, O., Hryhorenko, V., Shkola, O., Zhamardiy, V., Davydchenko, I., & Shevchenko, N. (2021). Innovative methods and information and communication technologies for training future teachers, educators, and social workers in the context of the COVID-19. *Journal for Educators, Teachers, and Trainers*, 12(3), 21-29. <http://dx.doi.org/10.47750/jett.2021.12.03.003>.
- Tavares, L. V. (2003). Foresight and governance: a problem-oriented methodology (GOVSIGHT). *International Transactions in Operational Research*, 10(2), 169-190. <http://dx.doi.org/10.1111/1475-3995.00402>.
- Tavares, L. V., Arruda, P., Ferreira, J. A., Saraiva, A. C., & Amaral, A. V. (2021). A multicriteria model to evaluate e-commerce websites under the perspective of the customer. *Journal of Internet Banking and Commerce*, 26(8), 1-18. <https://www.icommercecentral.com/open-access/a-multicriteria-model-to-evaluate-e-commerce-websites-under-the-perspective-of-the-customer.php?aid=89878>.
- Valverde-Berrococo, J., Fernández-Sánchez, M. R., Revuelta-Dominguez, F. I., & Sosa-Díaz, M. J. (2021). The educational integration of digital technologies preCovid-19: Lessons for teacher education. *PLoS ONE*, 16(8), e0256283. <https://doi.org/10.1371/journal.pone.0256283>.
- Zharinov, S. (2020). The Role of the Library in the Digital Economy. *Information Technology & Libraries*, 29(4), 1-17. <https://doi.org/10.6017/ital.v39i4.12457>.

Perceived Risks of the Data Economy: Autonomy and the Case of Voice Assistants

Uwe V Riss¹[0000-0003-0123-272X], Edith Maier¹ and Michael Doerk²

¹ Eastern Switzerland University of Applied Sciences, St. Gallen, Switzerland

² Lucerne University of Applied Sciences, Lucerne, Switzerland

uwe.riss@ost.ch

Abstract. The growing power of digital technologies such as artificial intelligence (AI) and digital business raises not only the question of privacy, but also the closely related question of customer autonomy. To understand the depth of the problem we look at the effects possible breaches of privacy may have. We find two principle ways; one is the transfer of confidential information (e.g. secrets), the other is the leakage of seemingly innocuous bits of data. The latter can be aggregated into powerful profiles by AI-based technology currently developed by big tech companies at a breathtaking pace. Whereas users are mostly aware of the first threat, this applies less to the second, which becomes more and more important for building a user profile. This can facilitate attacks on the autonomy of users through manipulation. To give more insight into the problem we present results from an ongoing research project that investigates the impact of voice assistants (VAs) on everyday life. The study reports on attitudes and perceptions of two groups of participants: 1) people who volunteered to use a VAs for four weeks at homes and 2) students who reflected on the use of VAs as part of their course work. We have observed that critical reflection leads to different perceptions compared to the actual use of VAs. We also look at recent legal initiatives and how they try to address the problem. Since the uncontrolled use of data is likely to turn into a serious problem for the data economy in the near future, the problems must be identified and addressed as soon as possible.

Keywords: Voice Assistants, Privacy, Autonomy, Data Aggregation, Artificial Intelligence

1 Introduction

Services and products that make massive use of customer data have been entering into our lives at a breathtaking pace as part of the digital economy; voice assistants (VAs) are just one example. They have initiated an era of data economy, which is largely dominated by big tech companies such as Google, Amazon, Facebook, Apple, Baidu, Alibaba, and Tencent. Several studies have recently discussed the influence that big

tech companies can exert on customer decisions using artificial intelligence (AI) (André et al. 2018, Bjørlo et al. 2021). On the one hand, users can benefit from applications that help them to deal with the plethora of available data, while, on the other hand, they can impair users' autonomy if such tools allow external parties to intervene in people's behavior and decision-making. Such intrusions have been classified as violation of privacy (Solove 2006, Murata & Orito 2021). The question of how the commercial use of data and AI affects (customer) autonomy is attracting increasing interest in research (Wirtz et al 2021). Here, we must look at both, the information or data leaving a person's zone of privacy and the resulting effects when this information data is used to influence that person. Both aspects are shaping the future of the digital economy, leading Davenport to identify the impact of digital technologies on autonomy as one of the most important open research questions in ethics. (Davenport et al. 2020).

What makes a person's data private often depends on how this data is used rather than the data as such; looking at user autonomy can show us more about the impact of such different types of use. We also want to distinguish between how data can leave a person's zone of privacy: one is the transfer of complete units of information such as secrets to external parties, the other is the leaking of tiny bits of often insignificant data to external parties who compile it into powerful profiles. The latter form of leakage users may only be partially aware of although the capabilities of AI in this respect are growing significantly. It is also this form of leakage that enables subtle manipulation and leads to a significant loss of users' autonomy.

An aspect that makes this matter more urgent is related to the fact that digital technologies are penetrating ever more deeply into people's private spheres. Mobile applications such as smartphones have become most people's ubiquitous companions.

Wearable applications have reached people's skins, and VAs have come to take part in conversations at home. The next goal is already in sight, namely to penetrate the human body (Fernandes 2016). As a result, human privacy is becoming increasingly blemished, which in turn makes it possible for technology to intervene more and more in personal autonomy.

In a 4-year project (2019 – 2023) funded by the Swiss Science Foundation, a team of researchers with expertise in human-centered design and interaction, home automation, digital services, data science, and behavioral economics has been investigating the influence of VAs on people's daily life and attitudes, as one example of technologies collecting large amounts of data. Users' understanding of their privacy with respect to VAs and of big tech companies' use of data from these VAs has been a focal point of the research. VA use illustrates the discomfort that many people have regarding the use of data by big tech companies and shows users' problem how to distinguish between beneficial and malicious use of novel data-based technology in the first place.

In the current study we want to slightly shift the focus of analysis from the core topic of data privacy and its protection to the question of how people actually see the danger in the effects of such data breaches and whether recent legal initiatives can deal with this problem. To this end we look at the concepts of privacy and autonomy in Section 2, before we examine the connection to data aggregation and manipulation

in Sec. 3. In Sec. 4 we then turn to a general consideration of the data economy and recent European legal initiatives to deal with issues of privacy and autonomy. In Sec. 5 we discuss insights from our studies on the perceptions of users before we finally conclude with a discussion and outlook of the study.

2 Privacy and Autonomy

Solove (2008) described privacy as a pluralistic concept that is best described as a “cluster of many distinct but related things”. A kind of common denominator is the view that people need privacy as a kind of protective zone around them; the further others enter this zone, the more vulnerable the person becomes (Altman 1975). A consequence of this is that the borderline between private and public, always remains negotiable since it is based on an interpersonal process of trust. This variability requires people to continuously reassess the situation and redefine the limit of privacy. Such reassessment also depends on the perceived risk that they associate with opening up to third parties.

Everything that enters a person’s zone of privacy unfiltered from the outside can affect his or her autonomy. Personal autonomy refers to the capacity of conducting self-directed actions (Buss & Westlund 2018) or making independent decisions, as part of the control over one’s own life (Becker 2019). The concepts of privacy and autonomy, while related, are distinct. Becker (2019) has pointed out that persons who are observed undressing lose their privacy but not their autonomy, whereas a person in captivity can enjoy a high degree of privacy but loses autonomy. Generally, however, it can be said that invasion of privacy is often accompanied by a loss of autonomy when that invasion is used to influence a person’s action.

All types of influence on a person might be considered as critical in terms of that person’s autonomy but also here the limit is variable and it might not necessarily result in a violation of privacy. For example, a person who seeks to persuade a friend to refrain from an action that he or she considers a mistake and is convinced that this action is not in the friend’s best interest will not be necessarily considered to have violated the friend’s autonomy. In the same way a digital service that directs a customer to the most suitable product for his or her particular need is not violating this person’s autonomy although such guidance can also be interpreted as manipulation, in particular if the reasons for a recommendation remain intransparent. In many cases such services rather increase the person’s autonomy because it helps them dealing with the large amount of available information. Indeed, mutual influence on each other’s decision, for example using persuasion, is in general considered a mostly legitimate means of social coexistence.

When it comes to the violation of autonomy is another issue is decisive: if we relate autonomy to self-directedness, we must not confuse autonomy with arbitrariness. Indeed, the second part of the word autonomy originates from the Greek word *nomos* (“law”) and indicates that autonomous decisions follow certain principles that are related to a person’s goals and values (Oshana 2016) or higher-

order attitudes (Buss & Westlund 2018). Thus, some influence that is consistent with these higher-order attitudes of a person does not violate his or her autonomy, whereas manipulation contrary to these higher-order attitudes does.

Furthermore, we have to distinguish between influence that is apparent or whether it is subliminal. Whereas blackmail has an obvious influence on persons' actions, a subconscious influence of their feelings is more difficult to recognize. Therefore, it is important to understand where subliminal influence might come from and how to deal with it.

3 Data Aggregation and Manipulation

The question if customer profiles, big data technologies and artificial intelligence for behavioral targeting can be used to manipulate customers has already been discussed in literature (André et al. 2018, Davenport et al. 2020). It has led to the development of so-called *behavioral targeting*, that is the use of personal behavioral data for the selection of displayed information or advertisement (Chen & Stallaert 2014). How behavioral targeting can be used to manipulate customers has already been discussed in literature (André et al. 2018, Davenport et al. 2020). Such personalized targeting (also called *micro-targeting*) makes use of all available user data including data about people's emotions, locations and moods with the intention to use this data to influence them in a very individual manner (Vold & Whittlestone 2019). It is generally rather difficult for the targeted individual to recognize this influence.

However, we must distinguish between different forms of manipulation, which can be rather obvious such as coercion or hidden by feigning false facts or bypassing people's rational decision (Noggle 2020). We will use the term manipulation for the hidden ways to undermine people's autonomy, excluding coercion from the current consideration. Use of profiling and AI can play a crucial role when used maliciously. Referring to the characterization of autonomy, we only use the term manipulation if it causes a person to act in a way that is inconsistent with their higher-level attitudes. Here, we acknowledge that mutual influence on each other's decision, for example using persuasion, as a mixture of argumentation and manipulation, is a general means of building consensus. It only becomes a problem when any argumentative control is bypassed.

Today we observe increasing possibilities to use AI-based information technology to manipulate people—also described as automated influencing—, which can take the form of online personalized or targeted advertising, recommendation systems, newsfeeds and search engines (for example, see Benn & Lazar 2021). It is based on a combination of big data collection and use of AI. The Internet of Things (IoT) plays a crucial role in this respect because it facilitates the aggregation of large amounts of data, which at first glance appears to be not really problematic but turns out to be quite critical with regard to tracking and profiling of users (Kröger 2019). The same holds for selftracking devices such as wearables (Lanzing 2016). Voice Assistants (VAs) have been identified as similarly intrusive (Chamarajnar & Ashok 2019). The problem of slow pervasive data aggregation has also been discussed in the

context of VAs (Chung & Lee 2018). Data from VAs can be used to gain very deep insights into a human psyche can be illustrated by a recent example given by Dickson (2019), who described that VAs can be used to analyze a person's voice utterances and predict when that person's current relationship is likely to end. It is obvious that similar algorithms can be used to influence people's emotions by addressing their vulnerabilities.

Data can leak from VAs and allow malicious agents to accumulate considerable information over time. In principle people can evaluate whether the expected benefits of a service are worth this risk, which leads us to the well-known choice between convenience and privacy (Chellappa & Sin 2005, Cao & Wang 2022). But, as Pascalev (2017) has pointed out, the large amounts of data and their complexity—data is transferred in tiny bits—as well as the growing ubiquity of data-producing services makes this virtually impossible.

The risks for human autonomy related to the availability of the high amounts of data have been discussed for some time (Calo 2014, Zuboff 2015, 2019, 2020). The more we rely on digital services that create personal profiles, the more vulnerable we become to manipulation and disinformation (Kertysova 2018), raising the question of how human autonomy can be preserved under these conditions. Vold and Whittlestone (2019) suggested criteria to distinguish between an appropriate and an inappropriate use for data for user profiling: (1) personalized targeting should be consistent with people's higher-order attitudes, (2) personalized targeting should be transparent, (3) companies should attempt to seek consent, (4) personalized targeting should not misinterpret reality, and (5) personalized targeting should not exploit personal vulnerabilities. The problem with these points is that they are mostly not accessible to the affected users.

4 Datafication and Data Economy

A new phenomenon of the data economy is the so-called datafication, that is the generation of value from data through the processing of data that are available everywhere. This development transforms data from a commodity to a form of capital (Sadowski 2019, Couldry & Mejias 2019). With the increasing accumulation of data as the driving force for new business models, profit-driven companies become data-driven companies. At the same time, there are calls for better privacy and user autonomy to protect customers from becoming pawns in the game they cannot control (Koskinen et al. 2019). A first step in this direction would be to increase customers' awareness of the challenges this entails.

Given the opportunities of datafication, it is not surprising that the aggregation of data and the development of increasingly powerful AI algorithms have become a core area of the digital economy and competition. Given the fine line between value-added service provision for the benefit of customers and the possibilities of malicious manipulation, the question of how to deal with the associated challenges is becoming increasingly important. This requires clear criteria which use is supportive and which is undermining customer's autonomy. This task cannot simply be accomplished by

leaving customers with the decision to give or deny explicit consent to the processing of personal data, as we currently see it with the General Data Protection Regulation (GDPR) (European Commission 2016). Customers consider this approach to be excessively complex and not very helpful. Moreover, it is not so much the data that makes the difference between valuable and malicious use of data but the use itself, that is, the same data can facilitate supportive assistance, as well as allow manipulative interventions. It is difficult to say in advance which data can be misused and in the hands of a malicious agent, in aggregated form even seemingly harmless data can be transformed into dangerous instruments.

Recently, the EU agreed on far-reaching new digital rules to curb the power of big tech companies. In March 2022, negotiators from the European Parliament and the Council reached an agreement after months of talks. The European Union's new set of rules must now be approved by the Council and the European Parliament. These rules are laid down in the Digital Services Act (DSA) (European Commission 2022a) and the Digital Markets Act (DMA) (European Commission 2022b). According to the DMA, big tech companies would face tighter restrictions on using people's data for personalized targeting. For example, companies would not be allowed to rank their own products or services higher than those of others in online search results or reuse data collected from different services. Moreover, customers' personal data cannot be aggregated for targeted ads unless explicit consent is given. Currently, the European Commission is working on a regulatory framework for Artificial Intelligences (also called AI Act) that is the first initiative to address the problem of manipulation (Kop 2021, Boine 2022). The DSA aims at the prevention of disinformation in terms of various negative effects such as in civic discourse or electoral processes. Its contribution to prevent manipulation concerns increased transparency even though it is mainly focusing on online service platforms and does not address other applications that might affect people's autonomy (Boine 2021).

Ultimately, the question remains as to how users perceive the challenges posed by subliminal manipulation and what options they see for dealing with them. This also raises the question of the extent to which they trust the regulatory options of legislators. Considering the broad spectrum of possible privacy violations, the manipulative use of data profiles appears to be a question that receives less attention.

5 Insights from User Studies on Voice Assistants

5.1 Approach

The empirical data collection was conducted in two different study groups. Group 1 included 31 participants who voluntarily took part in a four-week self-observation study in their homes. Information about their usage patterns was recorded in a diary app on the one hand and asked in interviews on the other. Further interviews were conducted 3 to 4 months after the in-home study to address longer-term effects of VA use. The age of participants in Group 1 ranged from 17 to over 70 and included both novice and tech-savvy individuals.

Group 2 consisted of a total of 95 students, mostly in their early twenties. Their data were collected as part of 13 weeks of coursework on VA technology. Their observations and responses are recorded via the "relax-concentrate-create" (rcc) platform of the Lucerne University of Applied Sciences, which is accessible to students and staff in Switzerland. The goal of the study was to capture students' attitudes and practices regarding the use of voice assistants in their personal lives. They could opt to explore VAs from one of three perspectives: Already using a VA; curious to find out more about VA; not interested in or opposed to using a VA.

The material from both the in-home study and the student study group has been analyzed and compared with the transcription of audio files from semi-structured interviews that were conducted via Zoom about 3-4 months after the observational study. For this purpose, we used MAXQDA and analyzed all the data using qualitative coding.

In the following analysis of the user data we investigate the topic of privacy and use of data with respect to users' attitudes towards the collection of data and how they perceive their situation in relation to big tech companies.

5.2 Results

In connection with privacy issues, both groups of participants were asked how they perceived the effect of VA use on their privacy. The responses reveal that most participants are aware that VAs give the big tech companies access to a very private area of their lives but tend to have diffuse ideas about the nature of the risks. Here the different setting for the two groups became relevant. Whereas Group 1 took part in a study with a focus on using the VA at home and in their family, Group 2 had the task to assume a reflective position on the use of VAs. They should develop a position towards a possible use of a VA. The different motivation of the two groups also led to different ways of looking at the potential threats.

The following quotes translations of the original German statements by the authors and assigned to Group1 or Group2, respectively. The selected quotes reflect the fundamental views of the participants regarding the use of VAs and the recording data. Typical quotes from Group 2 regarding the use of the collected data do not show any significant confidence in transparent use:

I find it difficult to trust the manufacturers. Also the lack of transparency of what happens directly to my data and whether this data is not being used against me is not entirely clear to me.

Due to the constantly activated microphone and the profile created of a person, the data could be misused, or not only used for what I actually want it at that moment.

I am a little uncomfortable with the thought that there will be more data floating around somewhere that can be used to analyze me and sell me things.

These quotes illustrate the skepticism towards big tech companies. However, regarding the possible use of misuse of their data, the participants did not utter clear ideas. They are aware that the data may be analyzed but not for which purpose. The most concrete assumption was that the data might be sold to other organizations. In most cases they also used the term *data* and not *information*, expressing a rather broad view of what might count as data

Despite the legislative initiatives by the European Commission public institutions are not perceived as decisive actors in the control of data handling. Users largely felt that they were left alone with the big tech companies.

Google may say what they want, but [I am convinced that] my data will still be used by third parties and, thus, I become a transparent citizen.

I read the terms and conditions, but you could not use some functions if you would set certain settings for your own data security. You actually buy functionality by giving Google your data.

Ultimately, Amazon, Apple, Google & Co. store and process the records of their voice assistants according to their own rules. In a country where a good part of the population has been systematically spied on by the state for decades, this kind of thing creates a lot of problems. by the state for decades, something like this inevitably creates a queasy feeling.

In particular the final quote shows that some participants, at least, were quite cautious about government actors and were worried that data could be passed on to government agencies for intelligence and similar purposes. Obviously these are not ideal conditions with regard to trust in public institutions.

If we compare this to the interviews with participants of Group 1 we find slightly different kinds of statements.

If someone comes to visit me, I should actually specifically tell them that my voice assistant might record conversations.

I don't think the voice assistant is allowed to randomly record things without consent. However, when I start a web search or a manual search, it is clear that this info is used.

The technology behind it is the same. Whether I type my information in or do it with my voice. It gets memorized by the company and I don't think voice makes it less secure.

What should they do with all of that information? Most of the things I say are not that interesting

These statements of Group 1 mainly deal with *information* and conversations and only to a minor degree with data. Often participants refer to other applications than VAs if they use the term data. The view of Group 1 is more influenced by the concrete use of the VA than by general considerations. Accordingly, participants of Group 1 related possible privacy breaches mainly to the conversations they had and the topics discussed e.g., health or financial matters.

Some interviewees of Group 1 suspected a linkage of their conversations with the VA to certain advertisements they were shown on the internet. This shows that, at least at a certain level, they expected manipulation based on the conversations with their VAs. Nevertheless, this was a diffuse feeling rather than clear evidence. The fact that companies use data from their VA for personalized advertisement was considered more of a nuisance than a real problem (cf. also Kitkowska et al. 2017).

Both groups partially realized to which extent personal data can be aggregated and analyzed. The diffusion of bits of data here and there absorbed by the VA providing companies might appear to be relatively innocuous. The proliferation of data fragments ingested here and there by VAs was not considered as especially frightening by Group

1 participants, even if aggregation over time may reveal a great deal about people's habits or preferences. The danger of influence and manipulation using data and the limitations to protection have been mostly discussed in relation to political issues (Bartlett 2021) but must also be seen as a concern about the future position of customers in the data economy.

6 Discussion and Outlook

Although people subconsciously assume that data profiles are used to manipulate them, for example, when it comes to targeted advertisements, they have no clear idea to which extent this might affect their decisions or actions. The ever growing power of AI technology combined with well-known psychology makes it difficult to assess the possibilities realistically. The interviews in the project have shown that participants still mostly think of privacy in terms of secrets and information, rather than the diffuse mass of tiny data bits that can be translated into personal psychological profiles. As long as these profiles and their use are hidden from the users they remain powerless.

A particular difficulty for users consists in the different kinds of transferred data. While it is rather clear to them what the recording and transfer of an utterance by the VA means, it is less clear what the transfer of a continuous data stream of voices and ambient noise, which often appear quite insignificant to users, might mean once they are aggregated and analyzed by the big tech companies. With small bits of data leaving the zone of privacy it is no longer clear which information is actually transferred. We could compare the problem with the so-called sorites or heap paradox (Keefe 2000); in the same way as it is unclear whether the addition of one grain turns an existing collection of grains into a heap, it is also unclear whether an additional bit of data leaving the zone of privacy via the VA generates information that would count

as a privacy breach. This means that users must get access to the profile information that results from their continuous data streams. It is actually a challenge to process the data in a way that such a profile would be comprehensible for human users.

In the further course of the project, we will investigate the topic of privacy and the effects of privacy and autonomy violations more deeply. We want to gain a deeper understanding of people's mental models of privacy and how they perceive data leaving their zone of privacy. In additional interviews we have already observed that this zone of privacy is variable because people agree that the same data might be used for different purposes. In this respect, most people call for more transparency. However, they perceive their influence on what happens with their data as rather limited—and it seems that they are not wrong in this respect either. The current protection measures that GDPR provides tend to be seen as too complicated and not really efficient. It is not the particular type of data that determines the harm but the aggregation on the side of the big tech companies, which is not transparent for users.

In the second half of this year, we are planning a further interdisciplinary event in which 20 students will engage in four creative workshops over the course of a week to carry out a future-oriented (inter)connected media project. In this context, the topic "data protection and privacy using virtual assistants in the year 2037" is to be addressed. On the one hand, these projects will look at the problems described above from a slightly different perspective; on the other hand, the more extensive treatment of the topic means that insightful ideas and approaches to solutions can be expected for the future in dealing with intelligent (in the worst case, manipulative) systems.

Acknowledgements

This study was funded by Swiss National Science Foundation (SNF) project VA-PEPR, ref. no. CRSII5_189955.

References

- Altman, I. (1975). *The Environment and Social Behavior: Privacy, Personal Space, Territory, and Crowding*. ERIC, Monterey, CA, USA.
- André, Q., Carmon, Z., Wertenbroch, K., Crum, A., Frank, D., Goldstein, W., Huber, J., von Boevn, L., Weber, B. & Yang, H. (2018). Consumer choice and autonomy in the age of artificial intelligence and big data. *Customer needs and solutions*, 5(1), 28-37. <https://doi.org/10.1007/s40547-017-0085-8>
- Bartlett, M. (2021). Beyond Privacy: Protecting Data Interests in the Age of Artificial Intelligence. *Law, Technology and Humans*, 3(1), 96-108. <https://doi.org/10.5204/lthj.1595>
- Becker, M. (2019). Privacy in the digital age: comparing and contrasting individual versus social approaches towards privacy. *Ethics and Information Technology*, 21(4), 307–317. <https://doi.org/10.1007/s10676-019-09508-z>

- Benn, C. & Lazar, S. (2021). What's Wrong with Automated Influence. *Canadian Journal of Philosophy*. <https://doi.org/10.1017/can.2021.23>
- Bjørlo, L., Moen, Ø., & Pasquine, M. (2021). The Role of Consumer Autonomy in Developing Sustainable AI: A Conceptual Framework. *Sustainability*, 13(4), 2332. <https://doi.org/10.3390/su13042332>
- Boine, C. (2021). AI-enabled manipulation and EU law. *SSRN Electronic Journal*. <https://dx.doi.org/10.2139/ssrn.4042321>
- Buss, S. & Westlund, A. (2018). Personal autonomy. Stanford Encyclopedia of Philosophy. Retrieved from <https://plato.stanford.edu/entries/personal-autonomy/> Accessed 10 May 2022.
- Cao, G. and Wang, P. (2022), Revealing or concealing: privacy information disclosure in intelligent voice assistant usage- a configurational approach, *Industrial Management & Data Systems*. <https://doi.org/10.1108/TMD-08-2021-0485>
- Calo, R. (2014). Digital Market Manipulation, *George Washington Law Review* 82(4), 995–1051.
- Chen, J., & Stallaert, J. (2014). An economic analysis of online advertising using behavioral targeting. *MIS Quarterly*, 38(2), 429-A7. <https://doi.org/10.25300/misq/2014/38.2.05>
- Couldry, N., & Mejias, U. A. (2019). Data colonialism: Rethinking big data's relation to the contemporary subject. *Television & New Media*, 20(4),336-349. <https://doi.org/10.1177/1527476418796632>
- Chamarajnagar, R. & Ashok, A. (2019). Privacy Invasion through smarthome IoT sensing, 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), pp. 1–9. <https://doi.org/10.1109/SAHCN.2019.8824933>
- Chellappa, R. K. & Sin, R.G. (2005). Personalization versus Privacy: An Empirical Examination of the Online Consumer's Dilemma. *Information Technology and Management* 6(2-3), 181–202. <https://doi.org/10.1007/s10799-005-5879-y>
- Chung, H. & Lee, S. (2018). Intelligent virtual assistant knows your life. Retrieved from <https://arxiv.org/pdf/1803.00466>
- Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*, 48(1), 24-42. <https://doi.org/10.1007/s11747-019-00696-0>
- Dickson, E. J. (2019). Can Alexa and Facebook predict the end of your relationship?. Retrieved from <https://www.vox.com/the-goods/2019/1/2/18159111/amazon-facebook-bigdata-breakup-prediction>. Accessed May 15, 2022
- European Commission (2016), "Regulation (EU) 2016/679 of the European parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (general data protection regulation)", *Official Journal of the European Union*, Vol. 59, pp. 1-88.
- European Commission. (2020a). The Digital Services Act: ensuring a safe and accountable online environment. Retrieved from <https://ec.europa.eu/info/strategy/priorities->

- 20192024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountableonline-environment_en
- European Commission (2020b). The Digital Markets Act: ensuring fair and open digital markets. Retrieved from https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_en
- Fernandes, T. (2016). Human augmentation: beyond wearables. *Interactions* 23(5), 66-68. <https://doi.org/10.1145/2972228>
- Keefe, R. (2000). Theories of vagueness. Cambridge University Press, Cambridge, UK.
- Kertysova, K. (2018). Artificial intelligence and disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered. *Security and Human Rights*, 29(1-4), 55-81. <https://doi.org/10.1163/18750230-02901005>
- Kitkowska, A., Meyer, J., Wästlund, E. & Martucci, L. (2017). Is It Harmful? Measuring People's Perceptions of Online Privacy Issues. In Thirteenth Symposium on Usable Privacy and Security, July 2017, Santa Clara, CA, USA.
- Kop, M. (2021). EU Artificial Intelligence Act: The European approach to AI. *Transatlantic Antitrust and IPR Developments*. Retrieved from <https://law.stanford.edu/publications/eu-artificial-intelligence-act-the-europeanapproach-to-ai/>
- Koskinen, J., Knaapi-Junnila, S., & Rantanen, M. M. (2019). What if we had fair, people-centred data economy ecosystems?. In 2019 *IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation* (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), IEEE, pp. 329-334). <https://doi.org/10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00100>
- Kröger, J. (2019). Unexpected Inferences from Sensor Data: A Hidden Privacy Threat in the Internet of Things. In: Strous, L., Cerf, V. (eds) *Internet of Things. Information Processing in an Increasingly Connected World. IFIP IoT 2018. IFIP Advances in Information and Communication Technology*, vol 548. Springer, Cham. https://doi.org/10.1007/978-3-030-15651-0_13
- Lanzing, M. (2016). The transparent self. *Ethics and Information Technology* 18(1), 9–16. <https://doi.org/10.1007/s10676-016-9396-y>
- Murata, K., & Orito, Y. (2021). The Privacy Paradox: Invading Privacy While Protecting Privacy. In [New] *Normal Technology Ethics: Proceedings of the Ethicomp 2021*, Universidad de La Rioja, pp. 199-201.
- Noggle, R. (2020). *The Ethics of Manipulation*, Zalta, E. (ed.) *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), Retrieved from <https://plato.stanford.edu/archives/sum2020/entries/ethics-manipulation/>.
- Oshana, M. (2016). *Personal Autonomy in Society*. Routledge, London, UK. <https://doi.org/10.4324/9781315247076>

- Pascalev, M. (2017) Privacy exchanges: restoring consent in privacy self-management. *Ethics and Information Technology* 19(1), 39–48. <https://doi.org/10.1007/s10676-016-94104>
- Sadowski, J. (2019). When data is capital: Datafication, accumulation, and extraction. *Big Data and Society*, 6(1), 1–12. <https://doi.org/10.1177/2053951718820549>
- Solove, D. J. (2006). A Taxonomy of Privacy. *University of Pennsylvania Law Review*, 154(3), 477-560.
- Solove, D. J. (2008). *Understanding Privacy*. Harvard University Press, Cambridge, MA, USA.
- Vold, K., & Whittlestone, J. (2019). Privacy, autonomy, and personalised targeting: Rethinking how personal data is used. In: Véliz, C. (Ed.), *Report on data, privacy, and the individual in the digital age*. University of Cambridge, Cambridge, UK.
- Wirtz, J., Hartley, N., Kunz, W. H., Tarbit, J., & Ford, J. (2021). Corporate Digital Responsibility at the Dawn of the Digital Service Revolution. *SSRN Electronic Journal*. <http://dx.doi.org/10.2139/ssrn.3806235>
- Zuboff, S. (2015). Big other: Surveillance Capitalism and the Prospects of an Information Civilization, *Journal of Information Technology* 30(1), 75-89. <https://doi.org/10.1057/jit.2015.5>
- Zuboff, S. (2019). *The Age of Surveillance Capitalism*. Profiles Books, London, UK.
- Zuboff, S. (2020, 01/24). *You Are Now Remotely Controlled*. New York Times, Retrieved from <https://www.nytimes.com/2020/01/24/opinion/sunday/surveillance-capitalism.html>

Mapping and understanding human factors in effective cybersecurity: a finance-sector organisation case study

Robin Renwick¹[0000-0002-2622-7823], Eliza Jordan¹[0000-0001-5534-9515], Eliseo Venegas Mayoral²[0000-0001-7995-5499], Amanda Segura Gonzalez²[0000-0003-2945-4412] and Leire Cubo Arce²[0000-0002-3551-5261]

¹Trilateral Research, Waterford, Ireland

²NTT Data, Madrid, Spain

robin.renwick@trilateralresearch.com

Abstract. The role of human factors in cybersecurity has gained increasing interest in the past decade. Individuals, and the interplay with their societal and organisational environments, have been highlighted as influencing the overall effectiveness of cybersecurity strategy. Social science research methods can provide a unique avenue for obtaining a holistic overview of how cybersecurity is both implemented and perceived within an organisation. One-to-one semi-structured interviews were conducted with 17 participants from a finance sector organisation. Interviews consisted of a range of questions about respondents' perceptions of cybersecurity in their workplace along with a series of hypothetical scenarios. Five themes emerged from the thematic analysis: *friction between cybersecurity and other processes; resource allocation; collaborative cybersecurity and professional responsibility; the role of training and awareness in effective cybersecurity; and employee and organisational trust*. Resulting themes were explored and discussed in relation to previous literature. Cybersecurity employees, general employees, and management had both conflicting and consistent priorities in relation to cybersecurity. Pressures such as time and lack of resources were noted as disablers of effective cybersecurity in the organisation. Moreover, the study details the potential benefit of social science research methodology being incorporated into existing risk assessment frameworks, to provide a holistic view of an organisation, and understanding cybersecurity as the symbiosis of technology, policy, process, and people.

Keywords: Cybersecurity; Human factors; Accountability; Trust

1 Introduction

It is widely recognised that cybersecurity is rooted in socio-technical constructs. Thus, cyber resilience should be viewed as a symbiotic compound of technology and behaviour (Carlton et al., 2019; D'arcy et al., 2009; Siponen and Vance, 2010; Whitman, 2018). Within cybersecurity, the individual is recognised as integral to operational security, with constructs such as awareness, perception, proficiency, and

personality viewed as determinants of success or failure (Oltamari et al., 2015). In recent years the discussion of human factors-based cybersecurity has begun to gain significant attention, as information security professionals seek to understand the root constructs of threats, vulnerabilities, and attack vectors.

While the individual is undoubtedly central when considering human factors, it is important not to lose sight of the fact that the individual almost always acts in engagement with societal and organisational structures (Greitzer et al., 2018). Factors such as teamwork performance (Buchler et al., 2018) and team dynamics (Jose et al., 2016) greatly influence organisational cybersecurity. At the management level, management ideology, upper management support of cybersecurity interventions, performance measurement and work design, amongst other factors, influence how incidents are handled (Mueller-Hanson & Garza, 2016; Zaccaro et al., 2016; Kraemer et al., 2009). For example, management responses that incorporate shaming have been shown to be counter-productive, negatively impacting on employee's work attitudes and employer relationships (Renaud et al., 2021). ENISA, the European Cybersecurity Agency have previously urged security professionals to 'think like social scientists' (ENISA, 2019b, p.6). However, it is still somewhat unclear what benefit this change in framing might have, or how it may work in practice.

In this study, social research methods have been used to uncover aspects that negatively impact overall cybersecurity resilience. It demonstrates how a social research methodology might be implemented to assess human factor-based cybersecurity to provide a critical communication avenue between employee and employer – mediated through an ethical and privacy-respecting social-science based method. Additionally, social research techniques allow a holistic view of the organisation to be developed through interaction with different departments, job roles, and functions. This holistic view draws out insights into how an organisation functions – as opposed to how various departments and roles perceive it to function.

This paper presents a case study conducted within a financial services sector organisation. A series of interviews were conducted to understand both individual and organisational processes that affect the cybersecurity environment. From the research analysis five themes were uncovered, that together provide a lens to understand the organisation more effectively, informing decisions regarding the development of cybersecurity policy and the effectiveness of its implementation. Crucially, the research reveals whether mismatches exist between how management believes cybersecurity is within their organisation, and how it actually is.

2 Hypothesis

The research team hypothesised that interplay between individual and organisational processes frequently yields contradictory cybersecurity incentives. In line with previous literature and gaps in research, three hypotheses were ideated:

- (i) Management, the cybersecurity department, and non-security employees have conflicting priorities regarding cybersecurity.

- (ii) Certain pressures exist that act as disablers of effective cybersecurity in the organisation.
- (iii) Viewing employees as threat vectors impacts trust levels within the organisation.

Cybersecurity literature provides us with a template through which to understand sources of threats or vulnerabilities – often framed through the term ‘localization’ (ENISA 2019a; Ruf et al., 2008). Operational security is nearly always viewed through concepts such as awareness, perception, proficiency, competence, and personality. These concepts ultimately determine whether a specific cybersecurity policy, or process, is effective at mitigating a particular or general cybersecurity threat (Oltamari et al., 2016).

The research presented in this paper demonstrates a method, used to gain deeper insight into an organisation, especially concerning daily operations. The goal is to provide a holistic view, as told by those responsible for implementing cybersecurity procedures. It is believed that a mixture of individual and organisational processes will be communicated by participants, which in turn will allow researchers to gain a detailed picture of how cybersecurity resilience manifests within the organisation.

3 Methodology

To address the hypotheses detailed above, qualitative research was conducted as part of the H2020 EU-funded project SOTER (Grant Agreement No. 833923). A total of 17 participants working in the finance sector took part in the study. Of the 17 participants, six worked in the cybersecurity department and eleven worked in other departments. There was a mixture of senior and middle managers interviewed, as well as a mix of cybersecurity and non-cybersecurity professionals. The broad coverage of employees allowed the researchers to test whether mismatches existed between departments and whether perspective differences existed both horizontally (across departments) and vertically (through the management structure).

One-to-one semi-structured online interviews were conducted, including questions related to participants’ perceptions and views of cybersecurity processes and procedures in their workplace. Additionally, participants were asked to reflect on a series of hypothetical scenarios (vignettes). These hypothetical scenarios allowed the researchers to understand participants’ perceptions and attitudes towards cybersecurity, without there being a defined emphasis on the individual and their specific role/tasks. The interview schedule consisted of six semi-structured questions. The six questions were in turn divided into three sections, with each of the three sections linked to one of the research hypotheses. The first section was concerned with cybersecurity processes – attempting to engage with what processes participants felt were effective and those that required some improvement. Following on from this, the next section included questions framed to allow participants to reflect on specific conflicts presented in their employment – especially any conflicts or pressures they encountered related to primary tasks and cybersecurity specific tasks.

The final section attempted to uncover any negative aspects of their employment, especially as related to psychological pressure or concern. Woven into the questions were three hypothetical scenarios. These scenarios were used to present a workplace situation. In the scenario, a character was presented with a cybersecurity related issue. Participants were asked to rate how much they agreed with the character as they responded. This determination was completed with a Likert scale of 1 (Not at all) to 5 (Very much). Moreover, participants were asked to reflect on their rating and expand on their views about the scenario.

Additionally, warm-up questions were included at the beginning of the interview to get participants to think about the topic and provide examples of cybersecurity processes in their workplace. A series of prompt questions were also provided for the researcher to prompt the participant to develop on their responses, where appropriate. At the end of the interview, participants were asked whether there was anything they wished to add.

After the interviews were conducted, the research team transcribed interviews for analysis. The analytic approach used was inductive thematic analysis (Braun and Clarke 2006). Interview transcripts were coded using semantic and latent coding. Researchers coded the interview transcripts twice and convened to discuss codes, inconsistencies of interpretation, and define resulting themes. The research team maintained a critical realist epistemological position (Bhaskar, 2008), accepting participants' reports as reflecting their reality, while simultaneously acknowledging that the reality may be influenced by their own socio-cultural environment.

From the analysis, five core themes (and related sub-themes) resulted: friction between cybersecurity and other processes; resource allocation; collaborative cybersecurity and professional responsibility; the role of training and awareness; and employee and organisational trust.

4 Results

The core themes (and related sub-themes) detailed above portray how cybersecurity was operationalised in the organisation. Understanding key themes and topics within the organisation, as told by the employees, provides an avenue for improving the overall cybersecurity resilience in the organisation, especially when concerned with identifying vulnerabilities in the organisation and understanding how better to mitigate potential (or ongoing) impacts.

From the thematic analysis it quickly emerged that certain sub-themes existed. Engaging with sub-themes allows a holistic understanding of the organisation to be developed. Below is an overview diagram of the primary themes, and related subthemes. These will then be discussed in the sub-sections below.

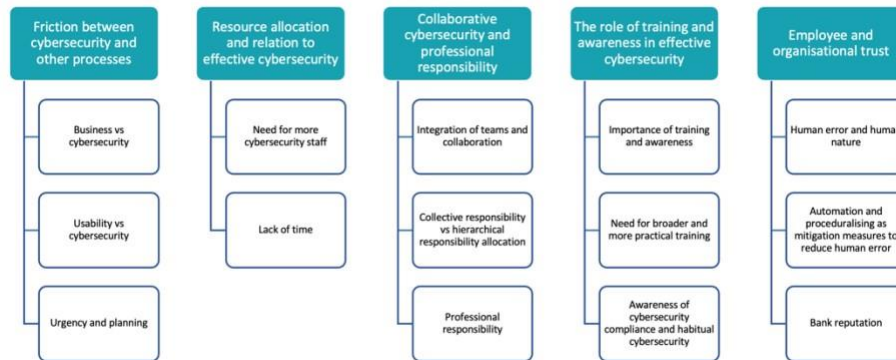


Figure 1. Overview of identified themes and sub-themes

4.1 Friction between cybersecurity and other processes

The friction between cybersecurity and other work tasks and procedures was noted. Achieving business objectives whilst complying with (sometimes) strict cybersecurity procedures was a discussed friction point, reported by both cybersecurity and the nonsecurity (general) staff. Most employees emphasised the need to find a balance between business/task-related demands and cybersecurity procedures.

“The difficult part is where you strike the balance between cybersecurity and business. Both worlds need to co-exist: let customers do the operations they want, whilst also complying with cybersecurity procedures in a secure way”.

General staff discussed that daily workplace performance may be hampered by compliance requirements, reducing usability of applications, or slowing down tasks. Information security staff discussed how business operations often required a high level of flexibility – which in turn meant that they had to adapt and change processes to meet specific company objectives. A basic business need for operational flexibility meant that employees believed specific situations existed where cybersecurity procedures were required to be more adaptive. Urgency also caused friction, as it was perceived to lead to a reduction in cybersecurity compliance. On many occasions employees mentioned that the actual execution of operations may be affected by circumstances or scenarios that require an acceleration of processes, and this acceleration often meant a trade-off against rigorous cybersecurity procedures.

“Rushing can cause many cybersecurity incidents. Human errors occur when you start a project and leave out some procedures in order to get the project done faster”.

4.2 Resource allocation

The importance of appropriate resource allocation is a recurring theme in cybersecurity research, as greater demands are placed on companies as they move further into digitization of their processes. The move towards larger and more complex IT systems places increasing strain on existing, sometimes outdated, IT support systems, while the overall attack surface widens as companies move services onto cloud or distributed infrastructures. Resource allocation was a central theme within the interviews, as both the general staff and the cybersecurity staff believed that greater emphasis on ensuring adequate resource allocation, including staff and time allocation, would provide for increasing returns to overall cybersecurity resilience.

Participants reported the importance of planning projects and tasks and factoring in cybersecurity processes into project timescales to ensure they are not omitted due to urgency.

“Before starting a project, time is spent thinking about how the project needs to be done. This initial loss of time in designing how to do the project later means the work is more efficient. If you don’t dedicate time to designing and you start working, in the end it will take longer to complete the project and with less preparation comes more probability of wasting time and risks occurring”.

They also mentioned both finding and retaining talent (especially security professionals) as a key driver of overall cybersecurity resource adequacy – made even more pertinent in an age where distinct human resource shortages existed, while human capital freedom of movement and increasingly prevalent ‘work from home’ policies meant that staff found it easier to change their employer.

“Cybersecurity and security departments today have a very big problem: talent retention. There is a lot of mobility of profiles of very talented people”.

4.3 Collaborative cybersecurity and professional responsibility

There was a sense of collective responsibility concerning the implementation of cybersecurity procedures. General staff and cybersecurity staff differed in their views on the division of responsibility, with some participants discussing top-down hierarchical structure and others discussing a mix of collective and hierarchical responsibility as being the most effective organisational structure for implementing effective cybersecurity processes.

The notion of professional responsibility was prevalent, with participants acknowledging that staff must have a global vision of what is best of the organisation, and this clear vision alleviates uncertainty regarding professional obligation and

responsibility. One participant when discussing whether to report a cybersecurity incident, stated:

“It is not a question of how you feel, but a question of responsibility. In that sense, you have to act responsibly”.

Over the course of the interviews, it was also made clear that close integration between the cybersecurity team and other teams was a distinct requirement, mainly to help employees manage cybersecurity practices and concerns in their work. The active involvement of cybersecurity staff throughout the lifecycle of projects was discussed. It was noted that having the cybersecurity team in continuous contact raised the effectiveness of decision making, provided staff with assurance that their work was secure and improved overall resilience in the organisation.

“(...) you involve the relevant departments early on, including the cybersecurity department, and then you start to shape it, so that we don't move in a direction that later becomes inoperative”.

“From the moment cybersecurity experts are involved in the projects, you can know that the steps you are taking in the project are secure and have all the guarantees”.

4.4 The role of training and awareness

Training and awareness were recurring topics in the interviews, as participants acknowledged the importance of both employees' (and clients') level of awareness and knowledge of cybersecurity issues. Multiple participants stated that training employees at a user level was of crucial importance. They noted that not all employees have technical backgrounds, and so needed support with technologically bound cybersecurity practices. They felt that training assists in making sure that employees have knowledge of *what to do* and *what not to do*, providing context for critical behavioural and procedural aspects of their job.

“The most important thing for me is not really cybersecurity as such, but the training we receive on cybersecurity at user level. I don't have a technical background, so there are many things that escape me, so I really appreciate the training (...)”.

Moreover, participants highlighted certain tests sent by the cybersecurity team which allow employees to apply their skills in a practical setting to condense and reinforce their knowledge. This was universally felt as beneficial, as it raised general state of readiness levels, and kept employees 'on their toes' as they went about their primary work-related tasks.

“(...) Another one that works very well is to make people in the company aware of phishing. We are always doing phishing campaigns, making people aware, when something doesn't seem right, they send us an email questioning whether an email is real”.

Participants reflected on the need for more training, that was broader and practical, to enhance their knowledge and awareness to apply of cybersecurity to their working routines and how to respond to potential cybersecurity incidents.

General staff also discussed that errors in the workplace were often due to lack of knowledge. Participants' unintentional oversights were not a product of maliciousness - highlighting the importance of training and awareness, as well as the provision of adequate and available information about incidence identification, handling, and response.

Furthermore, a sense of 'habitual cybersecurity' was noted from the interviews, with respondents mentioning cybersecurity as part of their daily life and routine, rather than being seen as a chore or inconvenience.

“[Cybersecurity] must be understood as something fundamental and incorporated day-to-day. Just like when you put on your seatbelt in the car and it is something you do automatically: security must be part of everyday life”.

4.5 Employee and organisational trust

The theme of employee and organisational trust also emerged from the interviews. There was consistent acknowledgement of the risk of human error, but it was also understood that human capital should be viewed as the most important resource. The importance of organisational reputation was heavily recognised across nearly all respondents, with participants highlighting that customers' trust is crucial for the organisation's overall reputation. Trust is a complex and contextually bound concept, but within the domain of cybersecurity it normally relates to the degree of confidence an entity has in people, systems, or technology – including any combination – to do what is expected of it. Within the research, the concept of trust situated itself through three dimensions – all of which raise complex questions about the nature of work, cybersecurity, and the impact that ineffective processes might have on business.

“(...) In the end it is a matter of trust. If there is an attack on a financial institution and many customers' bank funds are blocked, it would be difficult for them to trust the financial institution again”.

Participants were quick to acknowledge that the nature of humans meant that residual risk was always present in organisations – mainly as humans were prone to errors in their daily work.

“It is the quality of the people that makes them work well, rather than the processes themselves. You can have a good process but it is not followed. But you can have a good person who, even if there is no process, if he or she is a good professional, will solve the problem. I would emphasise the human resources much more than the process”.

Interviewees recognised that reliance on human capital meant that understanding the situated environment provided an avenue of understanding on what supports were in place, and what level of trust existed between organisation and individual. Providing an environment that is conducive to effective procedures was understood to be extremely important, and ultimately supported the development of habitual cybersecurity practice, timely incident response, and effective and accurate reporting.

“Tools that automatically monitor strange movements in clients’ bank accounts are also important to identify these actions proactively (...)”.

The question of automation was raised by both general level and cybersecurity staff, as it was understood to be an effective measure that could reduce the rate of error in an organisation. There was broad consistency on this theme, as employees agreed that automation could reduce workload and that this in turn could lead to an overall reduction in error rate.

This sub-theme can also be framed through the lens of trust, as it is related (at least in part) to the degree of trust that organisations have in their workforce. This discussion is especially relevant in the finance sector as algorithms are often used to identify fraud attempts – both identity and transaction-related. The ever-increasing dependency on machine or algorithm-based monitoring also has profound impact on the overall level of privacy afforded to both employee and customer – effecting levels of surveillance in the situated workplace environment.

5 Discussion

The conducted interviews provided an avenue for employees to communicate their perspectives, concerns, and ideas on cybersecurity in their workplace – communication that might not necessarily manifest in a normal working environment. Two of the three hypotheses were supported: hypotheses (i) and (ii). The analysis found both consistent and contradictory statements regarding cybersecurity when comparing general employees and cybersecurity employees. Additionally, participants reported pressures, such as urgency and lack of time, that impact effective implementation of cybersecurity. Although the concepts of trust and reputation were discussed, participants did not discuss employees being viewed as a threat vector negatively, thus hypothesis (iii) was not supported.

Obtaining perspectives from a range of employees in different teams and positions allowed the researchers to gather a holistic view of cybersecurity inside the

organisation. By doing so, it was possible to observe how differing views conjointly contributed to cybersecurity-related workplace dynamics, such as collaboration, communication, hierarchy, responsibility, as well as detail perceived and actual tensions and frictions. Interviewing staff directly furthers understanding of the effectiveness of cybersecurity procedures, policies, and practical limitations for implementation in the real world.

The study is not without its limitations. There was a moderate gender imbalance, with most participants being male. No front-office (customer facing) staff from the case study organisation were included in the sample which may have provided a different perspective into cybersecurity due to their direct contact with clients.

Notably, one of the hypothetical scenarios presented produced an array of responses, as it described an organisation which required employees' activity on their computer to be monitored and their cameras to be on. Various participants described the need to comply with the organisation's procedures, whereas others described a distinct invasion of privacy and potential violation of employee rights. This scenario presents an avenue for wider debate concerning professional responsibility and employer expectations – made more apparent in a post-COVID environment where remote working emerges as societal norm. Increasingly modern tools pose understandable concerns, as algorithmic processes are applied to workplace surveillance (Legg et al., 2015). Risks of surveillance normalization are real (Stanton and Stam, 2003; Taekke, 2011) as companies seek to reduce organisational risk profiles and mitigate potential impacts of insider threat (whether intentional or not).

The conducted study provides insight into how social research methods can contribute to better understanding of an organisation's current cybersecurity culture, to help identify vulnerabilities and risks in support of a more human-orientated organisational risk assessment. It also presents a method of how to involve employees in the overall design and deployment of cybersecurity practice, which has been noted as being beneficial (Heath et al., 2018). The goal was to understand the environment, its effect on individual compliance, and how best to incentivise modification of behaviours to improve overall cybersecurity culture and behaviour - in a trustworthy, inclusive and, ultimately, ethical manner.

6 Holistic risk assessment framework

Conducting risk assessments is a common and necessary practice in cybersecurity. It provides information to effectively determine actions required to identify, assess, and reduce risks that Information Technology (IT) systems present. IT risk assessments are usually conducted following the ISO 27005:2018 standard, which combines the risk assessment process defined in standard ISO 31000:2018 and the security considerations defined in ISO 27001:2013. Traditionally, experts who conduct these procedures focus solely on identifying and mitigating technical risks, but more efforts to incorporate the human factor element - increasingly relevant due to the increase of human-oriented attacks.

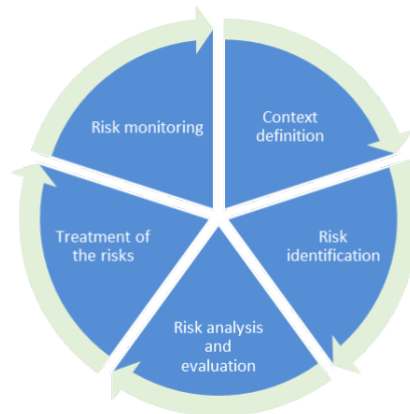


Figure 2. General risk assessment procedure

Understanding the human role in organisational security provides crucial context in the overall cybersecurity picture. Integrating human contexts shifts focus towards the situated environment, and focusses attention on one of the main attack vectors within any organisation. This holistic view provides a platform for a more encompassing cybersecurity strategy to be developed and allows a research team to uncover elements that might otherwise remain hidden to a cybersecurity department. Themes and subthemes may not be immediately apparent from surface level investigations, and providing a platform for employees to communicate perspectives, ideas, and concerns allows a far deeper understanding of an organisation to be developed.

The inclusion of social scientists, behavioural scientists, psychologists, and sociologists to conduct exploratory and investigative social research methods supports an empirical assessment framework for understanding existing practices, as well as suggesting mitigations and remedial actions. Integrating a social research-based exploratory exercise into the overall risk assessment process, an organisation has a method to understand what is happening within their firm, from the perspective of general staff, cybersecurity staff, as well as management. This leveraged knowledge allows them to build more resilient strategies and allows coordination and communication between varied departments - mediated by an impartial social research team. The process also allows an organisation to validate its initial risk assessment to understand if there are further vulnerabilities present that have not been immediately recognised.

The social research methodology can enhance the overarching risk assessment procedure by providing mechanisms for improving evaluation and understanding of the current cybersecurity status of an organisation. Integrating human and technical assessment can provide a host of benefits, including identifying bad practices; detecting malpractice and issues with policy compliance; supporting threat identification and developing potential mitigation strategies through collaborative discussion and ideating; improving internal and external risk assessment by providing an avenue for identifying mismatches and/or daily business activities that may be seen as vulnerabilities. Conducting social research through a third-party may also provide

an avenue for more honest feedback regarding policies, compliance, as well as providing a privacy respecting channel for improvement ideas that may not necessarily be existent in normal risk assessment practice.

7 Conclusion

By conducting qualitative research, this study has elucidated a deeper understanding into the perspectives and opinions of cybersecurity, from employees in different roles and departments in an organisation. This understanding includes interaction between roles, jobs, functions, and processes - providing a holistic view. General employees and the cybersecurity team noted conflicting priorities between their daily work tasks and cybersecurity tasks. This was mainly due to general employees' sense of urgency and pressure to meet business demands. Despite this, an overall feeling of responsibility to comply and obey cybersecurity measures was described. Training is a crucial pillar to supporting effective cybersecurity compliance within an organisation, as employees gain greater awareness of the impact of cybersecurity threats and attacks and incorporating cybersecurity measures and habits into daily work routines. Moreover, trust is a fundamental element, with a bidirectional trust relationship formed between the organisation and its employees and the clients and the organisation.

The current findings highlight the importance of understanding the human factor element in cybersecurity and serves as further insight into how cybersecurity is applied in real-life, and the challenges that surround its practical implementation. Incorporating social research methods as part of a holistic risk assessment framework to develop more resilient and detailed strategies and promote communication between departments is the proposed recommendation from this work. The combination of technical and social risk assessment iteratively conducted should provide an organisation of a rich picture of its attack surface, potential threats, and vulnerabilities, but also allow the firm to understand its own human resources and cybersecurity processes at a more granular level. This holistic understanding allows an organisation to think of its cybersecurity resources as a symbiotic compound of technology and behaviour - supporting the everincreasing literature that urges a human-focused approach to building cybersecurity resilience.

References

- Bhaskar, R. (2008). *A realist theory of science*. Abingdon, UK: Routledge
- Braun, V. & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3, 77-101
- Buchler, N., Rajivan, P., Marusich, L. R., Lightner, L., Gonzalez, C. (2018). Sociometrics and observational assessment of teaming and leadership in a cyber security defense competition. *Computers & Security*, 73, 114-136.

- Carlton, M., Levy, Y., & Ramim, M. (2019). Mitigating cyber attacks through the measurement of non-IT professionals' cybersecurity skills. *Information & Computer Security*, 27(1), 101-121.
- D'Arcy, J., Hovav, A. and Galletta, D. (2009). User awareness of security countermeasures and its impact on information systems misuse: a deterrence approach. *Information Systems Research*, 20(1), 79-98.
- European Union Agency for Cybersecurity (ENISA), (2019a). *Cybersecurity Skills Development in the EU*, The Certification of Cybersecurity Degrees and ENISA's Higher Education Database, December 2019. Retrieved from:
<https://www.enisa.europa.eu/publications/the-status-of-cyber-security-education-in-the-european-union>
- European Union Agency for Cybersecurity (ENISA), (2019b). *Cybersecurity Culture Guidelines: Behavioural Aspects of Cybersecurity*. Retrieved from:
<https://www.enisa.europa.eu/publications/cybersecurity-culture-guidelinesbehavioural-aspects-of-cybersecurity>
- Greitzer, F., Purl, J., Leong, Y. M., & Becker, D. S. (2018, May). Sofit: Sociotechnical and organizational factors for insider threat. In *2018 IEEE Security and Privacy Workshops (SPW)*, 197-206. IEEE.
- Heath, C. P., Hall, P. A., & Coles-Kemp, L. (2018). Holding on to dissensus: Participatory interactions in security design. *Strategic Design Research Journal*, 11(2), 65-78.
- Jose, I., LaPort, K., Trippe, D. M. (2016). Requisite attributes for cyber security personnel and teams. Cyber risk mitigation through talent management. In Stephen J. Zaccaro, Reeshad S. Dalal, Tetrick Lois E., Julie A. Steinke (Eds.): *Psychosocial Dynamics of Cyber Security*. New York, London: Routledge, 167-193.
- Kraemer, S., Carayon, P., Clem, J. (2009). Human and organizational factors in computer and information security: Pathway to vulnerabilities. *Computers & Security*, 28(7), 509-520.
- Legg, P. A., Buckley, O., Goldsmith, M., & Creese, S. (2015). Automated insider threat detection system using user and role-based profile assessment. *IEEE Systems Journal*, 11(2), 503-512.
- Mueller-Hanson, R., Garza, M. (2016). Selection and staffing of cyber security positions. In Stephen J. Zaccaro, Reeshad S. Dalal, Tetrick Lois E., Julie A. Steinke (Eds.): *Psychosocial Dynamics of Cyber Security*. New York, London: Routledge, 194-216.
- Oltamari, A., Henshel, D. S., Cains, M., & Hoffman, B. (2015). Towards a Human Factors Ontology for Cyber Security. In *STIDS*, 26-33.
- Renaud, K., Searle, R., & Dupuis, M. (2021, October). Shame in cyber security: effective behavior modification tool or counterproductive foil?. In *New Security Paradigms Workshop*, 70-87.

- Ruf, L.; Thorn, A.; Portmann, R.; Consecorn, A.; and Christen, T. (2008). Threat Modeling in Security Architecture – The Nature of Threats. *Information Security Society Switzerland*.
- Siponen, M., & Vance, A. (2010). Neutralization: new insights into the problem of employee information systems security policy violations. *MIS quarterly*, 487-502.
- Stanton, J. M., & Stam, K. R. (2003). Information technology, privacy, and power within organizations: A view from boundary theory and social exchange perspectives. *Surveillance & Society*, 1(2), 152-190.
- Taekke, J. (2011). Digital panopticism and organizational power. *Surveillance & Society*, 8(4), 441-454.
- Whitman, M.E. and Mattord, H.J. (2018). *Principals of Information Security*. (6th ed.). Boston, MA: Cengage Learning.
- Zaccaro, S. J., Dalal, R. S., Tetrick Lois E. & Steinke, J. A. (2016) The Psychosocial Dynamics of Cyber Security. An Overview. In Zaccaro, S. J., Dalal, R. S., Tetrick Lois E. & Steinke, J. A. (eds.) (2016) Psychosocial Dynamics of Cyber Security. Routledge, New York, London.

Probabilistic Analysis of Security Protocols Using Probabilistic Timed Automata

Olga Siedlecka-Lamch¹[0000-0001-9820-6629], Sabina Szymoniak¹[0000-0003-1148-5691]

¹ Department of Computer Science, Czestochowa University of Technology, Czestochowa, Poland

olga.siedlecka@icis.pcz.pl, sabina.szymoniak@icis.pcz.pl

Abstract. The amount of information transferred over the Internet has snowballed recently, and data security is most important for all users. In network communication, security is guaranteed by security protocols; we assume that they use secure cryptography. However, with the current number of keys used, administrative problems, new technologies, and applications implemented on the fly, the question remains, is it so? Our work analyses security protocols in a situation where we assume a certain probability of keys leaking or compromising. We analyse the weak points of the protocols and the possibility of introducing time constraints using a model based on probabilistic time automata. Thanks to the implemented tools, we can visualise the model, conduct tests, and make measurements.

Keywords: security protocols analysis, probabilistic timed automata, model checking

1 Introduction

In 2021, WhatsApp alone recorded 100 billion messages per day (Dean, 2021). Being able to communicate globally is considered the most dominant feature of the 21st century. Communication benefits all fields, from industry to critical domains, but with it comes a price. No user is interested in an easily penetrable communication platform. This urge to secure digital data introduced security protocols (Needham et al., 1978; Lowe, 1996; Abadi, 1999). The idea is that instead of allowing malicious users to obtain the data directly, the security protocol would be the first and foremost defence line. If the use of the protocol prevents attacks, we can start secure communication. Nonetheless, if the protocol fails, data are compromised, and the consequences can be devastating. It should be strongly emphasised in the context of our life, which is half in the real world and half in the virtual world, that it is not only the blocking of the Intruder's activities is essential. It is extremely important to detect its very presence. In the long run, this mere presence can negatively affect us. Based on this, it is imperative to investigate ways to make the protocol impenetrable (Armando et al. 2005; Blanchet 2016; Basin et al. 2018; Dolev et al. 1981; Li et al. 2016; Szymoniak et al. 2021).

Security protocols constitute a series of steps performed to ensure the users' authenticity and information safety. These steps include several values and an encryption-decryption system specific to the protocol. Aiming to infiltrate a protocol would require some or all these values depending on the value's importance. The values are ranked based on the chain reaction that follows if that value is compromised. For example, let us consider a protocol with several significant values. While the knowledge of one by the invader cannot be enough to bypass the protocol, another might be extracted. Therefore, the probability of acquiring that revealing value is essential, as the likelihood of succeeding in the attack rises with it. From this angle, we can see the benefits of including probability in the protocol verification methods (Dufлот et al. 2010; Kwiatkowska et al. 2012).

Time can equally be utilised as a parameter to check the users' honesty (Jakubowska et al. 2007; Szymoniak et al. 2021). Knowing the average execution time of each step might be enough to eliminate some slow or fast attacks. An attacker's primary goal is to deceive the users into believing nothing peculiar happens. He needs to run the protocol and satisfactorily complete it to infiltrate invisibly. The Intruder may need to perform encryption and decryption several times above what the standard provides. This Intruder behaviour causes the attack to take longer duration than usual. Similarly, the infiltrator might possess the values requested because of prior knowledge. This leak causes the attack to require less duration than expected. From this point, we can perceive how adding time into the formal verification method can benefit the protocol studied and its obscurity analysis.

1.1 Related Works

Our research stems from the bounded model checking (Henzinger et al. 2018; Baier and Katoen, 2008) and their use for security protocols (Basin et al. 2018, Lomuscio et al. 2008). Many tools allow writing protocol steps in dedicated specification languages, describing the examined features, and answering whether the protocol guarantees the specified features (Armando et al. 2005; David et al. 2015; Cremers 2008). The studies mentioned above most often concern the timeless version. The timed version that interested us appears in a smaller number of studies (Jakubowska and Penczek 2007; Szymoniak et al. 2019), but it indicates that time in the protocols is crucial (the authors of new protocols even intuitively use timestamps). Another aspect: resigning from the keys' security and accepting even the minimum probability of their taking over (not necessarily breaking) is just beginning to be examined (Ouchani et al. 2012; Mitchel et al. 2006). Finally, although there are tools for probabilistic analysis and modelling (Kwiatkowska et al. 2011; Alfaro 2000), they are not dedicated to security protocols. They do not allow quick and straightforward application in this field and do not provide easy-to-analyse results.

1.2 Our Contribution

The current security protocols approach suggests rewriting - e.g. by double encrypting or dropping the protocol if weaknesses are found.

Our approach allows us to add time constraints tailored explicitly to the protocol.

Encryption is considered secure with an appropriate key length, forgetting that even secure key management can be a challenge nowadays. Our solution allows us to determine which keys are essential for the security of the protocol under test.

The secure version of Andrew protocol (Boonkrong, 2014) presented in the article has not been analysed so far in terms of time constraints and the probability of taking over the keys.

1.3 The Work Structure

The following section will present what security protocols are, give an example, and define what types of attacks we can detect using our model. Then, in the section on probabilistic model-checking, we will discuss the reachability problem, probabilistic and time aspects. We will also present the definition of the probabilistic timed automata. Farther, we will present the proposed model in detail on the previously mentioned example of the Andrew protocol. Finally, we will briefly describe our tool and present the experimental results.

2 Security Protocols

Security protocols are algorithms sewn inside large communication protocols. Despite their size, they decide to ensure authentication, set keys for subsequent messages and data security, and further communication. Their correct design, implementation and verification determine our security in the network.

As a definition, a protocol is a sequence of operations performed by the users who want to communicate to prove their authenticity or the users who want to access restricted data (Dolev and Yao 1981). When user A (or Alice) wants to commune with user B (or Bob), they start performing the protocol steps by sending specific words to each other. Once A is sure he is talking to B and vice versa, they can set some additional keys and begin conversing with each other. This communication line initiated by the security protocol is now secure, and Alice and Bob can use it indefinitely until Alice or Bob closes it. These protocols are considered the first line of defence against intruders. If the Intruder does not trick the protocol, the protocol succeeds, and the communication is secure. The opposite is also true.

2.1 Andrew Secure Protocol as an Example

We will show our method on the example of the Lowe modified BAN concrete Andrew Secure Protocol [cite{L96some}](#) in the shortcut the Andrew Secure Protocol or the ASP. We chose the ASP protocol because of its compactness, readability and ability to fit the model in a small drawing format. More elaborate protocols would be difficult to depict in this article format. It does not change the fact that the whole method applies to all security protocols with similar elements (keys, nonce, timestamp).

Alice and Bob carry out four steps to authenticate each other and set up a new nonce and key combination for future communications. By manipulating the nonce N_a and the key K_{ab} , the users set up the new nonce N_b and the key K'_{ab} . The ASP procedure is as follows:

1. $A \rightarrow B : A, N_a$
2. $B \rightarrow A : \{N_a, K'_{ab}, B\}K_{ab}$ (1)
3. $A \rightarrow B : \{N_a\}K'_{ab}$
4. $B \rightarrow A : N_b$

In this protocol, Alice wants to communicate with Bob. She generates a nonce N_a and sends it to Bob. In return, Bob encrypts N_a , K'_{ab} , and his identity B using the key K_{ab} . Alice decrypts the message, acquires K'_{ab} , and sends N_a encrypted using K'_{ab} . At this point, Both Bob and Alice know they are honest. Bob generates and sends the new nonce N_b to Alice, and the communication line is securely open.

2.2 Types of Attacks on Security Protocols

There are many types of attacks against communication protocols, and there are also different types of intruders (Dolev and Yao 1981). In the case of security protocols, attacks on identity are interesting (the Intruder manages to impersonate a communication participant). An equally important attack is the attack on the secrecy of information. In our research, we also consider the possibility of obtaining the keys.

In turn, the man-in-the-middle behaviour is based on the positioning of the Intruder inside the communication. In this case, all information passes through his hands. This behaviour does not always allow for a full attack, but it is disordered behaviour. The Intruder considered in the research listens in the network and, if possible, joins the communication pretending to be an ordinary user or impersonating one of the websites. Unlike honest users, the Intruder may use the same nonce multiple times in different sessions and may collect single information and full ciphertexts.

$$\begin{aligned}
 \alpha^1_1 \quad A &\rightarrow I(B) : A, N_a \\
 \alpha^2_1 \quad I(A) &\rightarrow B : A, N_a \\
 \alpha^2_2 \quad B &\rightarrow I(A) : \{N_a, K'_{ab}, B\}K_{ab} \\
 \alpha^1_2 \quad I(B) &\rightarrow A : \{N_a, K'_{ab}, B\}K_{ab} \quad (2) \\
 \alpha^1_3 \quad A &\rightarrow I(B) : \{N_a\}K'_{ab} \\
 \alpha^2_3 \quad I(A) &\rightarrow B : \{N_a\}K'_{ab} \\
 \alpha^2_4 \quad B &\rightarrow I(A) : N_b \\
 \alpha^1_4 \quad I(B) &\rightarrow A : N_b
 \end{aligned}$$

In the example above, one can see an Intruder performing two simultaneous ASP executions (α^1 and α^2). Once with Alice, where the Intruder is impersonating Bob ($I(B)$). The second time with Bob, the Intruder impersonates Alice ($I(A)$). As a result, the Intruder gains only explicitly transmitted nonces. However, assuming that the keys can be compromised or leaked, such behaviour poses a real threat in the long run.

3 Probabilistic Model-checking

Model-checking is a technique used in almost all fields. Any control-based system is eligible for a model that can assist in testing it. Here is a non-exhaustive list of domains where model checking is sound: aerospace, railways, medical tools, factory automation, security protocols, plants. We are modelling the probabilistic and timed behaviour of the protocols.

Probabilistic model checking quantifies unpredictable behaviour and analyses the consequential properties of the system (Kwiatkowska and Parker 2012; Duflot et al. 2010; Ouchani et al. 2012; Jovanovic and Kwiatkowska 2018). It means that actions are not needed to follow a transition. Instead, a certain probability of transitioning to the next state represents a transition's likelihood. In the case of security protocols, probabilistic model checking can answer questions like: "What is the likelihood that honest users are performing this protocol execution?" This probability represents the ultimate goal of this paper. If we can lay all the information accurately, analyse it, and deduce an exact possibility of intrusion, this likelihood will represent the protocol's strength. With the protocol's strength known, constraints can be added, and their efficiency can be quantified.

3.1 Reachability Analysis of Probabilistic Model

We will show our method on the example of the LoweTypically, we stage model checking to test the qualitative properties of a system. Only lately, model checking found its way into examining quantitative properties in the shape of probabilities (Kwiatkowska et al. 2006; Kwiatkowska and Parker 2012; Duflot et al. 2010; Ouchani et al. 2012; Jovanovic and Kwiatkowska 2018). Probabilistic reachability is based on this novel model checking feature.

Probabilistic reachability is a probabilistic answer to whether a state is reachable or not. For security protocols, it allows us to inspect when the Intruder gains all the information necessary. We can also examine how long it takes to perform a step theoretically and cross-reference it with the actual running duration. An example of a probabilistic reachability property would be: "99% of the times, the invader does not manage to uncover all the information necessary".

Expected reachability is just as crucial for the analysis. Expected reachability allows us to compute the expected cost accumulated on the path towards a particular state (Stoelinga 2002). We model the trespasser's accumulated knowledge at each step and for each leaked key up to this point. A state in which the Intruder has obtained all the information necessary is a state to circumvent. After uncovering the probability of reaching these states, the analysis aims to add constraints to lower these specific probabilities.

Timed model-checking analysis - analysing the security protocol execution times may also be beneficial for its safety. Time properties allow us to discard odd times

and set constraints on standard times. Usually, time is calculated in terms of FLOPS because different hardware performs actions at different speeds. This speed difference implies that time is not entirely proportionate to the number of execution statements. Nonetheless, time does decrease or increase when large amounts of statements are omitted or added. This difference can form the tiebreaker between accepting an execution or rejecting it.

Probabilistic time-bounded reachability properties - these properties respect a specific deadline alongside a probabilistic minimum, or maximum (Kwiatkowska et al. 2006; Jovanovic and Kwiatkowska 2018). An example would be "94% of the cases, the run time of the protocol does not exceed 3 seconds." If we consider all run times below three seconds, to be honest, then the protocol is 94% secure.

Bounded response properties allow us to calculate the run time between a state and an inevitable one further down the path (Kwiatkowska et al. 2006; Duflot et al. 2010; Jovanovic and Kwiatkowska 2018). This type of property would be of the form "12% of the cases, and the Intruder will secure access to all the information necessary within 6 seconds." This probability entails that in 12% of the protocol executions, we reach a dangerous state in the Probabilistic Timed Automata (PTA) used to model the protocol.

4 Analysis of the Protocols Using the Probabilistic Timed Automata

A probabilistic timed automaton (PTA in short) is an automaton where transitions follow a probability distribution, and clocks enforce time constraints. As a formal definition, PTA formalises modelling systems whose behaviour incorporates both probabilistic and real-time characteristics. A system with Real-time features is a system that's bound by time constraints. Tracking the protocol running duration and flagging the suspicious run times is the basis for enhancing the protocol security. This feature of our analysis means that our protocol analysis does incorporate real-time characteristics. When analysing, we focus on what information the Intruder obtains if one of the probabilistic options becomes true. For model checking the protocols, these options represent each Intruder gaining access to one of the private keys. Since we analyse what would happen if each of the keys is compromised, our work is by nature probabilistic. We will be using PTA to perform the model checking of the security protocols.

4.1 Probabilistic Timed Automata (PTA)

To define PTA, we first need to define the time or actually clocks. By T we denote the time domain containing only non-negative real numbers. C will be a finite set of clock symbols. The values of the clocks increase with time and take the values from the set T . The clocks are evaluated by the function $v: C \rightarrow T$. T^C is the set of all evaluations.

We denote the time increment on all clocks as $v + t$. For any clock $c \in C$, the expression $v[c := 0]$ means resetting this particular clock, $v[C := 0]$ - resetting of all clocks (in short: 0). We restrict clocks to ensure reality. For example, they cannot go back. The constraints imposed on the clocks are in the form of the formulas' atomic conjunctions: $true, false, c \sim t$, where $c \in C, t \in T$ and $\sim \in \{<; \leq; =; \geq; >\}$. We denote the limitations as $G(C)$.

The definition of probabilistic timed automata is similar to classic timed automata (Alur, Dill 1994; Kwiatkowska et al. 2002, Norman et al. 2013). The probabilistic timed automaton is a tuple $PTA = \langle Q; \Sigma; C; inv; \delta; q_0; F \rangle$, where:

- Q is a finite set of states,
- Σ is an input alphabet,
- C is a finite set of clocks,
- inv is the function $inv: Q \rightarrow G(C)$,
- δ is the function of transition probability,
- $q_0 \in Q$ is the initial state,
- $F \in Q$ is a set of final states.

We can present the state of a PTA as a pair $(q; v)$, where $q \in Q$ and $v \in T^C$. At the beginning of the automaton operation, all clocks are reset, and the automaton is in its initial state. Each state is associated with the selection of whether a transition is made or time continues. The transition occurs to a different or the same state with a certain probability, provided that time conditions are met.

4.2 Modelling the Security Protocol

Let us transfer the PTA model to the level of security protocol analysis. The automaton states will correspond to the executable protocol steps and information about the Intruder's knowledge. Hence the entire step $B \rightarrow A : \{N_a, K'_{ab}, B\}K_{ab}; I_{know}=\{N_a\}; I_{unknown}=\{K_{ab}, K'_{ab}, N_b\}$ will be understood as one state.

The alphabet of the automaton consists of keys that must be broken, or rather obtained, in order to get secret information (for example $\Sigma = \{K_{ab}, K'_{ab}\}$). The state transition probability is the probability of breaking/acquiring a key or keys.

Next are time conditions. First, the clocks are reset to zero in the transitions corresponding to the commencement of communication in a given execution, as well as in the steps at which time stamps are generated. Then, in subsequent transitions, the time conditions associated with timestamps' expiration date t_{life} , estimated encryption time t_{enc} , network latency t_{del} , generation time t_{gen} , and decryption time t_{dec} , will be checked. Hence, we can make a specific transition if the cumulative time of the operations envisaged in the step is between the minimum and maximum times ($c_i \geq \min(t_{gen} + t_{enc} + t_{del})$ and $c_i \leq \max(t_{gen} + t_{enc} + t_{del})$). We can treat generation, encryption, decryption and latency times as parameters that can be changed depending on the properties of the servers and the network.

We will consider different probabilities according to the Intruder's capabilities at a given step, assuming that the probability of acquiring individual keys increases over time and the probability of not discovering anything decreases. Hence the following will appear: $p_{nothing}$ - the intruder did not discover anything; p_{Kab} - the Intruder got the key K_{AB} .

The initiating state q_0 is the state before the commencement of communication, in which the Intruder does not yet know about confidential data, and the users have the necessary keys to start the communication.

Final states are states in which the Intruder has acquired confidential information. The detection of such states in the model is associated with information about the protocol's weak points. Thus, we will know for which keys these states are reached, with which time constraints.

4.3 An Example Showcasing the Analysis' Benefits

The example security protocol we will showcase now is the Lowe modified BAN concrete Andrew Secure protocol (Boonkrong, 2014). As we showed earlier, the first part of the procedure is to draw the probabilistic time automata modelling the protocol's behaviour if private keys are exposed. Following that, the second step requires calculating the probability that this particular protocol execution is honest or not. Ultimately, the protocol execution is discarded or accepted according to the safety threshold we set.

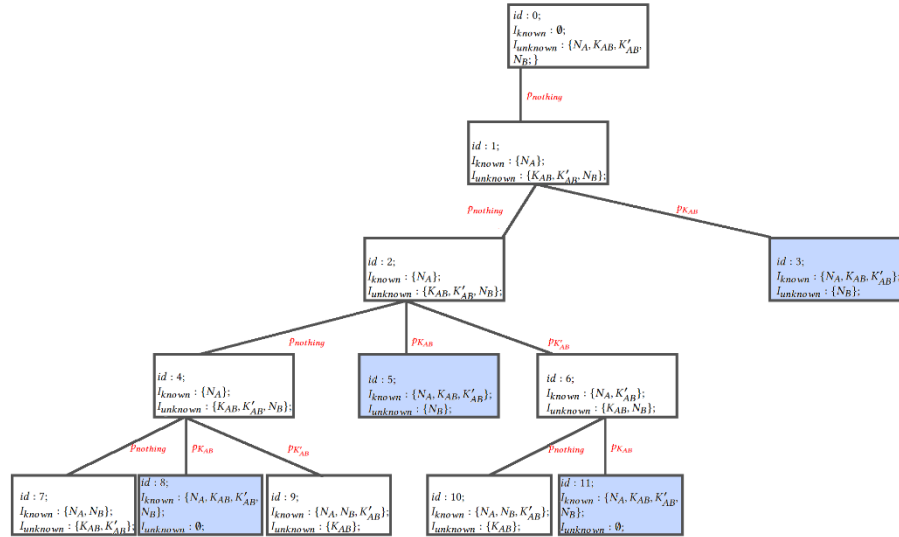


Fig. 1. The model of Andrew Security Protocol

For the ASP, the Intruder can succeed in his attack in two ways: getting N_A and K_{AB} or N_B and K'_{AB} . Nevertheless, for the second path to be possible, the Intruder must accomplish the protocol correctly. This dependency is because K'_{AB} is only

operational once the protocol execution is a success. Therefore, since the Intruder needs K_{AB} to perform the protocol correctly, we deduce that only the first path needs to be covered.

In figure 1, we see the model of the Andrew Secure Protocol. The figure shows the paths where the Intruder does not break any security or breaks/acquires individual keys with some probability: via red $p_{nothing}$, p_{KAB} , $p_{K'AB}$, respectively. States that are coloured in blue represents the situation in which the Intruder obtains all secret information. The execution should avoid these states. We coloured other states in white. Based on these blue states, we can see that the Intruder succeeds every time he obtains K_{AB} . This chain reaction is not equally valid for K'_{AB} . Hence, K_{AB} protection is more critical than K'_{AB} protection.

The next part of the procedure includes analysing the execution times of the protocol (figure 2). First, we adopt the notation to the Andrew protocol. The Andrew protocol possesses four steps, making $n = 4$. Two honest users, Alice and Bob, are performing the protocol, hence $m = 2$. Alice fulfils the first and the third step, while Bob conducts the second and the fourth.

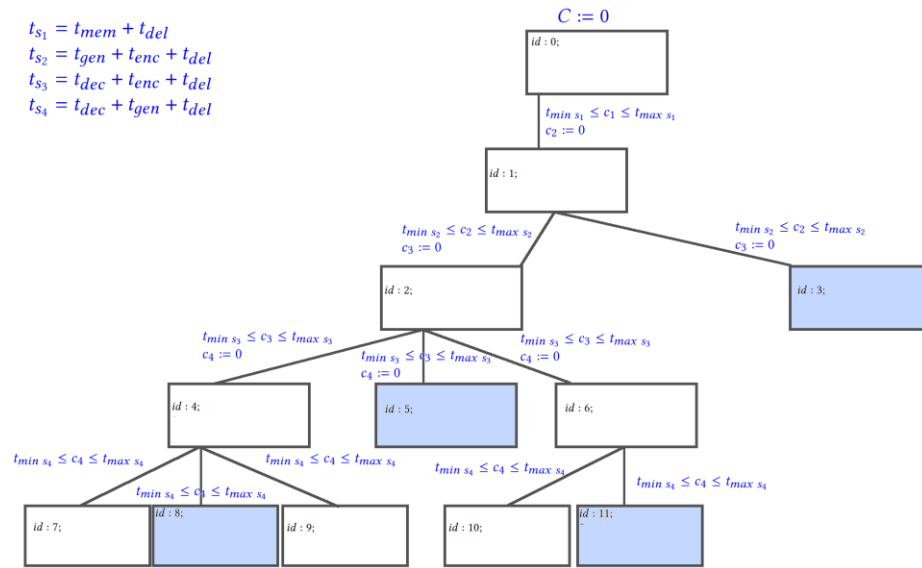


Fig. 2. Time dependencies in ASP

In the first step, Alice recovers N_A from the memory and sends the message to Bob. This step's duration is equal to $t_{step1} = t_{mem} + t_{del}$. In the second step, Bob generates the key K'_{AB} , encrypts the message and sends it to Alice. This step lasts $t_{step2} = t_{gen} + t_{enc} + t_{del}$. Similarly, step 3 and 4 duration are respectively:

$$t_{step3} = t_{dec} + t_{enc} + t_{del},$$

$$t_{step4} = t_{dec} + t_{gen} + t_{del}.$$

Since execution times are not proportionate, we do not expect the execution time of step 2 to be almost as equal to the step 4's time. Nonetheless, the encryption is the overwhelming time in the second step. If we omit the encryption, the step's execution time will surely drop. If this drop is enough to overcome the different hardware speeds, the execution would be suspiciously rejected.

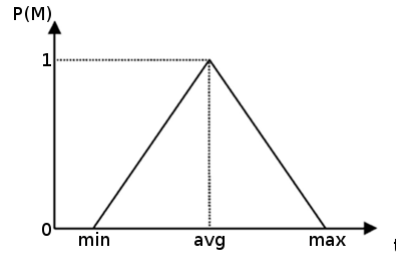


Fig. 3. Triangular function depicting the probability of honesty of an execution time

Previously, we save the minimum, average, and maximum execution time for each of the steps. We use these values to calculate the execution times of each step. The closer the execution time is to the average, the higher its chances of it being honest. If an execution time is above the maximum or below the minimum for particular step, its odds of being candid is zero. We depict this probability calculation in figure 3.

For each step s_i for $i \in \{1, 2, 3, 4\}$, we obtain an execution time t_{s_i} to test. To follow the membership function depicted in figure 2, we calculate the probabilities in this way:

$$P(t_{s_i}) = \begin{cases} 0, & t_{s_i} \leq t_{mins_i} \\ \frac{t_{s_i} - t_{mins_i}}{t_{avgs_i} - t_{mins_i}}, & t_{mins_i} \leq t_{s_i} \leq t_{avgs_i} \\ \frac{t_{maxs_i} - t_{s_i}}{t_{maxs_i} - t_{avgs_i}}, & t_{avgs_i} < t_{s_i} \leq t_{maxs_i} \\ 0, & t_{s_i} \geq t_{maxs_i} \end{cases} \quad (3)$$

At this point, we should have a probability of honesty for each of the steps' execution times. We record these execution times independently, one from the other. Therefore their probabilities are independent. User can group these probabilities. For user m to be straight, all the actions they performed must have a veritable execution time. Independent events (Meester 2008): Two events are separate if one event's incidence does not affect the other event's probability. $P(m) = \{\bigwedge_1^n P(t_{s_i})\}$ such that step i is achieved by user m .

For example:

$$P(Alice) = P(t_{s_1 Alice}) \cap P(t_{s_3 Alice}) = P(t_{s_1 Alice}) \times P(t_{s_3 Alice}),$$

$$P(Bob) = P(t_{s_2 Bob}) \cap P(t_{s_4 Bob}) = P(t_{s_2 Bob}) \times P(t_{s_4 Bob}).$$

We have the probability of honesty of each user and the likelihood that the total execution time is genuine. We also calculate the expectation that the total execution

time is truthful, denoted by $P(total)$. The users must be legitimate for the protocol procedure, and the full execution time must somewhat represent truthfulness. Represented by $P(protocol)$, we will then use this probability for the final output analysis.

$$P(protocol) = \{\bigwedge_1^m(P(m)) \cap P(t)\} \quad (4)$$

On our example:

$$P(protocol) = P(Alice) \cap P(Bob) \cap P(t) = P(Alice) \times P(Bob) \times P(t).$$

Finally, we compare $P(protocol)$ to the safety threshold. If $P(protocol)$ is equal to or higher than the safety threshold, the protocol execution is considered successful, and the output is 1. If $P(protocol)$ is less than the safety threshold S_T , the output is 0, and we discard the protocol execution.

$$Output = \begin{cases} 0, & P(protocol) < S_T \\ 1, & P(protocol) \geq S_T \end{cases} \quad (5)$$

5 Experiments

To visualise our model, we implemented a tool using Java programming language. This tool takes two inputs. The first input is the protocol itself. The second input is the combinations needed for the Intruder to succeed in his attack. For the ASP, the inputs are below:

1	N_A	n	A
2	N_A	K_{AB}	B
2	K'_{AB}	K_{AB}	B
3	N_A	K'_{AB}	A
4	N_B	n	B
	Combinations	K_{AB}	N_A

The notation used for the protocol implementation is different from the alphabet we used earlier. Therefore, we will detail this notation swiftly. Each line includes four different values. The first value represents the step number; the second represents the value obtained if the user possesses the third value. The last value is the person sending the message. In the second step, the user possesses N_A and K'_{AB} if he has K_{AB} . The last line represents the second input. If the Intruder has K_{AB} and N_A , he can call his attack a success. With the inputs shown and explained, the automata visualisation made by our tool is shown in figure 1.

We need some ASP execution times to develop the minimum, average, and maximum time for each step to test our theory. We have implemented an application simulating the execution of protocols. We recorded the execution times of each of the steps for one thousand runs. The calculated times are presented in table 1. These executions do not include the delay time. Since the delay time is constant, each user can automatically record the value and add it to the calculations.

Table 1. Minimum, Average, and Maximum Times Values for the ASP.

Step	t_{\min} [ns]	t_{avg} [ns]	t_{\max} [ns]
s ₁	1500	2640	6200
s ₂	70603500	83296850	99017900
s ₃	68583500	75312900	86040100
s ₄	68776100	78310540	85930900
Total	13221000	236922930	260880400

Safety thresholds ST have been established:

ST (s₁) = 2900ns,
 ST (s₂) = 86423440ns,
 ST (s₃) = 76249530ns,
 ST (s₄) = 77294532ns,
 ST (total) = ST (s₁) + ST (s₂) + ST (s₃) + ST (s₄) = 239970402ns,
 ST = 50%.

We calculate the probabilities of honesty. For this, we first check the conditions according to each step time and equation 3.

$$\begin{aligned}
 t_{avg s_1} &< S_T(s_1) < t_{max s_1}, \\
 t_{avg s_2} &< S_T(s_2) < t_{max s_2}, \\
 t_{avg s_3} &< S_T(s_3) < t_{max s_3}, \\
 t_{min s_4} &< S_T(s_4) < t_{avg s_4}, \\
 t_{avg total} &< S_T(total) < t_{max total}.
 \end{aligned}$$

Based on the conditions, we calculate the probabilities of honesty of the step according to equation 3.

P (s₁) = 0.9269662921348315,
 P (s₂) = 0.8011207902780031,
 P (s₃) = 0.9126864419419792,
 P (s₄) = 0.8934381043878823,
 P (t_{total}) = 0.8727965849482437.

Our next step is to calculate each user's probability of honesty according to the steps they performed and equation 4:

P (A) = P (s₁) × P (s₃) = 0.8460295669686887,
 P (B) = P (s₂) × P (s₄) = 0.7157518402517012.

The following is to calculate the probability of honesty of the entire protocol execution following the equation 4:

$$P(\text{protocol}) = P(A) \times P(B) \times P(t_{\text{total}}) = 0.5285195451741215$$

Since $P(\text{protocol})$ is bigger than the safety threshold of 50%, the protocol outputs 1. This output represents that the protocol execution is accepted. This chain of analysis is mounted by the program entirely. We first gathered the regular execution times recorded. Then, we input these execution times and the tested execution time to the second program. The output is 1 or 0 based on the honesty of the protocol performance.

5.1 Detection of Attacks and Attack Attempts

For testing purposes, we record the execution time for the attack performance shown on the equation 2 and calculate the probability of integrity. The execution times for this dispatch are:

$$\begin{aligned} ST(s_1) &= 2760\text{ns}, \\ ST(s_2) &= 81354720\text{ns}, \\ ST(s_3) &= 52576280\text{ns}, \\ ST(s_4) &= 79257460\text{ns}, \\ ST(\text{total}) &= 213191220\text{ns}, \\ ST &= 50\%. \end{aligned}$$

$ST(s_3)$ is below the minimum execution time calculated for that step, making the probability that the 3rd step is sincere equal to 0. This situation, in turn, reduces the likelihood that $I(A)$ is honest to 0. Hence, the odd of the protocol being straight also becomes 0. Thus, this attack is discarded automatically, and the protocol safety is enhanced.

6 Conclusions and Further Plans

Most research and analysis treat security protocols as purely deterministic. Actions are predefined, which take place in a specific order. At best, we can consider possible interleavings of different executions of the same protocol, which means that the analysed space increases significantly. The reality, on the other hand, is much more complex. The security protocol is embedded in much larger systems, which run on a network. Therefore, the protocol execution will be influenced by the hardware and its performance, network capacity, delays, the time of performing individual operations, and an Intruder's presence. Thus, we are dealing with a lot of probabilistic factors.

It is worth adding other protocol elements in further work, such as hashing functions. It is also worth introducing more time and probabilistic parameters. Tests can be carried out for different types of networks due to their load. Tests can also be carried out to fine-tune the function returning the probability of honesty of executions. This fine-tuning requires supervised learning and a specific analysis procedure. Our method generalises the analysis across all security protocols. On the other hand, fine-

tuning the function for each protocol based on actual data can reveal many concealed attacks. In the context of our social life on the Internet, it would also be beneficial to sensitise the methods, for example, by increasing the number of parameters so that the presence of an Intruder could be detected with greater precision.

We would also like to thank the student Koutayba Chaker especially for working on the experimental part and preparing the tool.

Acknowledgements

We would also like to thank the student Koutayba Chaker especially for working on the experimental part and preparing the tool.

References

- Abadi, M. (1999, March) *Security protocols and specifications*. In: Foundations of Software Science and Computation Structures, Second International Conference (FOSACS'99). LNCS, vol. 1578, pp. 1–13. Springer-Verlag
- Alfaro de, L., Kwiatkowska, M.Z., Norman, G., Parker, D., Segala, R. (2000) *Symbolic model checking of probabilistic processes using mtbdds and the kronecker representation*. In: Graf, S., Schwartzbach, M.I. (eds.) Tools and Algorithms for Construction and Analysis of Systems, 6th International Conference, TACAS 2000, Held as Part of the European Joint Conferences on the Theory and Practice of Software, ETAPS 2000, Berlin, Germany, March 25 - April 2, 2000, Proceedings. Lecture Notes in Computer Science, vol. 1785, pp. 395–410. Springer
- Armando, A., Basin, D., Boichut, Y., Chevalier, Y., Compagna, L., Cuellar, J., Drielsma, P.H., He am, P.C., Kouchnarenko, O., Mantovani, J., M'odersheim, S., von Oheimb, D., Rusinowitch, M., Santiago, J., Turuani, M., Vigan'o, L., Vigneron, L. (2005) *The avispa tool for the automated validation of internet security protocols and applications*. In: Etessami, K., Rajamani, S.K. (eds.) Computer Aided Verification. pp. 281–285. Springer Berlin Heidelberg, Berlin, Heidelberg
- Baier, C., Katoen, J.P. (2008) *Principles of Model Checking* (Representation and Mind Series). The MIT Press
- Basin, D., Cremers, C., Meadows, C. (2018) *Model Checking Security Protocols*, pp. 727–762. Springer International Publishing, Cham
- Blanchet, B.. (2016, Oct.) *Modeling and verifying security protocols with the applied pi calculus and proverif*. Found. Trends Priv. Secur. 1(1–2), 1–135
- Boonkronk, S. (2014) *A more secure and efficient Andrew Secure RPC Protocol*. Security and Communication Networks 7(11), 2063–2077
- Cremers, C.J.F. (2008) *The scyther tool: Verification, falsification, and analysis of security protocols*. In: Gupta, A., Malik, S. (eds.) Computer Aided Verification, 20th International Conference, CAV 2008, Princeton, NJ, USA, July 7-14, 2008, Proceedings. Lecture Notes in Computer Science, vol. 5123, pp. 414–418. Springer

- David, A., Larsen, K.G., Legay, A., Mikucionis, M., Poulsen, D.B. (2015) *Uppaal SMC tutorial*. Int. J. Softw. Tools Technol. Transf. 17(4), 397–415
- Dean, B. (2021, Oct. 19). *WhatsApp 2021 User Statistics: How Many People Use WhatsApp?* Retrieved from <https://backlinko.com/whatsapp-users>
- Dolev, D., Yao, A.C. (1981) *On the security of public key protocols*. In: Proc. of the 22nd Annual Symp. on Foundations of Computer Science. p. 350–357, SFCS '81, IEEE Computer Society, USA
- Duflot, M., Kwiatkowska, M., Norman, G., Parker, D., Peyronnet, S., Picaronny, C., Sproston, J. (2010) *Practical Applications of Probabilistic Model Checking to Communication Protocols*, FMICS Handbook on Industrial Critical Systems, pp. 133–150. IEEE Computer Society Press
- Henzinger, T., Veith, H. (2018) *Handbook of Model Checking*. Springer. <https://doi.org/10.1007/978-3-319-10575-8>
- Jakubowska, G., Penczek, W. (2007) *Modelling and checking timed authentication of security protocols*. Fundamenta Informaticae 79(3-4), 363–378
- Jovanovic, A., Kwiatkowska, M. (2018) *Parameter synthesis for probabilistic timed automata using stochastic game abstractions*. Theor. Comput. Sci. 735, 64–81
- Kwiatkowska, M.Z., Norman, G., Parker, D. (2011) *PRISM 4.0: Verification of probabilistic real-time systems*. In: Gopalakrishnan, G., Qadeer, S. (eds.) Computer Aided Verification - 23rd International Conference, CAV 2011, Snowbird, UT, USA, July 14-20, 2011. Proceedings. Lecture Notes in Computer Science, vol. 6806, pp. 585–591. Springer
- Kwiatkowska, M.Z., Norman, G., Parker, D., Sproston, J. (2006) *Performance analysis of probabilistic timed automata using digital clocks*. Formal Methods Syst. Des. 29(1), 33–78
- Kwiatkowska, M.Z., Parker, D. (2012) *Advances in probabilistic model checking*. In: Nipkow, T., Grumberg, O., Hauptmann, B. (eds.) Software Safety and Security – Tools for Analysis and Verification, NATO Science for Peace and Security Series - D: Information and Communication Security, vol. 33, pp. 126–151. IOS Press
- Li, L., Sun, J., Dong, J. S. (2016) Automated verification of timed security protocols with clock drift, in: J. S. Fitzgerald, C. L. Heitmeyer, S. Gnesi, A. Philippou (Eds.), FM 2016: Formal Methods - 21st International Symp., LNCS Vol. 9995, pp. 513–530
- Lomuscio, A., Penczek, W. (2008) *LDYIS: a framework for model checking security protocols*. Fundam. Informaticae 85(1-4), 359–375
- Lowe, G. (1996) *Breaking and fixing the needham-schroeder public-key protocol using fdr*. In: Margaria, T., Steffen, B. (eds.) Tools and Algorithms for the Construction and Analysis of Systems. pp. 147–166. Berlin Heidelberg: Springer
- Meester, R. (2008) *A natural introduction to probability theory*, second edition. Birkhauser Verlag
- Mitchell, J.C., Ramanathan, A., Scedrov, A., Teague, V. (2006) A probabilistic polynomial-time process calculus for the analysis of cryptographic protocols. Theor. Comput. Sci. 353(1-3), 118–164

- Needham, R.M., Schroeder, M.D.. (1978, Dec.) *Using encryption for authentication in large networks of computers*. Commun. ACM 21(12), 993–999
- Ouchani, S., Jarraya, Y., Mohamed, O.A., Debbabi, M. (2012) *Probabilistic attack scenarios to evaluate policies over communication protocols*. Journal of Software 7(7), 1488–1495
- Stoelinga, M. (2002) *An introduction to probabilistic automata*. Bull. EATCS 78,176–198
- Szymoniak, S., Siedlecka-Lamch, O., Kurkowski, M. (2019) *Network’s delays in timed analysis of security protocols*. In: Information Systems Architecture and Technology: Proceedings of 39th International Conference on Information Systems Architecture and Technology – ISAT 2018. pp. 19–29. Springer International Publishing, Cham
- Szymoniak, S., Siedlecka-Lamch, O., Zbrzezny, A.M., Zbrzezny, A., Kurkowski, M. (2021) *Sat and smt-based verification of security protocols including time aspects*. Sensors 21(9)

Choosing the Right Cybersecurity Solution: A Review of Selection and Evaluation Criteria

Rafał Leszczyna¹[0000-0001-7293-2956]

¹ Gdańsk University of Technology, Faculty of Management and Economics, Gdańsk, Poland
rle@zie.pg.edu.pl

Abstract. Information technologies evolve continuously reaching pioneering areas that bring in new cybersecurity challenges. Security engineering needs to keep pace with the advancing cyberthreats by providing innovative solutions. At the same time, the foundations that include security and risk assessment methodologies should remain stable. Experts are offered with an extensive portfolio of solutions and an informed choice of a particular one becomes problematic. Transparent criteria are the instrument that answers this issue by laying the ground for evidence-based justifications. Within the framework of systematic literature analysis, this study reviews the criteria proposed in the relevant literature. Based on the outcome, a consolidated set of criteria that should help in choosing a cybersecurity solution is proposed. Ethical questions posed by certain cybersecurity assessment activities are discussed. Consequently, new criteria related to the ethical application of a solution are introduced in the consolidated set.

Keywords: cybersecurity, management, solutions, methods, ethics, organisation management

1 Introduction

Information technologies evolve continuously demarcating new areas of applications and solutions. Edge computing, Software 2.0, digital twins, remote working at a massive scale, broad AI application or 5G are only selected examples of emerging technology trends (Accenture, 2021b; Deloitte Insights, 2021; Duggal, 2021; McKinsey & Company, 2021). Together with great opportunities, these changes bring in new challenges. Innovative hardware and software architectures open new paths for cyberattacks. The contemporary cyberthreat landscape is marked by phenomena such as cybercrime as a service (ENISA, 2021), commodity malware (Accenture, 2021a), and the ransomware crisis (Accenture, 2021a; Sophos, 2021) or refined supply chain infringements (ENISA, 2021). Cybersecurity needs to keep pace with the advancing cyberthreats by providing innovative solutions. The developments include situational awareness (Alcaraz & Lopez, 2013; Bolzoni et al., 2016; Tadda & Salerno, 2010) and threat intelligence platforms (Leszczyna & Wróbel, 2019), facilitated cyberincident

information sharing (Leszczyna et al., 2019) or embedding artificial intelligence into defensive mechanisms.

In parallel, there are elements of cybersecurity that independently of the complexity or innovativeness of the technology shall remain stable. One of them is *risk assessment* – the process devoted to the identification, analysis and evaluation of cybersecurity risks (ISO/IEC, 2018; NIST, 2011; Wangen et al., 2018) – that constitutes the central part of cybersecurity management (ISO/IEC, 2013). Also, *cybersecurity assessment* should be a solid and steady component. It investigates the cybersecurity state of an assessed entity (Dalalana Bertoglio & Zorzo, 2017; Qassim et al., 2019; Rogers & Syngress Media, 2004) and determines how effectively the entity fulfils specific security objectives (Scarfone et al., 2008).

Experts are offered an extensive portfolio of solutions, both the innovative ones and related to the fundamental risk and cybersecurity assessment activities (Gritzalis et al., 2018; Ionita & Hartel, 2013; Leszczyna, 2021; Wangen et al., 2018). In this regard, an informed choice of a particular solution becomes challenging. Transparent criteria are the instrument that answers this issue by laying the ground for evidence-based justifications. *Selection criteria* enable identifying the methods applicable to a specific area, while *evaluation criteria* facilitate comparing the identified methods and eliciting the most suitable one.

This study reviews the criteria proposed in the relevant literature during a systematic process that implements the (Webster & Watson, 2002) and (Kitchenham & Brereton, 2013) guidelines. Based on the outcome a consolidated set of criteria that should help in choosing a cybersecurity solution is proposed. Because cybersecurity management and cybersecurity assessments, in particular, may include activities that pose ethical questions, the collection is extended with criteria related to the ethical application of a solution.

The main contributions of the research are as follows:

- Criteria proposed in the relevant literature are identified during a systematic review process.
- A consolidated set of criteria that should facilitate informed decisions on the choice of cybersecurity solution is proposed.
- The collection is extended with criteria related to the ethical application of a solution.

The paper is organised as follows. Section 2 presents the research method applied in the study. The outcome of the analysis i.e. the identified criteria are presented in Section 3. After that, the ethical questions related to cybersecurity management are discussed (Section 4) and the consolidated set of criteria is introduced (Section 5). The paper closes with concluding remarks.

2 Research method

This study adopts the approaches of Webster & Watson, 2002 and Kitchenham & Brereton, 2013 to systematic literature surveys. Its main components are presented in Figure 1. During the *literature search*, relevant publications were searched for using

the keywords “security assessment”, “review” and “survey”. To reduce the number of results this step was repeated several times. Also, selection and evaluation criteria were applied to facilitate the process. Depending on the functionalities provided by a search engine, the initial iterations focused on titles, abstracts, keywords or other metadata. Then, the descriptions of the publications were read (*manual search*), to finally browse the contents of the documents in the concluding iteration (*in-depth analysis*). When possible, the search was restricted to computer science or a cognate domain.

The literature sources included journals, books and the databases of established publishers that address the topics of cybersecurity, communication systems, computer science and related i.e. the ACM Digital Library, Elsevier, Emerald, IEEE Xplore, Springer and Wiley. Also, collective databases that contain records of various publishers – EBSCOhost, Scopus and Web of Science were utilised. In addition to that, the search was completed with a short search of conference proceedings and the Internet. When discovered papers mentioned other relevant articles, also the latter were subject to the analysis (*backward analysis* (Webster & Watson, 2002)).

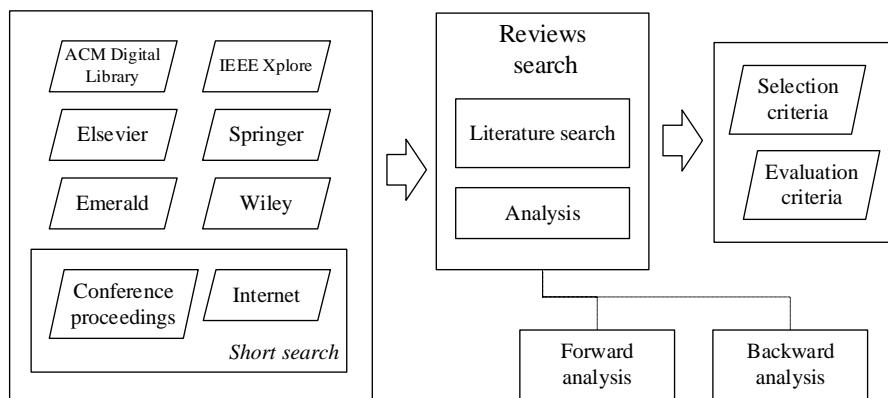


Fig. 1. The key tasks and data sources employed during the review process.

3 Identified criteria

Qassim et al., 2019 apply basic criteria to depict a standard or a guideline as most relevant to the topic of the study: availability free of charge, English documentation, publication by a standard body or governmental agency, implementation or application in the context of industrial control systems and the presence of detailed descriptions. The documents’ analysis criteria regard the method’s coverage of cybersecurity management processes, the assessment mode (active or passive), and compliance to the NERC CIP requirements.

Cherdantseva et al., 2016 took advantage of an adaptation of the literature review approach of Kitchenham & Brereton, 2013 to obtain a structured search. As inclusion criteria, the authors utilised the method’s coverage of risk assessment processes, its industrial control systems-specific design and scientific, cybersecurity-related origin.

The analysis criteria embraced the method's aim, its application domain, addressed phases and concepts of risk management, the impact assessment scheme, sources of data for deriving probabilities, the method's evaluation procedure and availability of supporting tools. Based on the analysis criteria a series of methods' categorisations have been introduced, comprising, for instance, formula-based and model-based, low-level and high-level detail coverage or qualitative and quantitative, the latter including probabilistic, non-probabilistic and undefined. Also, a content coverage-related criterion is applied by Hahn & Govindarasu, 2011 who compared assessment methods based on the coverage of 13 NERC CIP requirements relevant to cybersecurity assessment. Shahriar & Zulkernine, 2009 defined seven criteria for comparing testing methods. The criteria encompass vulnerability coverage, source of test cases, test generation method, test level, the granularity of test cases, tool automation and application domain.

A structured method for comparing and evaluating risk assessment techniques is applied by Gritzalis et al., 2018. The authors thoroughly analysed the criteria utilised in earlier studies to derive a set of common criteria that were discussed by industry experts and contain a strong practical component. The criteria are categorised into four groups, namely validity, compliance, cost and usefulness. They are presented in Table 1. In addition, as selection criteria, the presence of a risk assessment method description, the focus on information security risks and the availability of English documentation are used.

Table 1. Gritzalis et al., 2018 evaluation and comparison criteria for risk assessment methods.

Validity	Compliance	Cost	Usefulness
Completeness			Ease of use
Preparation (1)			
Risk identification (2)			Usability (Interface,
Risk analysis (3)			handle errors,
Risk evaluation (4)			documentation)
Type of analysis			Scope
Qualitative	Compliance	Support cost	Target organisation
Quantitative	with standards	Software cost	(type, size), Focus
Risk calculation class			Life cycle
Class A			Release
Class B			Last update
Class C			Adaptability
Class D			Software support
Class E			Training

Wangen et al., 2018 developed a dedicated framework for comparing risk assessment methods and evaluating their completeness as far risk assessment

constituent tasks are concerned. The development process was bottom-up and incremental, based on extracting and combining tasks specified in the methods. Consequently, more than 40 tasks and 10 associated concepts are distinguished, coverage of which pertains to the frameworks' criteria. The tasks are grouped into three descriptive categories related to the incumbent stages of risk assessment (risk identification, estimation and evaluation). As selection criteria, the large number of citations (exceeding 50), the coverage of particular risk subjects (incentives risk, cloud risk and privacy risk), the coverage of three compulsory risk assessment processes, timeliness (published during the last 15 years) and availability of English or Norwegian documentation were utilised.

Ionita & Hartel, 2013 advocate deriving selection criteria directly from the scope and assumptions of the research. As a result, introduced inclusion criteria include the presence of a method description that comprises all obligatory risk assessment stages, application to an existing system or a system design, specific audience (chief security officers or other decisive personnel), the focus on information security risks, the availability of comprehensive English documentation as well as practical application in more than one country. Exclusion criteria embraced certification purpose, orientation towards a concrete product or system and only high-level (managerial or governance) specifications. When comparing risk assessment methods the authors apply the following criteria: method class (from one to five), method type (quantitative or qualitative), sponsor, focus, supported risk assessment phases, release date, price, type of target users (management, operational or technical), required skills, availability of supporting tools (paid, free), availability of a standalone version and the target organisation type (government agency, large company, small or medium enterprise). Besides that, a categorisation of methods is proposed based on risk measurement (quantitative and qualitative), risk model (five classes) and goal (e.g. certification, audit or internal control).

Felderer et al., 2016 performed an extensive analysis of existing evaluation criteria and classifications of security testing methods to construct a taxonomy of model-based security testing techniques. Based on the research, the authors proposed a structured set of classification criteria presented in Figure 2. The criteria are divided between filter and evidence types. The former validated the existence of a system security model, an environment security model and explicit test selection criteria. The latter focused on examining the maturity of the assessment object, evidence measures and the evidence level. The selection criteria used in the study embraced the documentation in a peer-reviewed paper written in English and the coverage of a model-based security testing approach. Earlier on, Felderer & Schieferdecker, 2014 presented a compound taxonomy focused on risk-based testing that distinguishes three top-level classes related to risk drivers, risk assessment and the risk-based test process. It contains more than 40 classification criteria. Giannopoulos et al., 2012 propose analysis criteria that result from the practice of conducting multiple impact assessments, namely the method's scope, objectives, target users, applied techniques and standards, the coverage of interdependencies, addressing of cross-sectoral risks and relevance to resilience.

To compare risk analysis methods, Meriah & Rabai, 2018 use the following criteria: purpose, inputs, outcome, the structure of the security management process, supporting

tools and type of system application. Fabisiak et al., 2012 proposed a substantial set of comparison criteria that are used to evaluate methods that support various aspects of cybersecurity management, including cybersecurity assessment and risk assessment. The criteria are presented in Table 2. Shah & Mehtre, 2015 classify vulnerability discovery techniques into manual testing, automated testing, static analysis and fuzz testing. A taxonomy of automated cybersecurity assessment based on D³ (Discovery, Description, and Detection) approach is proposed by Barrere et al., 2014. Steffen Weiss (Weiss, 2008) introduces a basic categorisation of cybersecurity assessment methods into measurement approaches and combines approaches depending on the breadth of evaluation (components – organisation), as well as based on the meticulousness' level of measurements – algorithmic approaches and guidelines. For instance, algorithmic measurement approaches include vulnerability analysis.

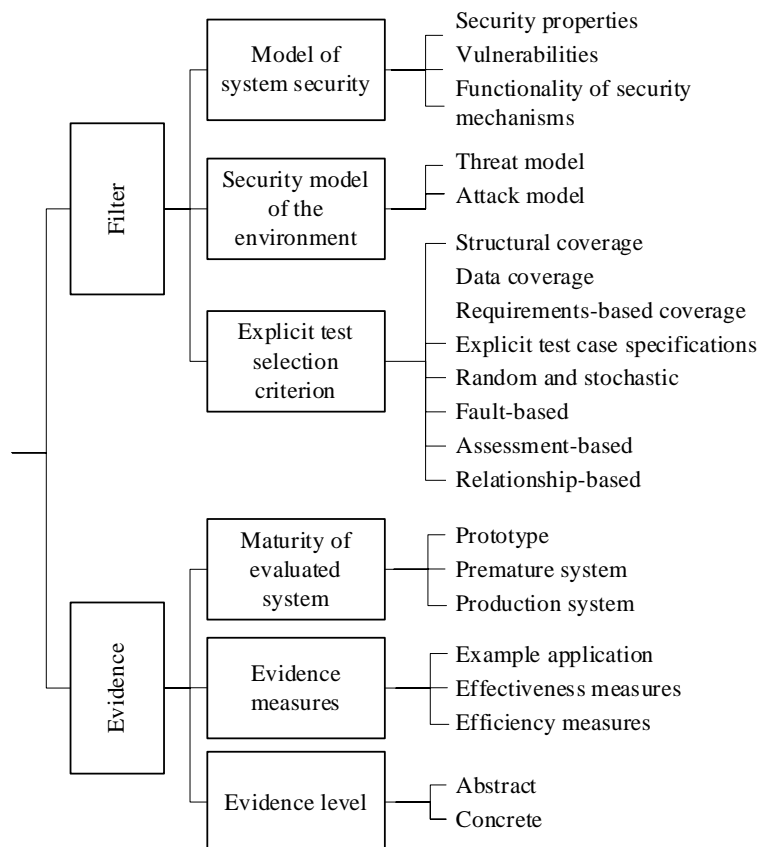


Fig. 2. Classification criteria for security testing methods proposed by (Felderer et al., 2016).

The study of Li et al., 2019 focuses on the software used in vulnerability identification. Thus, selection criteria, besides the availability free of charge, are strictly

technical. They refer to stand-alone, self-contained operation, detection of Java code vulnerabilities and identification of security weaknesses extending beyond code bugs. Similarly, the metrics applied to compare the applications are software-oriented. They comprise the vulnerability coverage (the number of detected flaws), the recall, precision, and discrimination rates as well as 15 usability measures including the tool output quality, the averaged false positive rate or extendability. The analogous study of Holm et al., 2011 concentrated on the tools' functionality and accuracy. 12 metrics associated with the former attribute embraced software flaws detection, configuration errors detection, scanning mode (active or passive), ports coverage, patch deployment ability and others. The latter property was linked to the number of false negatives. Lykou et al., 2019 compare vulnerability identification tools using 13 metrics, namely the tool type, its developer, origin, description, the number of stages in the evaluation process, survey method, required security expertise, standards compliance, presence of standards' compliance checking functionality, database of industry available cybersecurity practices, sector average score, presence of a recommendation list and the type of result.

Table 2. Comparison criteria for cybersecurity assessment, risk assessment and cybersecurity management methods introduced by Fabisiak et al., 2012.

Cost	Number of standard scenarios	Risk metric
English documentation	Analysis of scenarios dependencies	Choice of countermeasures
National standard	Data gathering method	Analysis of countermeasures' dependencies
International standard	Data verification	Analysis of countermeasures' influence
Declared compliance with standards	Basis for risk calculation	Estimation of risk treatment efficiency
Target group	Number of risk levels	Risk monitoring
Sophistication of usage/implementation	Basis for probability estimation	Detection of new risks
Popularity	Number of probability levels	Automatic correction of dependant risk for security policy framework
Flexibility	Basis for cause estimation	Procedures generation support
Method's scope of action	Number of cause levels	
Method of risk identification	Cause metric	
Risk completeness verification	Probability metric	

The selection and analysis criteria applied in the studies (besides the criteria of Felderer et al., 2016 and Gritzalis et al., 2018 which are presented in Figure 2 and Table 1) are summarised in Tables 3 and 4. Based on them, common criteria for choosing a cybersecurity management solution were derived. The criteria are presented in Section 5.

Table 3. Method selection criteria from several studies (Cherdantseva et al., 2016; Felderer et al., 2016; Gritzalis et al., 2018; Ionita & Hartel, 2013; Qassim et al., 2019; Wangen et al., 2018).

(Qassim et al., 2019)	(Cherdantseva et al., 2016)	(Gritzalis et al., 2018)	(Wangen et al., 2018)	(Ionita & Hartel, 2013)	(Felderer et al., 2016)
Availability free of charge	Coverage of risk assessment processes	English documentation	Large number of citations	English documentation	English documentation
English documentation	Industrial control systems-specific design	Focus on information security risks	English or Norwegian documentation	Practical application in more than one country	Documentation in a peer-reviewed paper
Publication by a standard body or governmental agency	Scientific, cybersecurity-related origin	Presence of a risk assessment method description	Coverage of particular risk subjects	Application to an existing system or a system design	Coverage of a model-based security testing approach
Implementation or application in the context of industrial control systems			Coverage of three compulsory risk assessment processes	Presence of a method description that comprises all obligatory risk assessment stages	
Detailed descriptions			Timeliness	Specific audience	Focus on information security risks

4 Ethical questions

Cybersecurity management is perceived as an ethical process. By protecting cyberassets, security of human beings that depend on them is improved. Moreover, this is an ethical obligation for cybersecurity professionals to protect their organisations' infrastructure from intrusions and attacks (Vallor & Rewak, 2018). Consequently, cybersecurity assessments, as a constituent of a cybersecurity management process should be perceived as ethical. During assessments, the overall level of protection is determined and vulnerabilities are discovered (Dalalana Bertoglio & Zorzo, 2017; Qassim et al., 2019; Scarfone et al., 2008). This, in turn, enables the introduction of controls that aim at improving the recognised situation.

One of the techniques, broadly instructed in cybersecurity guidelines is the emulation of hacking techniques. There, the analysts employ the same tactics and tools as hackers do (Harper et al., 2018). Already this poses ethical questions, as hacking is generally considered unethical (Best, 2006; Falk, 2004), but the situation gets even more complicated when the analyses are directed toward the human component. *Social engineering* refers to manipulating individuals to perform specific actions or to reveal sensitive information (Lohani, 2019; Sargent & Webb, 2020). Social engineering attacks intertwine human interactions and technical measures (Klimburg-Witjes & Wentland, 2021). In cybersecurity assessment, social engineering exercises aim at testing the resistance of the human element in an organisation to this type of hacking technique. It facilitates determining the level of cybersecurity awareness and enables identifying weaknesses in user behaviour, including not following cybersecurity policies (Scarfone et al., 2008).

Table 4. Evaluation and comparison criteria from several studies (Cherdantseva et al., 2016; Giannopoulos et al., 2012; Ionita & Hartel, 2013; Meriah & Rabai, 2018; Qassim et al., 2019; Shahriar & Zulkernine, 2009).

(Qassim et al., 2019)		(Cherdantseva et al., 2016)	(Shahriar & Zulkernine, 2009)	(Ionita & Hartel, 2013)	(Giannopoulos et al., 2012)	(Meriah & Rabai, 2018)
Assessment mode (active or passive)	Aim	Test level	Class	Scope	Purpose	
	Application domain	Source of test cases	Type (quantitative or qualitative)	Objectives	Outcome	
Compliance to the NERC requirements	Addressed phases and concepts of risk management	Test generation method	Type of target users (management, operational or technical)	Applied techniques and standards	Type of system application	
Coverage of cybersecurity management processes	Impact assessment scheme	Vulnerability coverage	Supported risk assessment phases	Coverage of interdependencies	Inputs	
	Sources of data for deriving probabilities	Granularity of test cases	Availability of supporting tools (paid, free)	Addressing of cross-sectoral risks	structure of the security management process	
	Evaluation procedure	Tool automation	Release date	Target users		
	Availability of supporting tools	Application domain	Availability of a standalone version	Relevance to resilience	Supporting tools	
			Price			
			Required skills			
			Sponsor			
			Focus			

Commonly adopted cybersecurity assessment frameworks describe social-engineering-related testing operations. The Information Systems Security Assessment Framework (ISSAF) explains several social engineering techniques to be applied during cybersecurity assessment. When evaluating compliance with *handling sensitive information* or *password storage* policies, testers are instructed to look for documents left on users' desktops, around office devices or to seek written down passwords, notepapers, or keys attached to monitors that enable opening lockers that often contain message pads with passwords. Also, they should approach users' workstations and check if they can access them due to users' negligence. Another technique recommended for detecting cybersecurity "incompliance" of employees is "shoulder surfing" i.e. stealthy observing users when they type in passwords on keyboards. Also, the fundamental social engineering activity i.e. wastepaper analysis is indicated to be performed during cybersecurity assessments (Rathore et al., 2006).

On a weekly basis, evaluators should call users and impersonate IT Helpdesk analysts to identify employees who would disclose their passwords. According to the framework, the employees should be dismissed from their work or at least severely sanctioned. The auditors can phone the audited organisation and present themselves as an employee requesting assistance. They may also act as a monitoring or maintenance unit and offer help in resolving some fictional problem. They may even cause a real system disruption to make the situation more realistic. In all the cases sensitive information from administrators or personnel in charge would be required (Rathore et al., 2006). The ISSAF framework instructs the evaluators not to inform the employees about the social engineering activities. Only the necessary personnel indicated in the organisation's procedures will be briefed on them. This is explained by the potential presence of "malicious insiders" in the organisation i.e. male employees that collaborate with attackers or are attackers themselves (Rathore et al., 2006).

Similarly the NIST Special Publication 800-115 "Technical Guide to Information Security Testing and Assessment" (Scarfone et al., 2008) indicates misrepresentation as a help desk worker or an employee that requests assistance to be employed during assessments. Also, phishing or sending e-mail with a malicious attachment is indicated in the guideline. The document reflects on the delicate nature of the exercise in the sense that it may cause unwanted bias and negative emotions among workers. Thus, the demand for conducting experiments in an indiscriminating way is emphasised.

The Open Source Security Testing Methodology Manual (OSSTMM) (Herzog, 2010) devotes Chapter 7 to testing human security. There, in section 7.6 *Trust Verification*, impersonating a member of the internal support or delivery personnel, a manager or external support or delivery agent is thought to determine users' susceptibility to reveal sensitive data based on social engineering attacks. Also, phishing exercises need to be performed in this respect. What is more, evaluating the organisation's resistance to extreme or mass reactions, such as revolt, violence or chaos caused by disruption of personnel and the use of misinformation or other psychological abuse is discussed. The Open Web Application Security Project (OWASP, 2020) Testing Guide points out social engineering steps when testing a web application's security. The Penetration Testing Execution Standard (PTES) (*PTES Technical Guidelines – The Penetration Testing Execution Standard*, 2019) directs to the Social-

Engineering Toolkit (SET) as a means for simulating social-engineering attacks and determining their effectiveness in a given environment.

These testing activities may raise serious doubts as to their ethical dimension. To address them, certain requirements to the analysis procedure are introduced. One of them is performing experiments in a controlled and secure way (Harper et al., 2018), so no damage is incurred. Another requires obtaining the authorisation for carrying out the tests from the analysed organisation and involved parties (e.g. employees or contractors) (Oriyano, 2017). Within the latter boundaries, the activity should become an “*ethical*” *hacking*. The former is associated with so-called *penetration testing* (Dalalana Bertoglio & Zorzo, 2017; Gordon, 2016; Oriyano, 2017). Penetration testers use the techniques of malicious attackers, but in a controlled and safe manner (Harper et al., 2018). These “*ethicising*” operations influence the course of experiments and may have an impact on the results. While the ethical component needs to be considered when planning the testing exercises, a proper balance between these two dimensions needs to be found by each individual organisation. The decision can be facilitated by providing the ethical “parameters” of a cybersecurity assessment framework. Such parameters, for instance, could take the form of an ethical discussion of each assessment activity. The presence of the ethical attributes needs to be taken into account when selecting a specific solution.

5 Criteria for choosing a cybersecurity solution

In this section, a consolidated set of criteria derived from the publications identified during the literature review is proposed. The criteria should help in choosing a cybersecurity solution. They are presented in Tables 5-11. Table 5 comprises selection criteria that facilitate a quick decision on a method choice. The criteria include the presence of the discussion of ethical questions in the documentation of a method.

Table 5. Selection criteria.

Focus and scope	Release	Documentation	Application
Particular focus	Publication by a standard body or governmental agency	Documentation in a specific language	Implementation or application in a particular sector or domain
Specific audience	Scientific, cybersecurity-related origin	Detailed descriptions	Application to an existing system or a system design
Coverage of particular risk subjects	Timeliness	Documentation in a peer-reviewed paper	Practical application in more than one country
Coverage of all assessment processes	Availability free of charge	Discussion of ethical questions	
Coverage of specific assessment processes		Large number of citations	
Coverage of a model-based security testing approach			

Tables 6-11 contain criteria that enable more detailed analyses and comparisons between various frameworks. They are related to the scope, application, design, compliance and specific features of the solutions. The criteria are provided with literature references, where their definitions and descriptions can be found. Based on the discussion presented in Section 4, the ethical criterion related to the discussion of security assessment activities has been introduced into the group of application-related criteria.

Table 6. Scope-related criteria.

Criterion	Reference
Aim	(Cherdantseva et al., 2016)
Scope	(Fabisiak et al., 2012; Giannopoulos et al., 2012; Gritzalis et al., 2018)
Objectives	(Giannopoulos et al., 2012)
Purpose	(Meriah & Rabai, 2018)
Focus	(Gritzalis et al., 2018; Ionita & Hartel, 2013)
Inputs	(Meriah & Rabai, 2018)
Outcome	(Meriah & Rabai, 2018)
Type (quantitative or qualitative)	(Fabisiak et al., 2012; Ionita & Hartel, 2013)

Table 7. Application-related criteria.

Criterion	Reference
Release (date)	(Gritzalis et al., 2018; Ionita & Hartel, 2013)
Last update	(Gritzalis et al., 2018)
Application domain	(Cherdantseva et al., 2016; Shahriar & Zulkernine, 2009)
Type of system application	(Meriah & Rabai, 2018)
Target users	(Fabisiak et al., 2012; Giannopoulos et al., 2012; Ionita & Hartel, 2013)
Target organisation (type, size)	(Gritzalis et al., 2018)
Availability of a standalone version	(Ionita & Hartel, 2013)
Availability of supporting tools (paid, free)	(Cherdantseva et al., 2016; Ionita & Hartel, 2013; Meriah & Rabai, 2018)
Software support	(Gritzalis et al., 2018)
Tool automation	(Shahriar & Zulkernine, 2009)
Required skills	(Ionita & Hartel, 2013)
Training	(Gritzalis et al., 2018)
Popularity	(Fabisiak et al., 2012)
National standard	(Fabisiak et al., 2012)

International standard	(Fabisiak et al., 2012)
English documentation	(Fabisiak et al., 2012)
Ethical discussion of individual assessment activities	
<i>Cost</i>	(Fabisiak et al., 2012; Gritzalis et al., 2018)
Price	(Ionita & Hartel, 2013)
Support cost	(Fabisiak et al., 2012)
Software cost	(Fabisiak et al., 2012)
Sponsor	(Ionita & Hartel, 2013)

Table 8. Criteria related to detailed design features of a solution.

Criterion	Reference	Criterion	Reference
Impact assessment scheme	(Cherdantseva et al., 2016)	Number of probability levels	(Fabisiak et al., 2012)
Sources of data for deriving probabilities		Basis for cause estimation	
Evaluation procedure		Number of cause levels	
Source of test cases	(Shahriar & Zulkernine, 2009)	Cause metric	
Test generation method		Probability metric	
Vulnerability coverage		Risk metric	
Method of risk identification	(Fabisiak et al., 2012)	Choice of countermeasures	
Risk completeness verification		Analysis of countermeasures' dependencies	
Number of standard scenarios		Analysis of countermeasures' influence	
Analysis of scenarios dependencies		Estimation of risk treatment efficiency	
Data gathering method		Detection of new risks	
Data verification		Automatic correction of dependent risk	
Basis for risk calculation		Structure of the security management process	(Meriah & Rabai, 2018)
Number of risk levels		Addressed phases and concepts of risk management	(Cherdantseva et al., 2016; Gritzalis et al., 2018; Ionita & Hartel, 2013; Qassim et al., 2019)
Basis for probability estimation		Risk calculation class	(Gritzalis et al., 2018; Ionita & Hartel, 2013)

Table 9. Compliance criteria.

Criterion	Reference
Declared compliance with standards	(Fabisiak et al., 2012; Gritzalis et al., 2018)
Applied techniques and standards	(Giannopoulos et al., 2012)
Compliance to the NERC CIP requirements	(Qassim et al., 2019)

Table 10. Criteria related to general characteristics of a solution.

Criterion	Reference
Completeness	(Gritzalis et al., 2018)
Adaptability	(Fabisiak et al., 2012; Gritzalis et al., 2018)
Usability (Interface, handle errors, documentation)	(Gritzalis et al., 2018)
Ease of use	(Gritzalis et al., 2018)
Sophistication of usage/implementation	(Fabisiak et al., 2012)
Flexibility	(Fabisiak et al., 2012)
Validity	(Gritzalis et al., 2018)
Usefulness	(Gritzalis et al., 2018)

Table 11. Criteria related to additional features of a solution.

Criterion	Reference
Coverage of interdependencies	(Giannopoulos et al., 2012)
Addressing of cross-sectoral risks	(Giannopoulos et al., 2012)
Support for security policy framework generation	(Fabisiak et al., 2012)
Procedures generation support	(Fabisiak et al., 2012)
Risk monitoring	(Fabisiak et al., 2012)

6 Conclusions

Based on a systematic literature review process around ninety criteria that facilitate an informed choice of a cybersecurity management solution have been identified. The criteria have been consolidated into selection criteria that enable a quick choice of a framework primarily based on its scope and applicability to a specific domain. In addition, six groups of attributes have been distinguished that support thorough analyses and comparisons between different solutions. The six categories are related to the scope, application (including the cost), design features, compliance, general characteristics and supplementary features of a proposal. Around eighty different criteria have been classified into the categories.

Cybersecurity assessment methodologies may include activities that raise ethical questions. Widely used cybersecurity guidelines instruct to emulate hacking techniques to identify vulnerabilities in the organisation's cybersecurity posture. Even more, with the aim of verifying the resistance of employees to social engineering, exercises that employ this hacking technique are advised. Namely, the testers should try to manipulate individuals to perform specific actions or to reveal sensitive information. In this way, the level of cybersecurity awareness and the weaknesses in user behaviour can be identified. Although valuable from the point of cybersecurity, it can have a negative impact on social relations, the level of trust in the organisation and individual situations of employees. For these reasons, the ethically questionable activities need to be

transparently indicated in cybersecurity assessment frameworks and the associated ethical component discussed. Consequently, criteria related to the ethical application of a methodology have been proposed and included in the consolidated set.

References

- Accenture. (2021a). *2021 Cyber Threat Intelligence Report* (Issue July). Accenture.
- Accenture. (2021b). *Technology Vision 2021*.
- Alcaraz, C., & Lopez, J. (2013). Wide-area situational awareness for critical infrastructure protection. *Computer*, 46(4), 30–37. <https://doi.org/10.1109/MC.2013.72>
- Barrere, M., Badonnel, R., & Festor, O. (2014). Vulnerability assessment in autonomic networks and services: A survey. *IEEE Communications Surveys and Tutorials*, 16(2), 988–1004. <https://doi.org/10.1109/SURV.2013.082713.00154>
- Bolzoni, D., Leszczyna, R., Wróbel, M. R., & Wrobel, M. R. (2016). Situational Awareness Network for the electric power system: The architecture and testing metrics. In M. Ganzha, L. Maciaszek, & M. Paprzycki (Eds.), *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, FedCSIS 2016* (pp. 743–749). IEEE. <https://doi.org/10.15439/2016F50>
- Cherdantseva, Y., Burnap, P., Blyth, A., Eden, P., Jones, K., Soulsby, H., & Stoddart, K. (2016). A review of cyber security risk assessment methods for SCADA systems. *Computers & Security*, 56, 1–27. <https://doi.org/https://doi.org/10.1016/j.cose.2015.09.009>
- Dalalana Bertoglio, D., & Zorzo, A. F. (2017). Overview and open issues on penetration test. *Journal of the Brazilian Computer Society*, 23(1), 1–16. <https://doi.org/10.1186/s13173-017-0051-1>
- Deloitte Insights. (2021). *Tech Trends 2022*.
- Duggal, N. (2021, December 7). *Top 9 New Technology Trends for 2022*. https://www.simplilearn.com/top-technology-trends-and-jobs-article#9_cyber_security
- ENISA. (2021). *ENISA Threat Landscape 2021* (Issue October). ENISA. <https://doi.org/10.2824/324797>
- Fabisiak, L., Hyla, T., & Klasa, T. (2012). Comparative Analysis of Information Security Assessment and Management Methods. *Studia i Materiały Polskiego Stowarzyszenia Zarządzania Wiedza / Studies & Proceedings Polish Association for Knowledge Management*, 60, 55–70.
- Felderer, M., & Schieferdecker, I. (2014). A taxonomy of risk-based testing. *International Journal on Software Tools for Technology Transfer*, 16(5), 559–568. <https://doi.org/10.1007/s10009-014-0332-3>
- Felderer, M., Zech, P., Breu, R., Büchler, M., & Pretschner, A. (2016). Model-based security testing: A taxonomy and systematic classification. *Software Testing Verification and*

- Reliability*, 26(2), 119–148. <https://doi.org/10.1002/stvr.1580>
- Giannopoulos, G., Filippini, R., & Schimmer, M. (2012). Risk assessment methodologies for Critical Infrastructure Protection. Part I: A state of the art. In *European Commission JRC (Joint Research Center) Technical notes*. <https://doi.org/10.2788/22260>
- Gordon, A. (2016). The Official (ISC) 2® Guide to the CCSP SM CBK ®. In *The Official (ISC) 2® Guide to the CCSP SM CBK ®*. John Wiley & Sons, Inc. <https://doi.org/10.1002/9781119419198>
- Gritzalis, D., Iseppi, G., Mylonas, A., & Stavrou, V. (2018). Exiting the risk assessment maze: A meta-survey. *ACM Computing Surveys*, 51(1), 1–30. <https://doi.org/10.1145/3145905>
- Hahn, A., & Govindarasu, M. (2011). An evaluation of cybersecurity assessment tools on a SCADA environment. *IEEE Power and Energy Society General Meeting*. <https://doi.org/10.1109/PES.2011.6039845>
- Harper, A., Regalado, D., Linn, R., Sims, S., Spasojevic, B., Martinez, L., Baucom, M., Eagle, C., & Harris, S. (2018). *Gray hat hacking: the ethical hacker's handbook* (Fifth). McGraw-Hill Education.
- Herzog, P. (2010). *OSSTMM 3 - The Open Source Security Testing Methodology Manual*. <http://www.isecom.org/mirror/OSSTMM.3.pdf>
- Holm, H., Sommestad, T., Almroth, J., & Persson, M. (2011). A quantitative evaluation of vulnerability scanning. *Information Management and Computer Security*, 19(4), 231–247. <https://doi.org/10.1108/09685221111173058>
- Ionita, D., & Hartel, P. (2013). *Current Established Risk Assessment Methodologies and Tools*.
- ISO/IEC. (2013). ISO/IEC 27001:2013: Information technology -- Security techniques -- Information security management systems -- Requirements. In *ISO/IEC 27001* (p. 23).
- ISO/IEC. (2018). *ISO/IEC:2018 Information technology — Security techniques — Information security management systems — Overview and: Vol. 5th edit* (p. 38).
- Kitchenham, B., & Brereton, P. (2013). A systematic review of systematic review process research in software engineering. *Information and Software Technology*, 55(12), 2049–2075. <https://doi.org/https://doi.org/10.1016/j.infsof.2013.07.010>
- Klimburg-Witjes, N., & Wentland, A. (2021). Hacking Humans? Social Engineering and the Construction of the “Deficient User” in Cybersecurity Discourses. *Science, Technology & Human Values*, 46(6), 1316–1339. <https://search.ebscohost.com/login.aspx?direct=true&db=bsu&AN=152626306&lang=pl&site=ehost-live&scope=site>
- Leszczyna, R. (2021). Review of cybersecurity assessment methods: Applicability perspective. *Computers & Security*, 108, 102376. <https://doi.org/10.1016/J.COSE.2021.102376>
- Leszczyna, R., Wallis, T., & Wróbel, M. R. (2019). Developing novel solutions to realise the European Energy – Information Sharing & Analysis Centre. *Decision Support Systems*, 122. <https://doi.org/10.1016/j.dss.2019.05.007>

- Leszczyna, R., & Wróbel, M. R. (2019). Threat intelligence platform for the energy sector. *Software: Practice & Experience*.
- Li, J., Beba, S., & Karlsen, M. M. (2019). Evaluation of open-source IDE plugins for detecting security vulnerabilities. *ACM International Conference Proceeding Series*, 200–209. <https://doi.org/10.1145/3319008.3319011>
- Lohani, S. (2019). Social Engineering: Hacking into Humans. *International Journal of Advanced Studies of Scientific Research*, 4(1).
- Lykou, G., Anagnostopoulou, A., Stergiopoulos, G., & Gritzalis, D. (2019). Cybersecurity self-assessment tools: Evaluating the importance for securing industrial control systems in critical infrastructures. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11260 LNCS, 129–142. https://doi.org/10.1007/978-3-030-05849-4_10
- McKinsey & Company. (2021). *The top trends in tech - executive summary*.
- Meriah, I., & Rabai, L. B. A. (2018). A survey of quantitative security risk analysis models for computer systems. *Proceedings of the 2nd International Conference on Advances in Artificial Intelligence (ICAAI 2018)*, 36–40. <https://doi.org/10.1145/3292448.3292456>
- NIST. (2011). NIST SP 800-39 Managing Information Security Risk Organization, Mission, and Information System View. In *Nist Special Publication* (Issue March). <http://csrc.nist.gov/publications/nistpubs/800-39/SP800-39-final.pdf>
- Oriyano, S.-P. (2017). Penetration Testing Essentials. In *Penetration Testing Essentials*. John Wiley & Sons, Inc. <https://doi.org/10.1002/9781119419358>
- OWASP. (2020). *OWASP Testing Guide v4.2*. <https://github.com/OWASP/wstg/releases/download/v4.2/wstg-v4.2.pdf>
- PTES Technical Guidelines -- The Penetration Testing Execution Standard*. (2019). <http://www.pentest-standard.org>
- Qassim, Q. S., Jamil, N., Daud, M., Patel, A., & Ja'afar, N. (2019). A review of security assessment methodologies in industrial control systems. *Information and Computer Security*, 27(1), 47–61. <https://doi.org/10.1108/ICS-04-2018-0048>
- Rathore, B., Brunner, M., Dilaj, M., Herrera, O., Brunati, P., Subramaniam, R. K., Raman, S., & Chavan, U. (2006). *Information Systems Security Assessment Framework (ISSAF) Draft 0.2.1B*.
- Rogers, R., & Syngress Media, I. (2004). *Security assessment: case studies for implementing the NSA IAM*. Syngress.
- Sargent, S. A., & Webb, J. P. (2020). The Key to Trust: Social Engineering Fraud and Modern Threat Detection. *Benefits Magazine*, 57(1).
- Scarfone, K., Souppaya, M., Cody, A., & Orebaugh, A. (2008). *NIST SP 800-115 Technical Guide to Information Security Testing and Assessment*. NIST.
- Shah, S., & Mehre, B. M. (2015). An overview of vulnerability assessment and penetration

- testing techniques. *Journal of Computer Virology and Hacking Techniques*, 11(1), 27–49.
<https://doi.org/10.1007/s11416-014-0231-x>
- Shahriar, H., & Zulkernine, M. (2009). Automatic testing of program security vulnerabilities. *Proceedings - International Computer Software and Applications Conference*, 2, 550–555.
<https://doi.org/10.1109/COMPSAC.2009.191>
- Sophos. (2021). *The State of Ransomware 2021* (Issue April). Sophos.
<https://doi.org/10.2824/324797>
- Tadda, G. P., & Salerno, J. S. (2010). Overview of Cyber Situational Awareness. In S. Jajodia, P. Liu, V. Swarup, & C. Wang (Eds.), *Cyber Situational Awareness* (Vol. 46, pp. 15–35). Springer US. <https://doi.org/10.1007/978-1-4419-0140-8>
- Vallor, S., & Rewak, W. J. (2018). *An Introduction to Cybersecurity Ethics*. <https://www.scu.edu/ethics/focus-areas/technology-ethics/resources/an-introduction-to-cybersecurity-ethics/>
- Wangen, G., Hallstensen, C., & Snekkenes, E. (2018). A framework for estimating information security risk assessment method completeness: Core Unified Risk Framework, CURF. *International Journal of Information Security*, 17(6), 681–699.
<https://doi.org/10.1007/s10207-017-0382-0>
- Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: writing a literature review. *MIS Quarterly*, 26(2), xiii–xxiii.
- Weiss, S. (2008). Industrial approaches and standards for security assessment. In I. Eusgeld, F. C. Freiling, & R. Reussner (Eds.), *Dependability Metrics: Vol. 4909 LNCS* (pp. 166–175).
https://doi.org/10.1007/978-3-540-68947-8_14

Bullshit Blockchain Mining?

Kazuyuki Shimizu¹[0000-0002-4847-0197]

¹ Meiji University, Tokyo/1018301, JAPAN
shimizuk@meiji.ac.jp

Abstract. Smartphones and high-speed computers have advanced our productivity sufficiently and are vital to our daily lives. However, our everyday environment is not much different from before and is surrounded by more "bullshit jobs". It's as if someone were out there making up bullshit jobs to keep us all working. Here, let us consider the meaning of "bullshit". Indeed, Bitcoin has been touted as an innovative method of information security as distributed ledger technology (DLT). GAFA's "centralization of information" can be decentralized by this DLT technology. However, the spread of blockchain technology is considered "bullshit" for the following three reasons. First of all, the issue of electricity consumption in Blockchain mining and their industry, and secondly, whether the blockchain technology is adaptable to the organizational structure and what is the bureaucratic "bullshit job"? Finally, there are some applicable fields in blockchain technology and their limitations.

Blockchain mining cannot wipe out uncertainties about the future. Still, breaking through the current situation, blockchain mining will be a ray of light in the darkness for a globally distributed economy. To try to automate "bullshit jobs" through human enhancement such as the Internet of Ability (IoA), it is necessary to carefully analyze, break down "bullshit jobs" and render unto man's jobs.

Keywords: Blockchain governance, corporate and stakeholder governance, Internet of Ability (IoA), institutional Crypto-economics

1 Introduction

Satoshi, 32 years old, has been doing Ethereum mining since 2020 as a side business between his main business¹. He downloaded the software, tried mining, and got a few dollars in terms of the conversion rate of Ethereum for the first time. He started thinking that it would work somehow. He built a specialized P.C. for mining, bought a video card RXT2080Ti from Nvidia, and started mining. However, the P.C. dedicated to mining did not work correctly and was initially challenging. After all, mining heats his PC about 70 degrees, and it needs to be released with airconditioner. His total revenue will be about \$70, excluding the electricity bill spending of about \$30 per month. This means he can then make about \$40 profit per month². Of course, there are possibilities where the data is rejected in the mining pool³. This story is an individual-based story of Satoshi.

Thus, mining is not heavy labour using limbs like a gold rush or a coal mine worker. Mining is simply setting up a computer for crypto-assets⁴ and letting the computer machine do calculations 24 hours a day and 365 days a year, using electricity constantly. This does not involve any physical labour work. Can this be called "Mining"? It is a kind of typical "Bullshit job".

The rest of the paper is structured as follows. Chapter 2 addresses the theoretical background of what a "bullshit job" is in a bureaucratic organization. Chapter 3 investigates the power utilization of mining, an incentive for Blockchain. Finally, we will present a new way to use Blockchain for management. Next, I would like to explain the definition of "Bullshit job".

2 Theoretical Background

2.1 What is a "Bullshit job"?

The word is quoted from David Graeber. His definition for

"a bullshit job is a form of paid employment that is so completely pointless, unnecessary, or pernicious that even the employee cannot justify its existence even though, as part of the conditions of employment, the employee feels obliged to pretend that this is not the case." (David, 2018)

If such a job is a typical bullshit job, why do we kind of mine in front of a computer which could increase productivity? C. Marx believes that capitalism needs a Bullshit Job to keep it running. However, if such a bullshit job of mining could change this system that keeps us working, it would be necessary to spend our time destroying this system.

Labour benefits are for obtaining livelihood (salary). Converting labour into capital is detrimental to recognizing the work one really wants to do such as comfortable jobs. It is not a scientific perspective to discuss whether you like your job. We do not go into the psychological point of view here. Blockchain mining is proofed by the node every 10 minutes. Here we need to identify between the volatile price trends of crypto assets and the various pioneering management cases of inspired blockchain technology. The point is that we need to investigate the management cases for their suitability, not instability. Zheng, Xie and Dai roughly categorize the blockchain applications into five domains, which are 1. finance, 2. IoT, 3. public and social services, 4. reputation system, and 5. security and privacy, see Fig. 1. (Zheng, Xie, Dai, Chen, & Wang, 2018) Therefore, it is necessary to reconfirm the significance of data confirmation in Blockchain other than letting the computer data encryption that this work is not Bullshit.

Table 1. Representative application domains of Blockchain

<i>Blockchain Applications</i>	Fields	brief description
	1. finance	<ul style="list-style-type: none"> Financial Services <u>Enterprise Transformation</u> P2P Financial Market Risk Management
	2. IoT	<ul style="list-style-type: none"> E-business Safety and Privacy
	3. public and social services	<ul style="list-style-type: none"> Land Registration Energy Saving Education Free-Speech Right
	<u>4. reputation system</u>	<ul style="list-style-type: none"> <u>Academics</u>⁵ <u>Web Community</u>
	5. security and privacy	<ul style="list-style-type: none"> Security Enhancement Privacy Protection

Gordon pointed out that the lives and productivity of most Americans had changed to an enlargement that would have been previously unimaginable between 1870 and 1940. The impact of mechanization, urbanization, automation etc., shifted many into jobs that paid better but gave them on demand, increasing the amount of time they had to pursue activities other than working, enabling them to receive consumer goods (Gordon , 2017). What is essential is that productivity was promoted by the technology developed at that time and the extra time gained by used for consumption activities. So is consumption the job we are looking for?

2.1.1 DAO (Decentralized Autonomous Organization) as a reputation system

Blockchain technology needs to be reliable for those who want to use it. Therefore, the reputation system in the blockchain application will be described, which is above mentioned fourth factor. Google has built a system to search for useful information on the Internet with well-known PageRank. Other social network services (SNS) analyze the relationships within this network and use them to provide the information that users want.

Mitari examines the relationship between the fame of works of art such as paintings and music with the regions (U.K., Italy, Germany, and the United States) by age group. Again, this study reveals that fame is associated with social networks.

- What is important here is that the shape of the network ensures the suitability, not instability of the Blockchain system. This is because a system structure called DAO (Decentralized Autonomous Organization) constructs its reputation and consensus which lead by the governance (control).

2.2 Why is Weber's bureaucracy so appealing?

For example, suppose a country's military system is inefficient. The military system is a typical bureaucracy. Therefore, the flow of sub-, sub- sub-, and sub- sub- sub-contracting jobs to other private companies to reform the inefficient bureaucracy. It's considered to be exactly the privatization of the bureaucratic system. M. Weber uses the concept of legitimacy to ensure that the law is essential in modern society. This legitimacy includes the concept of democracy. From this point of view, Freeman's stakeholder theory can be adopted for legitimacy (Freeman, Harrison, Wick, Parmar, & Colle, 2001). Stakeholder theory has been developed to counter this dominant mindset, such as the bureaucratic system. One side considers the centralized authority in Weber's bureaucracy as a client/server model, and the other considers Freeman's stakeholder theory as a network comparison with a decentralized P2P model. The decentralized P2P model has the control methods such as public, private, and consortium Blockchain.

The "centralization of information" by huge global ICT companies such as GAFA has reached various technological limits. To do this, we should return to the original Internet with decentralization by DLT technology (George, 2018). The Internet service, which is created by decentralized network architecture, is free, but there is such a crisis here if you want to conduct commercial use, such as financial transactions. Therefore, in order to use the Internet for commercial use, internet security is essential to protect privacy data. This requires enormous costs to protect the "walled garden". DLT brings the solution for this situation. In other words, the information is fixed from the beginning and created as a database to not be changed.

Recently, several blockchain projects have been created to transform the government system in place in more than 30 countries. This government blockchain is a technology that is directly related to the social system approach. The consensus mechanism forms the core of blockchain technology. Traditionally, a consensus mechanism is not in the field of machinery, but in humankind. Blockchain operates through various types of consensus algorithms with human intervention (MyungSan, 2018). He defines the nature of bureaucracy as a social technology that works as an "information processing machine" for the community. In other words, bureaucracy is a social technology dedicated to distributing and processing information needed in a specific community. It could create a consensus by the multi-stakeholder system. Therefore, it is theoretically not problematic to claim that the blockchain technology would replace the role of bureaucracy.

2.3 Did IoT (Industrie 4.0) make us happy?

Keynes, J. M. said that predicted in the 1930s after World War I, the standard of living of progressive countries 100 years later (1930s + 100 years = 2030) would be four to eight times higher than that time. The world's population has more than tripled from 2 billion in the 1930s to 7 billion in the 2020s. Indeed, his conjecture is correct, given that food and energy allocations are optimized. (Maynard, 1930)

Organizations such as the bureaucracy mentioned above have improved our standard of living. Given this, David G. "Bullshit Job" is, What does it mean? Is capitalism just keeping humans working? Table 1 below describes technological developments that have improved living standards. Is Blockchain Mining to ensure data trust (such as Privacy data protection, GDPR) described below as a way to automate "Bullshit Job"? Before humans become cyborgs via A.I., there is an idea from IoT (Internet of Things) to IoA (Internet of Ability). Applying this idea, IoA automates the "Bullshit Job".

Table 2. the technological developments

Year	brief description
1700	Manual Labor Production flexibility was highest when manual labour dominated production. "high-mix low-volume" production.
1800	The first industrial revolution (Mechanization) was marked by a transition from manual labour to the water- and steam-powered mechanical loom. Steam engines were powered by coal and coal mining.
1900	The second industrial revolution followed the introduction of electrically powered mass production based on the division of labour. It increased unemployment, as machines replaced many factory workers. The Factory Acts were passed by the Parliament in U.K.
1960	The third industrial revolution also known as the digital revolution, occurred after the two great wars. High-speed computers were required during the Cold War. The most important innovation was the Internet. It was created by the Pentagon under the name ARPANET.
1970	The concept of (A.I.) was emerged. It is an entirely different period when computer technology used as an interface. The importance of human power decreased, and production speed increased considerably through computer systems.
1980	Machine learning is a subset of the broader field of A.I. that focuses on teaching computers how to learn without needing to be programmed. To educate a machine, three components are required: Big data-set, high-speed computer and an first algorithm.
2010	Deep learning is an advance class of machine learning. Different types of algorithm exist inspired by the human brain's neural networks. Since 2015, the Central Processing Unit (CPU) has had a counterpart in the Graphics Processing Unit (GPU). Deep learning adopted the structure of the human neural network, using CPU and GPU.
2022 third quarter	Blockscale ; Intel launches new Bitcoin mining chip. Bitmain's Antminer S19 Pro has a hash rate of 110 TH/s and consumes 3,250 watts of power, and Blockscale chips would have a total hash rate of 148.5 TH/s and consume the same energy efficiency ⁶ .

(Kazuyuki, 2020)

3 Mining work

3.1 The geographic shift of Bitcoin mining

We have already introduced Satoshi's mining activity by the introduction. What about commercial miners doing large-scale mining? 1,660 firms are operating in multiple jurisdictions report their activities in 2020. According to the Cambridge Alternative Finance Center (CCAF), China was the preferred mining site until 2020, but due to strong regulations ban in the first half of 2021, most of the operations were suspended, and in 2022 the United States and Canadian mining companies are increasing their presence (Cambridge Center for Alternative Finance, 2021). One of the reasons for China's mining ban may be the problem of using abundant coal to generate energy. Similarly, a giant ICT company such as GAFA uses a lot of electricity to manage their huge data center to utilize and develop artificial intelligence.

Bitcoin reached one trillion U.S. dollars market cap in 2022. One of the largest bitcoin mines is Whinstone U.S. in Rockland, Texas, in America. Whinstone U.S. has measured its 300 MWh⁷ per year in developed capacity. Their subsidiary Riot currently has a deployed hash rate capacity of 3.0 EH/s⁸ utilizing approximately 91 MWh of energy. Global data centre electricity use in 2020 was 200-250 TWh or around 1% of global final electricity demand. This excludes energy used for crypto assets-mining, which was around 100 TWh in 2020 [The IEA, 2022].

3.2 Iceland as a mining site

Küfeoğlu, S. and Özkuran, M. suggest the historical peak of power consumption of bitcoin mining during the several week period of time on 18 December 2017 with a demand of between 1.3 to 14.8 GW. This maximum demand figure was between the installed capacities of Finland (~16 G.W.) and Denmark (~14 G.W.). (Küfeoğlu & Özkuran, 2019) With a world population of about 7 billion, electricity demand in developed countries is much higher than in emerging countries. Based on Wikipedia calculations, it is characteristic that the population of about 5.5 million in Finland (relative to the world population, 0.7%) is the fifth-largest in the world in terms of average power per capita (1,740 watts per person). In particular, Iceland (5,898 watts per person)⁹ is the most average power consumption per person, and Norway (2,648 watts per person) is the second (Wikipedia, 2022). Bitcoin mining consumes a lot of power, and a place with low temperature is preferred for the computer to work properly. Science said data centres are energy-intensive enterprises, estimated to account for around 1% of worldwide electricity use. These trends have clear implications for global energy demand and must be analyzed rigorously. The worldwide energy use of data centres had grown from 153 terawatt-hours (TWh) in 2005 to between 203 and 273 TWh by 2010, totalling 1.1 to 1.5% of global electricity use. Since 2010, however, the data centre landscape has changed dramatically. By 2018, global data centre workloads had increased more than sixfold, whereas data centre internet protocol (I.P.) traffic had increased by more than 10-fold. But since 2010, electricity use per computation of a typical volume server has dropped by four, mainly due to processor-efficiency

improvements. These data suggest that, although global data centre energy has increased slightly since 2010, energy growth has been substantially decoupled from growth in data centre compute instances over the same period (Masanet, Shehabi, Lei, Smith, & Koomey, 2020).

3.3 Which is the more value, data or coal mining?

According to the CCAF mentioned above, China was the most significant mining site until 2020, but most operations were suspended due to a strong regulations ban in the first half of 2021. And many Chinese miners moved to the United States and Canada in 2022. Akyildirim E., Corbet S. and Lucey B. suggest that global coal prices, especially Chinese coal mining, are more linked to Bitcoin price fluctuations than real mining conditions. (Masanet, Shehabi, Lei, Smith, & Koomey, 2020) The reason for the results of this paper is the use of the Dynamic Conditional Correlation (DCC) model, which considers systematic risk as the basis for its calculation. The DCC model systematically measures other effects. If environmental problems are ignored, the price of coal as a commodity is subject to price fluctuation factors due to Blockchain Mining. In other words, if this frees us from "Bullshit Jobs", rational market participants rely more on value-added data mining than on natural resource (coal) mining. In addition, the basic Bitcoin ownership structure is that 6.28% of addresses own 93.72% of total assets, which is a distribution like Pareto optimal (Dániel, Márton, István, & Gábor, 2014). Bitcoin's anonymity and visualization of transaction history are different from banking transactions that involve centralized control as an institution.

However, the ownership structure mentioned above increases the utility of large ownership. For mining in other management applications, a new way of incentives methods such as reputation applies to increase the utility of decentralization.

4 Essential mining

Even if the giant ICT companies such as GAFA's dominance changes somewhat with Blockchain technology, it is unclear whether our way of the bullshit job will change. However, the mental struggle to escape and win freedom from such a dominant situation seems healthy.

"Markov chains tell you the statically likely future without knowing the past; blockchains enable the real-life future by indelibly recording the past." (George, 2018)

The future in which distributed ledgers support decisions is more independent than A.I. (artificial intelligence), which disconnects probabilistic states. The difference between a bullshit job and an essential job is trust. For example, data mining is a term similar to digging gold. Because the currency as a payment method requires the trust of "gold". This trust recognizes currency = Bitcoin. Where is the trustworthiness of digital data in

this sense? Once you trust a trustworthy database, even if data mining is bullshit mining, the data that supports decisions can be trusted, and mental stability can be rewarded.

4.1 IoT to IoA (Internet of Ability)

Usually, the Internet means a personal computer identifies and exchanges digital information in the network. This identification is mainly made by people who use these PCs. In this sense, IoT (Internet of Things) means a thing positions in the network. IoT began to be used in a study called RFID (Radio Frequency Identification) at MIT for the first time in 1999. Identification-product management, which uses barcodes was positioned on the Internet using these radio waves. Also, IoT has a very similar meaning to ubiquitous (everywhere) computing.

Blockchain technology is mainly utilized for data management operations in healthcare and IoT, specifically to improve data security, including data integrity, access control, and privacy protection. Blockchain and IoT, including health IoT, can be similar meanings on the Internet of Ability (IoA). The most recognized example of an IoB is a pacemaker, a small device placed in the chest to help patients with heart conditions control abnormal heart rhythms with electrical impulses (Adere, 2022).



	<i>Fields</i>	<i>Example</i>
<i>IoT</i> (Internet of Things)	Supply chain	RFID, Connected Cars
<div style="display: flex; justify-content: space-around; align-items: center;"> <div style="text-align: center;">  </div> <div style="text-align: center;"> <u>Automate "Bullshit Jobs"</u> </div> <div style="text-align: center;">  </div> </div>		
	Human Augumentation	HappinessCounter (^^)
<i>IoB</i> (Internet of Bobby)	Healthcare, Medical equipment	Wearable health monitors
<i>IoA</i> (Internet of Ability)	Business collaboration, innovation	Supply chain

Fig. 3. Automation of "bullshit job" through the Human Augmentation (Human Augmentation Research Initiative, 2019)

5 Conclusion

C., Frey & M. , Osborne investigated jobs which are replaced by the supercomputers like using artificial intelligence. They estimate that about 47 per cent of total U.S. employment is at risk (Osborne & Frey, 2013). Their suggestion shocked many workers. Should we seriously feel uncomfortable being "meaningful job" replaced by A.I.? Aren't there more "Bullshit Jobs" out there than you have to do?

1. First, clarify the "Bullshit job" in the bureaucracy,

2. Clarifying the consensus system by DAO.
3. The important thing here is to accept the instability of the blockchain system and find compatibility.
4. Through that, Create an incentive system called reputation, which people believes.
5. Reputation isn't a way of "large cargo" incentive, that is a value transformation.
6. Instead of probabilistically creating a reputation from conventional analogical data, create it with data trusted by Blockchain.
7. This helps DAOs (Decentralized Autonomous Organizations) use reputation to find governance structures.

References

- Adere, E. M. (2022, July 14). Blockchain in healthcare and IoT: A systematic literature review. *Array*. doi:<https://doi.org/10.1016/j.array.2022.100139>
- Akyildirim, E., Corbet, S., & Lucey, B. M. (2021). *China, Coal, Calamities and Cryptos*. SSRN. Retrieved from <https://ssrn.com/abstract=3851253>
- Cambridge Center for Alternative Finance. (2021). *The 2nd Global Alternative Finance Market Benchmarking Report*. Retrieved from <https://www.jbs.cam.ac.uk/wp-content/uploads/2021/06/ccaf-2021-06-report-2nd-global-alternative-finance-benchmarking-study-report.pdf>
- Cambridge Centre for Alternative Finance. (2021). *Bitcoin network power demand*. Retrieved from <https://cbeci.org/>
- Dániel , K., Márton , P., István , C., & Gábor , V. (2014). Do the Rich Get Richer? An Empirical Analysis of the Bitcoin Transaction Network. doi:<https://doi.org/10.1371/journal.pone.0086197>
- David, G. (2018). *Bullshit Job*. Penguin Random House.
- Freeman, E. R., Harrison, J., Wick, A., Parmar, B., & Colle, S. (2001). *A Stakeholder Approach to Strategic Management*. Massachusetts: Blockwell.
- George, G. (2018). *Life After Google: The Fall of Big Data and the Rise of the Blockchain Economy*. Regnery Gateway . Retrieved from <https://www.amazon.co.jp/Life-After-Google-Blockchain-Economy/dp/1621575764>
- Gordon , R. J. (2017). *The Rise and Fall of American Growth: The U.S. Standard of Living Since the Civil War*. Princeton Univ Pr.
- Human Augmentation Research Initiative. (2019). *Human Augmentation Research*. Retrieved from <https://humanaugmentation.jp/>
- Jaime , G.-J., & Alfonso , G.-C. (2011). *Overview and Challenges of Overlay Networks: A Survey*. International Journal of Computer Science & Engineering Survey. Retrieved from https://www.researchgate.net/publication/50199321_Overview_and_Challenges_of_Overlay_Networks_A_Survey
- Kazuyuki, S. (2016). Socio-cybernetic approach into the triumvirate: stakeholder governance between management, shareholders and employees. *An Enterprise*

- Odyssey. International Conference Proceedings*, 273-280. Retrieved from <https://search.proquest.com/docview/1815362010?pq-origsite=gscholar&fromopenview=true>
- Kazuyuki, S. (2020, Oct). Digital transformation of work and ESG: Perspectives on monopoly and fair trade. *Risk Governance and Control: Financial Markets & Institutions*, pp.75-82. doi:<http://doi.org/10.22495/rgcv10i3p6>
- Küfeoğlu, S., & Özkuran, M. (2019, May). Energy Consumption of Bitcoin Mining. doi:<https://doi.org/10.17863/CAM.41230>
- Masanet, E., Shehabi, A., Lei, N., Smith, S., & Koomey, J. (2020, Feb). *Recalibrating global data center energy-use estimates*. (Science, Editor) doi:10.1126/science.aba3758
- Maynard, K. J. (1930). Retrieved from 2. Economic Possibilities for our Grandchildren: http://www.gutenberg.ca/ebooks/keynes-essaysinpersuasion/keynes-essaysinpersuasion-00-h.html#Economic_Possibilities
- MyungSan, J. (2018). *Blockchain government - a next form of infrastructure for the twenty-first century*. Journal of Open Innovation: Technology, Market, and Complexity. doi:DOI 10.1186/s40852-018-0086-3
- Norbert, Wiener. (1950). *The Human Use of Human Beings; Sybernetics and Society*. Retrieved from <http://21stcenturywiener.org/wp-content/uploads/2013/11/The-Human-Use-of-Human-Beings-by-N.-Wiener.pdf> / 2014.05.10
- Osborne, M., & Frey, C. B. (2013, Sep). *The future of employment: How susceptible are jobs to computerisation?* Retrieved from https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf
- The IEA. (2022). *Data Centres and Data Transmission Networks*. Retrieved from <https://www.iea.org/reports/data-centres-and-data-transmission-networks>
- Wikipedia. (2022). *List of countries by electricity consumption*. Retrieved from https://en.wikipedia.org/wiki/List_of_countries_by_electricity_consumption
- Yermack, D. (2017). *Corporate Governance and Blockchains*. Review of Finance, Volume 21, Issue 1. doi:<https://doi.org/10.1093/rof/rfw074>
- Zheng, Z., Xie, S., Dai, H.-N., Chen, X., & Wang, H. (2018, Oct). Blockchain challenges and opportunities: a survey. *Article in International Journal of Web and Grid Services*. doi:DOI: 10.1504/IJWGS.2018.10016848

¹ Two types of mining methods. There are two main types of mining: ASICs and GPUs. Initially, an ASIC, a device designed specifically for mining, can dig Bitcoin (BTC). And with GPU mining using a graphics board, you can earn Ethereum and Altcoin. Due to the peculiarities of ASIC devices, Bitcoin miners are oligopolies by several companies worldwide.

² Calculated by <https://www.nicehash.com/profitability-calculator/nvidia-rtx-2080-ti>

³ Looking at individual mining work, there is a “mining pool” method. Mining is becoming increasingly popular with high-speed computer devices compatible with home computers in recent times. However, the chances of realistically profiting from individual mining are diminishing.

For example following URL; <https://www.nicehash.com/profitability-calculator>

The blockchain mining process is configured so that if more miners are digging, the difficulty level goes up, while a decline in the number of miners eases the difficulty level. The higher the mining power in the entire network, the more difficult it becomes for miners to mine a new block because the Bitcoin program adjusts the mining difficulty.

As mining pools represent more significant groups of miners, they play an essential role also a potentially dangerous position in the network. To guarantee the security and trustworthiness of the Bitcoin network in the long term, none of the mining pools (or the combination of a few pools) should dominate the market. When one or more collaborating pools gain the majority of the total mining power, they can perform a 51 % majority attack and decide to validate invalid transactions. Only a few large and persistent mining pools can survive as significant and specific investments are needed to mine successfully. The danger of a 51 % attack is real. The potential risk raises questions on which factors lead to mining pool concentration and how miners collectively react when concerned about an attack.

⁴ The phenomenon of crypto-assets is digitally representing anything that we regard as anything having value putting encryption around it and putting it in the distributed ledger where the ownership of that asset can be assigned and transfer ownership can take place in viable ways.

⁵ As a basic example, the blockchain of education here is to open the learning records of students to the public and manage them as a whole like a supply chain. Furthermore, as an application example, the mutual transaction of education between teachers and students is considered as blockchain learning, and the idea is that teachers mine the possibilities of students.

⁶ Global Bitcoin Mining Data Review - Q1 2022, URL; https://bitcoinminingcouncil.com/wp-content/uploads/2022/04/2022.04.25-Q1_2022_BMC_Presentation.pdf

⁷ Megawatts per hour are used to measure the output of a power plant or the amount of electricity required by an entire city. For example, a typical coal plant is about 600 MWh, which means Whinstone U.S. uses half of a coal plant. The unit of electricity is watts per hour, kilowatt per hour, megawatts per hour, gigawatt per hour and terawatt per hour. For example, New, energy-efficient refrigerators use about 300-400 kilowatt-hours per year.

⁸ Hashrate is the speed of mining. A unit of hash/second means how many calculations per second can be performed. Machines with a high hash power are highly efficient and can process a lot of data in a single second. Hashrate is usually measured in units of k (kilo, 1,000), M (mega, one million), G (Giga, one billion), T (tera, 1 trillion), P (Peta, one quadrillion), or E (Exa, one quintillion).

⁹ According to the government of Iceland, about 85% of the total energy supply in Iceland is derived from domestically produced renewable energy sources. This figure is the highest share of renewable energy in any nation.

The Hidden Side of Digital Technologies for Family Use: Privacy and Other Related Ethical Issues in Digital Distance Education and Telecommuting at Home

Ryoko Asai¹ and Sachiko Yanagihara²

¹Ruhr University Bochum, Bochum, Germany ryoko.asai@rub.de

²University of Toyama, Toyama, Japan

sachiko@eco.u-toyama.ac.jp

Abstract. Due to the lockdown induced by the coronavirus pandemic, children were forced to study at home. Many parents of school-age children had no choice but to telecommute, while tending for their youngsters. In this study, we primarily examine the extent to which telecommunication proved effective for taking care of children. It also examines the issues that stem from the attempts to balance professional and family life, especially when professional and educational activities take place within the same household, concurrently. Focusing on the aspect of privacy, the study observes the extent of the impact of telecommuting work patterns on a child's development. In the general technological context, privacy of information is of utmost importance. However, informational privacy is just one of the many complex characteristics of privacy. In this research, we aim to elucidate the complexity of privacy from the viewpoint of care, and to develop a certain know-how for the improvement of children's wellbeing in the digital environment.

Keywords: care, child wellbeing, digital distance education, privacy, telecommuting, telework

1 Introduction

The pandemic caused by COVID-19 (hereafter referred to as the “coronavirus pandemic”) greatly restricted not only people's work styles, but also their daily activities, completely changing their daily lives. Restrictions on people's behavioral patterns in their daily lives became essential as a measure to prevent infection, with people telecommuting/working from home with the use of ICT (Information and Communication Technology), and children staying and studying at home due to the closure of schools. Many families with school-aged children telecommute while taking care of their children, and survey results highlight the advantages and disadvantages of telecommuting (Meiji Yasuda Life Insurance, 2020).

In this study, we examine the extent to which telecommuting can effectively aid in caring for children. Further, the study scrutinizes the various aspects that need to be

focused on, to maintaining a balance between work and family life at home, along with the effect of telecommuting work patterns on a child's development. First, we summarize advantages and disadvantages of traditional telecommuting, and of distance learning. Second, we explain the ongoing situation of distant education during the nationwide school closure, due to the declaration of emergency from March to May 2020 in Japan, and explore how the current educational situation is different from the one before the pandemic. Third, we examine how the current situation, where parents' telecommuting and children's distance education are conducted at home at the same time, influences family relationships, especially focusing on information-sharing between parents and children, and children's privacy at home. In addition, we will also address some ethical concerns when children are considered a target user of digital technology actively introduced along with the spread of the corona pandemic.

2 Impacts of the Coronavirus Pandemic on Family Life in Japan

2.1 Extensive Changes in Lifestyle

Since the beginning of the corona pandemic in Japan, many companies have introduced telework as a working style to maintain social distance and safety protocols. The education sectors, the government and the education ministry have also accelerated introducing ICT in educational places under the GIGA (Global and Innovation Gateway for All) school project. ICT made it possible to work and study remotely as well as to contribute social distancing under the pandemic situation. Many people, thereby, learned how digitalization can function efficiently or inefficiently in their working place and society.

However, telecommuting, and online homeschooling pose the problem of boundaries between public and private sphere, with the intersection and blurring of boundaries. This means, along with introducing telecommuting in working places and distance education in schools, people's activities in the public sphere (work in the office and education at a classroom) are brought into the private sphere (personal/family life space or at home). This drastic change in the lifestyle has caused conflicts between the "public" and the "private," especially in the private sphere (at home).

To improve their work-life balance, individuals have worked at home even prior to the pandemic, regardless of whether they were taking care of children or older family members. The government has also been focusing on and promoting telecommuting, and other forms of telework, such as "workation/workcation" (combination of working and taking a vacation), which makes it possible for employees to work using the Internet while taking a vacation.

While some flexible working styles were promoted by the government and companies, the impact of telecommuting on families, especially on children, has not been paid much attention to. Even when this kind of issue is picked up as a social

topic, it has mostly been discussed from the workers' perspective, such as an "unpaid overwork problem," the way to manage working hours, security, technological problems in working from home, and so on. However, it is unlikely that the impact on children is considered. In other words, there has been little attention paid to the impact of a shift in parents' working style on children. The strict protocols during the coronavirus pandemic led to children being restricted from going to school, playing outside, and meeting friends, leading to their communication activities occur in a virtual space, supported by digital technology.

At present, although vulnerability to the virus remains, vaccinations have progressed nationwide, and various restrictions have been progressively lifted, to resume normal activities in work and education restarting as much as possible. While telework, including telecommuting and remote work, became widespread during the current pandemic, whether this new style of work will become standard in the future, remains elusive. To adapt to it as a normal work style, the government, and industries need to coordinate and conduct it in a manner suitable for them and their workforce (Yanagihara, 2021).

2.2 Distance Education in Japan

Prior to the coronavirus pandemic, the Ministry of Education, Culture, Sports, Science and Technology (MEXT) had been promoting the "GIGA (Global and Innovation Gateway for All) School Concept" in elementary and secondary education, and ICT-based distance education was already stipulated in the School Education Law. However, only a few examples proved worthy enough, to carry out the project in schools.

In elementary and secondary levels, distance education did not attract people to use ICT for daily academic activities, and the implementation ICT in school administration, did not prove to be a smooth process. However, the nationwide school closures at the end of February 2020, due to the coronavirus pandemic, became a game changer for ICT use in compulsory education. Children needed continued education provided by schools, despite closed schools, as a basic human right. To compensate the absence of classroom education, the government accelerated the implementation of the GIGA project in schools.

Meanwhile, the project required teachers to additionally work toward management of such projects, without reforming their working style. The workload increased for teachers, with their requirement to acquire digital skills for managing the project, provide digital forms of education to kids in class, and perform their usual tasks. Now it is time to reconsider, socially, about what education is, how ICT needs to be utilized for education, how the workload and tasks of teachers should be managed etc.

The GIGA school project radically promoted digitalization in schools. Simply put, the GIGA school project aimed to provide a digital tablet to each pupil studying in public primary and secondary school nationwide by the year 2021. The reason the project was rapidly conducted, was to improve ICT use in school settings. While one digital tablet per each pupil was allocated in public primary and secondary schools, pupils were not allowed to use their tablet as they wished. Schools offered pupils very

limited time and ways when used in classrooms and at home. Even some local governments do not allow children to use their tablet freely anywhere (CNET Japan, 2021). In other words, limited use of digital devices helps teachers and schools prevent school children from causing or getting involved in online security problems beforehand, such as leaking personal information, accessing inappropriate websites and so on.

Moreover, schools rely on parents and families for establishing and maintaining the ICT environment, including internet access at home. Under the circumstances, even if schools distribute digital tools to all pupils, each family has a different ICT environment and the difference between families would affect children's learning process and ability. Furthermore, parents' and guardians' attitudes toward education and ICT would generate some unfair learning conditions for children. Some children can get the sufficient support and learning environment from their family, and some cannot. Therefore, to avoid any potential risks for damaging equal education opportunities for children, many schools allow children to use digital devices under the very limited conditions.

3 Telework and Distance Education as a Concurrent Event at Home

3.1 Telework as a Means of Improving Work-Life Balance

Over the past decades, research on telecommuting and distance education has been conducted separately. Normally, in Japan, it was not expected that parents and children would simultaneously engage in their "public" activities (telecommuting for guardians and distance education for children) at home. Specifically, distance education has not been considered in the context of children's private life at home. That is because education for children generally and mostly happens at school, where is a public sphere for children, until the coronavirus pandemic.

For adults, children are persons to be protected by guardians, adults, and society, and have been recognized as part of the private lives of their parents. The younger the child is, the more involved the parents need to be in the child's life. Telecommuting has been considered as one way to enable parents to be more flexible in maintaining balance between work and family-care. Especially, in many families in Japan, there are children who have to stay at home by themselves and wait for their parents coming back home from work.

While telecommuting is socially promoted as a means for work-life balance, concerns regarding information leakage to family members and online security at the home have traditionally been big concerns for telecommuters and companies. We can easily imagine that some families, no matter how careful they are, cannot separate between a working room and other rooms in the same house/apartment. In fact, along with applying distance education based on the online network, information leakage has been recognized as a big risk in telecommuting as well. The Japanese education ministry had to issue various urgent recommendations, guidelines and policies in

order to protect personal information in conducting distance education digitally (Information Education Division, Lifelong Learning Policy Bureau, Ministry of Education, Culture, Sports, Science and Technology, 2016; Ministry of Education, Culture, Sports, Science and Technology, 2017;2021)

In many cases, parents feel extremely psychologically and physically pressured to work efficiently, both at home and the office, while working with their children or elderly parents in the same residence. However, we barely consider how parents' telecommuting would influence children's life and how distance education at home would affect children's development psychologically and physically. Under the situation that parents' telecommuting and children's distance education are simultaneously conducted at home, children could feel some stressed studying at home. For example, parents are sometimes watching their children studying or interacting with teachers and others via internet while they are working at the same time. It is not hard to imagine that this kind of situation could stress children or influence them mentally.

In Japan, nuclear families have been increasingly common, and it is very ordinary that children, whose both parents work in the office until the late evening, usually must return home by themselves from the primary school and stay home alone until their parents get home. The absent of parents at home could cause stress or psychological impacts on children, and, in the worst-case scenario, some incidents might occur due to children being home alone. According to previous research on telework, telework allows parents to stay home with their children, and contributes to keep a home environment safe and stable for school-aged children. Even if parents are working at home without sitting with their children in the same room, the fact of staying with children at home could give peace of mind to children and their communities as well as to parents (Yanagihara, 2019).

However, there are only few studies that have examined children's perception of their parents' telecommuting and the impact of telework on them. This means that conventional telecommuting research has been conducted from the workers' perspectives, regardless of marital status and family composition. Under the current circumstance that telecommuting, and distance education are carried out at home at the same time, it is necessary to examine privacy issues and ethical issues with child development over digitalized family life at the home setting. Furthermore, the practice of telecommuting under the coronavirus pandemic can be seen as a touchstone of telecommuting in the DX (digital transformation) era.

3.2 An Expectation Gap Between Children and Parents

While accompanied by confusion and perplexity, telework has been perceived positively to some extent (Japan Productivity Center, 2021a: pp.16-21). However, official restrictions and requests for self-restraint to stop the spread of COVID infection have been prolonged, and a negative trend toward satisfaction with and effectiveness of telework has emerged (Japan Productivity Center, 2021b: pp. 17-22). There is also the problem of telework further increasing the burden of household

chores and childcare for women when both husband and wife or a husband work from home at the same time (FNN, 2021). The survey by Japan Productivity Center points out the lack of "systems and mechanisms to reduce the burden of housework and childcare" and "family cooperation" as the issues of telework.

This situation is also stressful for the children, who are unable to go out or socialize with friends, and must study at home, while being concerned about the coronavirus infection. A survey about health focusing on children and their parents during the pandemic indicates that the children's mental and physical health conditions is declining (National Center for Child Health and Development, 2021). While children's "wish to go to school" and "enjoyment" are greatly reduced, many children feel an increased "difficulty in studying" (pp. 49-53). Furthermore, due to the increase in family time at home, parents feel that they have more parent-child time, while children feel that they have fewer conversations with their families. This indicates that there is a difference between parents and children in their ideas of how they expect "parent-child time" to be spent (pp. 49-58).

In other words, parents are satisfied with being in the same space as their children, and children expect not only to be in the same space as their parents but also to do activities with their parents. Telecommuting allows parents to be in the same space as their children, but there are difficulties for them to enjoy conversations with their children and to help them with homeschooling. The coronavirus pandemic has prompted a rethinking of the benefits of telecommuting for families and the role parents should play for their children.

4 Privacy at Home

4.1 Children's Privacy at Home

Generally, when we discuss the concerns in conducting telework and distance learning in the home setting, privacy is considered as one of the most significant issues, socially. In this case, the main subject of discussion about privacy is the risk that the personal information of each family member, or important information of organizations (workplace or school) would be leaked to and/or violated by third parties. The concept of privacy, however, consists of three different elements: informational privacy which is equivalent to personal information; spatial privacy which is related to physical space and integrity; decisional privacy which influences a person's decision making and a freedom of choice (Kang, 1998; Levesque, 2017).

Privacy issues caused by the concurrent occurrence of teleworking parents and distance-learning children at the same residence, need to be examined not only from an informational privacy aspect focusing on information-leak, but also from the conjunct perspective of three different privacy elements. This approach would help us to get deeper insights about ethical concerns and implications of telework and digital distance learning. The latter perspective can enlighten us about of the process of respecting individual privacy should be respected even in families, and contribute to improving children's wellbeing by considering the characteristics and necessities of

children's privacy. Additionally, it can lead to developing educational practices for cultivating a sense of privacy and learning to respect others' privacy.

Generally, children's informational privacy is often regulated by parents at home. Many digital devices and applications including television and broadcasting services, have special features to regulate children of certain age, from accessing them and their contents. However, even if strict measures are taken by parents from facing any risks online, many a child faces the dangers of privacy-leak, phishing, sexual abuse, and so on. Under this precarious online situation, parents try to increasingly control and regulate their children's use of digital tools, and consider it their responsibility to protect their offspring's privacy. Especially in Japan, children have been traditionally considered as an extension of parents. This traditional family norm limits a child's independence and freedom.

Regarding digital distance learning, it is natural and rational to want to regulate children's online access and use of digital devices a priori, to protect them from cyber security risks. Parents and schools believe that they protect children under their responsibility. Furthermore, families and teachers who care for the children will be satisfied that they can encourage children's rapid adjustment to family, social or group life by intervening in children's privacy. However, children also have their own life and privacy depending on where they are. We can imagine that children take different attitudes at home, at school, in front of their friends and so forth.

However, when children have to stay at home with parents, with no exposure to the external world, unavoidably, certain unseen parts of their behaviors and personalities are exposed. This results in the lack of privacy within the family at home. Involuntary exposure to family under an unintended environment could stress children as well as parents. As parents do, children also need spatial and decisional privacy as well as informational privacy. In this context, spatial and decisional privacy give us a clue to see the family privacy problem from the different aspect. For example, do children need special protection? How much freedom do children need? How effective are excessive restrictions on children's development?

4.2 Complexity of Privacy in Family Life

The impact on privacy from telecommuting and distance learning, which can occur at home at the same time, requires consideration of the mutual impact on privacy within a family, as well as of information leakage. It is undeniable that the spatial privacy of both parents and children is affected when work and study, which used to be activities in public places outside the home, are brought into the same residence. It is a situation in which one must undeniably show one's family members something that one would not normally show them. In addition to the issue of personal information leakage, it promotes an understanding of how personal privacy should be respected, even among family members. For parents, it provides an opportunity to understand their children's need for privacy, and for children, it provides an opportunity to develop a sense of privacy and learn to respect privacy of others. In this situation, the value of information shared within the family is examined through privacy.

According to Fried, the concept of privacy allows individuals to control their own information and knowledge about themselves, including the quantity and quality of the information (Fried, 1968). Privacy could lead to emergence and development of the moral capital in relationships involving love, friendship, and trust. This in turn indicates its importance in building an intimate relationship, with a partner or family, and in developing independence of an individual. However, privacy consists of complex characteristics. When we care about someone and try to establish a relationship based on love, friendship, and trust, we inevitably step into their privacy and increase the risk of invading their privacy. Simply put, the more you care for them, the more you face the risk of violating their privacy. Then, how do we know how to create a balance between care and violation of privacy? According to Fried, it will depend on the balance between privacy that is compromised and the value of something that is gained (Fried, 1968).

Privacy in family life holds higher complexities than privacy in a general context, as it is dependent on family relationships and circumstances. In other words, , due to the inherent love, trust, and care present in the family, there is constant violation of each other's privacy. Public activities like work and education occurring in the external world, were inadvertently pushed into the domestic environment as a result of nation-wide lockdowns. This situation highlights the need for protection of privacy even in family relationships. Therefore, it is very important for parents to understand their children's need for privacy, while children need to develop an awareness of “privacy” and learn to respect privacy of others including family members.

5 Conclusion

With the aim of improving wellbeing, which encompasses not only people's physical and mental health but also a better relationship with the environment they live in, ethical considerations for digital wellbeing based on human-computer interaction will become even more important in the future (Floridi, 2014; Burr, Taddeo, & Floridi, 2020). The emerging technologies are changing not only our living standards dramatically, but also our values and norms related to the way we live with others. Family life is not exceptional either, in this social stream. Among many ethical concerns in the high-tech era, privacy is still very essential for our existence and needs to be reconsidered in the context of living with emerging technologies.

For example, the coronavirus pandemic resulted in the rapid growth in the sales of social robots that serve as a playmate for children and the accelerated use of educational applications (Asai, 2021). In Japan, each of all school-aged children (in mandatory education) gets a digital tablet distributed by schools, for its mandatory usage in educational activities regularly in school and at home. However, the development of children's digital literacy and the positive effects of ICT, are dependent on its usage. The education ministry along with the educational division of city offices and schools, prepared various proposals, guidelines, and policies about the usage of these digital devices.

The ministry and schools mostly focus on how to prevent the information leakage such as personal information, among other issues (Information Education Division, Lifelong Learning Policy Bureau, Ministry of Education, Culture, Sports, Science and Technology, 2016; Ministry of Education, Culture, Sports, Science and Technology, 2017). In an extreme case, due to strong concerns of the information leakage and cyber security, schools allow pupils to use a digital tablet only in a specific subject time, with all pupils' account information and passwords are controlled by school. In contrast, developing children's ethical competence, which is often included in "digital media literacy," should be considered as a part of ICT use at school. Now, it is time to seriously consider how to interact with technology at home as well as in school in order to enhance children's digital wellbeing.

Through this study, we tried to situate privacy concerns at home as invisible but serious ethical problems in child development, and elucidate how to comprehend children's privacy in the context of family by referring the three elements of "privacy." Heretofore, a lot of research, especially in computer ethics and business information management areas, have discussed on the practical prevention of the information leak. In other words, children's privacy has been overlooked, as it was perceived as a part of parents' privacy. However, excessive restrictions based on security considerations can have negative effects in terms of fostering a sense of privacy and autonomy, which are important for children's development and wellbeing. In the digital society, we need to understand privacy from the perspective of "care," and find ways to develop children's digital media literacy including awareness of privacy and respect for others' privacy.

From the viewpoint of care, telecommuting could affect children's privacy because of the complexity of privacy. Whereas we promote telework as a solution to take a work and life balance, it needs to be reconsidered from the children's perspective. In order to enhance family wellbeing, we need to further explore more about the impacts of parents' telework at home and their work-life balance on children's wellbeing as a next step.

The rapid social change caused by the coronavirus pandemic, required personal and social reassessments of the relationships between people and between human and technologies. This reassessment has extended to family relationships between parents and children, and between families and technologies. the extend of ICT's effective support on family life, and it counted as one of life's essential commodities. Meanwhile, the positive impacts make invisible ethical concerns obscure, and we didn't pay proper attention to the negative effects of ICT use at the home setting. Therefore, there is a to repeatedly examine how emerging technologies affect children, those who take over this world from older generations and create their own world in the future.

Acknowledgements

This research was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant-in-Aid for Scientific Research (C) Number JP20K12551 and JP21K01650.

References

- Asai, R. (2021). Ethics and Social Robots: How do I live with a social robot? *Proceedings of ETHICOMP 2021*, pp.281-284.
- Asahi Shimbun online. (2020). "Telework, Child Day-Care at Home Stresses Parents. What Can We Do to Reduce Stress at Home?" <https://www.asahi.com/articles/ASN4T6KCWN4RUPQJ003.html>. Accessed 15 May 2022.
- Burr, C., Taddeo, M., & Floridi, L. (2020). The Ethics of Digital Well-Being: a Thematic Review. *Sci Eng Ethics* 26, pp.2313-2343.
- CNET Japan. (2021). "Not Bring Your Tablet to Home. A Shortage of Informatics Expert in School: How is GIGA School Project Going?" <https://japan.cnet.com/article/35169670/>. Accessed 15 May 2022.
- Floridi, L. (2014). *The Fourth Revolution*. Oxford: Oxford University Press.
- FNN Prime Online. (2021). "House-Keeping disparity between husband and wife," <https://www.fnn.jp/articles/-/196346>. Accessed 15 May 2022.
- Fried, C. (1968). Privacy. *The Yale Law Journal*, vol. 77, no.3, pp.475-93.
- Gender Equality Bureau. (2021). *Committee Report on the Impacts and Concerns on Women under the Coronavirus Pandemic*. https://www.gender.go.jp/kaigi/kento/covid19/siry o/pdf/post_honbun.pdf. Accessed 15 May 2022.
- Information Education Division, Lifelong Learning Policy Bureau, Ministry of Education, Culture, Sports, Science and Technology (2016), "Urgent Proposal for Educational Information Security", https://www.mext.go.jp/a_menu/shotou/zyouhou/detail/1377772.htm. Accessed 15 May 2022.
- Japan Productivity Center. (2021a). *The 5th Workers Report*. https://www.jpcnet.jp/research/assets/pdf/5th_workers_report_0617.pdf. Accessed 15 May 2022.
- Japan Productivity Center. (2021b). *The 6th Workers Report*. https://www.jpcnet.jp/research/assets/pdf/6th_workers_report.pdf. Accessed 15 May 2022.
- Kang, J. (1998). Information Privacy in Cyberspace Transactions. *Stanford Law Review*, 50, pp.1193-1294.
- Levesque, J. R. (2017). *Adolescence, Privacy, and the Law*. Oxford: Oxford University Press.

- Meiji Yasuda Life Insurance Company (2020), "Emergency Questionnaire Survey on Households Raising Children in Corona Peril," Meiji Yasuda Life Insurance Company,
https://www.meijiyasuda.co.jp/profile/news/release/2020/pdf/20200707_01.pdf.
 Accessed 15 May 2022.
- Ministry of Education, Culture, Sports, Science and Technology (2021), "*Guidelines for Educational Information Security Policy*" Handbook, Ver.May-2021,
https://www.mext.go.jp/content/20210630-mxt_jogai02-000011648_052.pdf.
 Accessed 15 May 2022.
- Ministry of Education, Culture, Sports, Science and Technology (2017), *Training materials on information security and ICT environment maintenance in schools*,
https://www.mext.go.jp/a_menu/shotou/zyouhou/detail/1403502.htm. Accessed 15 May 2022.
- National Center for Child Health and Development. (2021). *The 5th Survey Report on the Coronavirus Pandemic * Children*, https://www.ncchd.go.jp/center/activity/covid19_kodomo/report/CxC5_repo_20210525.pdf. Accessed 15 May 2022.
- Yanagihara, S. (2019). "Organizational citizenship behavior of telecommuters whose working hours are strictly controlled," *Journal of the Japan Society for Information Management*, 39(1), pp.57-67.
- Yanagihara, S. (2021). "How Human Coexistence with ICT in the Era of Telework Enabled Should Be: From the Viewpoint of Now and the Future of Telework in Japan," *Japan Labor Issues* 5(29), pp.21-36.
- Yokomaku, T. (2020). "Sharing Household Chores and Childcare at Home under the state of emergency," in Mitsubishi UFJ Research and Consulting Report,
https://www.murc.jp/wp-content/uploads/2020/05/survey_covid-19_200526.pdf.
 Accessed 15 May 2022.

Smart patches in mass-casualty incidents

Katerina Zdravkova¹ [0000-0002-9674-3081] and Ana Madevska Bogdanova² [0000-0002-0906-3548]

^{1,2} University Ss. Cyril and Methodius, Faculty of Computer Science and Engineering, Skopje,
N. Macedonia

katerina.zdravkova@finki.ukim.mk

Abstract. The favourite Macedonian greeting for birthdays and new births is: “Be alive, healthy and happy”. Mass-casualty incidents severely endanger them all mainly due to the unavoidable overcrowding where treating of one patient might threaten the welfare of those who are also waiting for treatment. The prevention of such situations during emergencies and disasters is fixed by efficient emergency triage process. The triage identifies life threatening conditions and their severity and determines whether the patient needs an urgent intervention or not. Objectivity of judgment is crucial and depends primarily on the patients’ heart rate, respiratory rate, oxygen saturation, and blood pressure. Smart patches keep track of all these vital parameters. They are an affordable emerging technology remotely connected to healthcare institutions. To be massively used, their drawbacks need to be carefully identified and reduced as much as possible. The goal of our research was to detect which are the major ethical challenges of this emerging technology and to propose solutions that will avoid them. To accomplish the goal, four research questions were identified, thoroughly reviewed, and the results of the ethical analysis were highlighted for each question. They embrace privacy, security, reliability, responsibility issues and the usefulness of smart patches. Based on the review, we propose recommendations for the development of new patch-like devices that will support the efficient detection of respiratory and cardiovascular changes of triage-labelled victims, without interfering with human rights and dignity.

Keywords: Bio-medical ethics, Mass-casualty incidents, Wearable devices

1 Introduction

In the last years, everyday life has become very fragile. Recent European floods; deadly Asian, Peruvian and Hispaniola earthquakes; North India landslides; North Pacific and Indian Ocean tsunamis; and the COVID-19 pandemic have posed significant risks to life. They place “a significant demand on medical resources and personnel” (Lee, 2010). If early diagnosis is crucial to prevent organism failure and death caused by the virus (Zhai et al, 2020), efficient triage systems are critical to prevent mass-casualty incidents due to emergencies and disasters (Bazvar et al, 2020). There are many different triage systems and methods. START (Simple Triage and

Rapid Treatment) is a simple system that is sufficiently effective to be used by lightly trained emergency response personnel (Gebhart and Pence, 2007).

Innovations in wireless communications and low-power electronics have contributed to the creation of effective wearable medical devices that support monitoring of vital signals (Khan et al, 2016). Fabricated of flexible and lightweight materials, they are less irritable for people who wear them. The results obtained from their sensor systems are wirelessly connected with the healthcare institutions.

Many new multifunctional smart patches have proven their efficiency, accuracy and relevance for advanced healthcare (Hwang, 2018). Their diagnostic abilities can lead to more personalized care, and give instant feedback of patient's condition, before it becomes serious and life threatening. Faster the reaction is, greater are the chances to save victim's life.

Unfortunately, frequent malware attacks against wireless sensor networks (Queiruga-Dios, 2016) and severe ransomware attacks against hospitals (Spence, 2017) can seriously endanger the security and reliability of health monitoring, threatening the health of already vulnerable patients. Apart from security problems, privacy of medical data has also been compromised (Ching and Singh, 2016). These problems affect smart patch devices similarly to all other wearable devices.

To overcome their occurrence, particularly during mass-casualty incidents, considerable attention should be paid to invent technical measures that will minimize the risk of threatening patient's life, instead of saving it.

Discrimination in providing a medical care is another problem (Williams, 2019). This is the first ethical challenge of the prospective emergency triage. Medical and biomedical ethics should "respect for autonomy, beneficence, nonmaleficence and justice" (Aacharya, 2011).

Based on these initial premises, the goal of our research was to investigate the following questions:

1. Can continuous monitoring of the essential health parameters become a privacy and security threat for their users?
2. Do smart patch devices ethically challenge the emergency triage more than traditional medical devices?
3. If smart patches fail to accurately present the health condition of the triage labelled victims due to technical or cyber challenges, who will be held liable for negligence?
4. If smart patches are approved in the triage process, who and how should make the decisions whether the triage-labelled victim can use them or not?

To answer the defined research questions, smart medical patches, their advantages and disadvantages were thoroughly researched, aiming to define technical and ethical recommendations for developing new patch-like devices that will be accurate, reliable and efficient. Thanks to their wide availability and affordable price, these medical devices will be able to reduce the detection time of respiratory and cardiovascular changes of the triage labelled victims, which are crucial in high-stress situations, such as mass-casualty incidents.

2 Related work

Emergency department triage aims to improve the quality of emergency care by quickly sorting patients to determine priority of further evaluation of care (Aacharya et al., 2011). During mass-casualty incidents, it is affected by massive overcrowding, thus it is crucial to decide who needs immediate treatment and avoid the two extremes: undertriage or underestimation of the severity of the patient's condition, and over-triage, i.e. assigning a higher acuity rating than necessary (Fernandes et al., 2005).

Depending on the observable vital signs, demographics, medical history and the injured part of the body, patients are classified with one of the five-level triage categories: resuscitation, emergent, urgent, semi-urgent, and non-urgent (ClayWilliams, 2020). Within the triage system, they get a visible triage tag with four colours: green for minor injuries; yellow for non-life threatening injuries that can get a delayed treatment; red for life threatening injuries that need an instant treatment; and black for expected death that need a pain medication only (Coleman et al., 2011).

START is a simplified triage strategy that is suitable for mass-casualty incidents. It is designed to identify problems that could cause death to the patient within one hour, typically breathing problems, head injury or significant bleeding. The next step is the need of observing if the victim is hemodynamically stable (if the blood pressure and heart rate are stable) before he/she is transported to a hospital. The idea is to use a specific design of a smart patch that includes electrocardiogram (ECG) and photoplethysmogram (PPG) sensors, providing easy calculation of heart rate, respiratory rate, oxygen saturation (SpO₂), and blood pressure (Lehocki et al, 2021). To be massively used, it has to be available, cheap, reliable and with few side effects. (Santos et al, 2018).

During some life-threatening situations, blood circulation is limited to the torso and head and thus monitoring blood oxygenation from finger can result in corrupted blood saturation reading due to poor blood circulation (Schreiner, 2010). Integration of PPG sensor into patch device placed on the chest reduces the delays of SpO₂ measurement by several seconds comparing to finger-based sensor, which is of great importance in emergency situations.

In 2014, IEEE standard for wearable cuffless blood pressure measuring devices was introduced (IEEE, 2014). They are embedded in the Internet of Medical things applications, enabling flexible, accurate and continuous and non-invasive blood pressure monitoring (Ibrahim & Jafari, 2019).

Smart patches, as wearable electronics can be equipped with different biosensors and can be used in different situations. They provide an opportunity for everyone to own personal healthcare systems. Most of the smart patches are self-powered with wireless transmission ability to monitor the temperature and motion status of individuals (Shi, 2017).

3 Ethical challenges of medical smart patches

Following four subsections elaborate the research questions defined in the introduction of this paper in more detail aiming to deduce how to decrease the probability of future occurrence of the perceived ethical challenges of medical smart patches used in the emergency triage system.

3.1 Can continuous monitoring of the essential health parameters become a privacy and security threat for their users?

Collecting data is an issue that needs special attention. Whenever possible, the data is kept locally and used with no recording. The main approach in developing multisensory wearable patch-devices is to maintain the functionality within the device. If the parameters can be obtained using the patch memory and processing power, no data is recorded but only their momentarily heart and respiratory rate values, needed for the triage process. However, the remaining two parameters, SpO₂ and BP, need further data processing on a mobile device – tablet or on a remote server. The first approach uses a Bluetooth connection to the sensor and the tablet serves as a platform to the model for SpO₂ and blood pressure estimation, so the data stay locally on the tablet (Fedor, 2021).

Privacy and confidentiality of medical devices should be carefully safeguarded to avoid accidental disclosure of sensitive health information (Berlan and Bravender, 2009). Unfortunately, most wearable medical devices keep the collected data in their original form without any encryption (Vijayalakshmi et al., 2018). They continuously monitor the changes of patient's condition and send them remotely to computers in medical institutions for further processing (Vijayalakshmi et al., 2018).

Remote monitoring systems are still not appropriately secured allowing access to transmitted sensitive data and enabling their interception, generation or modification (CISA, 2021). Sun et al. (2018) claim that healthcare organizations are usually indifferent about security and privacy, forgetting that they produce a vast amount of highly sensitive data (Sun et al., 2018). According to their review, major privacy and security challenges are related to insecure networks, rather light security protocols, and insufficiently protected data sharing. To prevent capturing of sensitive medical data, they suggest encryption in the phases of their storing, accessing, searching and transferring. Additionally, they recommend trusted third party auditing and data anonymization. Their impression that “more successful exploration is needed” is vital for the embedded remote medical devices.

How do poor protection of stored data and vulnerable communication affect smart patches? Smart patches are self-adhesive or tattooed on patient's skin, thus they cannot be accidentally lost or stolen. This is their great advantage to traditional wearable medical devices. However, they use the same data transfer MQTT protocol for remote communication as IoT (Montgomery et al., 2018). The authentication is either optional or not encrypted, making the protocol “highly susceptible to man-in-the-middle attacks” (Combs, 2022). Since 2014, around 90 vulnerabilities of this protocol have been identified (Combs, 2022). This might be one of the major reasons

why in September 2021, more than 60 million fitness tracker records of Apple and Fitbit users were compromised, revealing many private non-medical information (McKeon, 2021).

Vulnerable data transfer protocols and observed risks of massive data breaches prove that continuous monitoring of essential health parameters might cause accidental disclosure of confidential medical information that are collected in smart patches. This undesirable challenge is additionally amplified by the fact that most patients trust that their personal information is well protected (Cilliers, 2019).

Synergy of inevitable continuous monitoring of the triage-labelled victims, privacy and security risks and patients' lack of awareness about these risks might obstruct massive use of smart patches during mass-casualty incidents.

3.2 Do smart patch devices ethically challenge the emergency triage more than traditional medical devices?

The major challenge of mass-casualty incidents is the unavoidable overcrowding. It inevitably causes delays in providing care or prioritization of some patients, which threatens the welfare of clinically urgent patients who have wait for a treatment (Aacharya et al., 2011). Wearable medical devices, including smart patches are part of the pre-hospital triage and they are crucial to determine patients' flow when clinical needs exceed capacity of emergency departments (Aacharya et al., 2011). Their obligation is to comply with the four principles of biomedical ethics: autonomy, nonmaleficence, beneficence and justice (Hall & Smithard, 2021).

The principle of respect for autonomy is the right of a patient to accept or reject medical treatment, even when the patient has lost consciousness (Gillon, 1994). The decision is based on patient's personal values, fears and beliefs (Aacharya et al., 2011). To overcome the problem, good physician - patient communication should be established, which is difficult in emergency cases. Those patients who accept medical treatment usually become impatient, hindering the treatment of other patients and challenging the principle of nonmaleficence.

Nonmaleficence can be simplified with the principle "do no harm", and it is part of the Hippocratic Oath (Aacharya et al., 2011). In overcrowded mass-casualty situations, the obligation of emergency department is to provide the reasonably best care without discrimination and ill intent from the medical staff. Unfortunately, high stress situations usually worsen patients' psychological state, which includes "stress, fear, feeling neglected or not being taken care of" (Aacharya et al., 2011).

Beneficence is also part of the Hippocratic Oath and represents a moral obligation to contribute to the benefit and well-being of people (Aacharya et al., 2011). One of the main challenges of this principle in the mass-casualty situation is overtriage, i.e. overestimation of the urgency and incorrect prioritization of less urgent patients to those who need more urgent care. Overtriage triggers inefficient use of medical staff and resources, increasing the cost and reducing the effectiveness of urgent care.

Justice is related to distributive justice realizing the ethical values of equality, utility and priority to the worst-off (Marlink, 2017). Although triage systems strive to

equally and justly distribute all the resources, in overcrowded situations, patients will get “a fair share based on appropriate criteria and principles” (Aacharya et al., 2011).

Review of metrological properties of wearable medical devices discovered that there is still not a standard test protocol for their validation (Cosoli et al., 2020). Accuracy of wearable heart rate monitors in cardiac rehabilitation is inferior to accuracy of clinical electrocardiographic monitors (Etiwy et al., 2019). Additionally, they may be affected by transmission failures caused by “communication channel disconnection, power loss, power off, and interruption of user biometric information sensing” (Lee, 2021). Finally, wearable devices are prone to failures (Koydemir, 2018).

These reliability problems significantly affect the four principles of biomedical ethics. Although autonomy is not explicitly disturbed, nonexistence of an accurate information about patient’s health condition, if such an information exists, seriously devastates the remaining three principles: nonmaleficence, beneficence and justice. Autonomy depends on the patient’s awareness of own health state. All the data stored on medical smart patches are accessed by the clinicians only (GlobalData, 2109). It means that patients do not know their current state. It is actually a double-edged sword: on one hand, lack of knowledge about a critical health state keeps the patient calm and prepared to wholeheartedly collaborate with the emergency department, on the other, lack of information affects the feeling of helplessness and increases panic, even when there the condition is not life-threatening.

It is extremely difficult to determine whether medical smart patches ethically challenge the emergency triage more than the traditional ones. They are still in their infancy, and their reliability is still developing. Their major advantage is that they are affordable, widely available and accurate enough to support the emergency department triage and reduce overcrowding of mass-casualty incidents.

3.3 If smart patches fail to accurately present the health condition due to technical or cyber challenges, who will be held liable for negligence?

Nash (2021) examines the liability threats resulting from smart devices hacking emphasizing the lack of comprehensive legislative framework to improve their security. He claims that many manufacturers deliver solutions with embedded security vulnerabilities (Nash, 2021). He also examines product liability associated with product failures or defects, stressing that “failures in technology can result in direct physical discomfort, harm and mortality” (Nash, 2021).

Liability of medical smart devices in Europe is regulated according to Regulation 2017/745 (EUR-Lex, 2017). Although not explicitly stated, liability of smart patches, which use nanoparticles refers to manufacturers. According to this regulation “such devices should be subject to the most stringent conformity assessment procedures” (EUR-Lex, 2017). Together with manufacturers, all stakeholders in the supply chain may also be subject to liability.

Typical side effects of wearable devices to patients’ health include: headaches, dizziness, discomfort, and musculoskeletal disorders (Xue, 2019). None of these symptoms have been registered in smart patches. Due to their minuscule dimensions,

safety concerns embracing awkward postures, forceful exertions, physical fatigue, and mental vibration are not applicable to smart patches (Schall et al., 2019). A small problem can be the durability of the sensor and the risk of damaging the smart patch by sweating. Whenever failures occur, smart patches create a potential risk of harming patients (Parimbelli, 2018). Their medical liability in the emergency triage predominantly refers to manufacturers and to medical staff. Manufacturers include knowledge engineers, system developers and maintenance engineers. Medical staff consists of physicians, nurses, and the hospital that is responsible for monitoring the results. For communication failure, responsibility passes on the network service provider. In all these examples, the patient is excluded. The patch is attached to the patient's chest while the patients are in need of medical attention. The patch can be used only if it becomes a part of an emergency medicine protocol, so it is agreed upon the utilization of the patient's data. Patients do not interact with the patch, unless they intentionally remove or destroy them. In such case, they use their autonomy right to refuse medical care, and are therefore no longer part of the emergency triage.

3.4 If smart patches are approved, who and how should make the decisions whether the triage-labelled victim can use them or not?

In the beginning of COVID-19 pandemic, most health systems prompted rapid transition towards virtual care. Wearable smart medical devices played a vital role in self-diagnosing of coronavirus and enabled observing of the health condition of people with visible signs and symptoms. During this massive and long-lasting mass-casualty incident, people with mild or moderate symptoms managed to recover without hospitalisation, preventing the instant collapse of health care system. They could rely on the devices they already had or purchased online. The only prerequisite for such brave decision was the ability to get access to available medical smart devices.

Similarly, to most wearable accessories available on the market, medical smart patches are affordable, so the first prerequisite for their widespread use during masscasualty incidents is met (Omerov et al. 2021). Their penetration is still limited to developed countries, mainly due to network availability and regulatory constraints. However, it is expected that global patient monitoring market will increase tremendously, with a projected value of 30 billion US\$ by 2026 (Businesswire, 2022).

Although used by almost one quarter of US population, very few wearable medical monitoring technologies are FDA-approved (Phaneuf, 2022). It is expected that their compliance with existing legislation will soon increase, enabling their broader adoption for real-time monitoring of health conditions, which is a key requirement for improving the survival rate during mass-casualty incidents.

If they are approved, the decisions whether the patient can use them or not should be made without introducing any kind of discrimination (Camporesi & Mori, 2021). According to Montgomery et al. (2018), profiling algorithms and techniques for selecting the eligible candidates and usually biased, enabling discrimination on the basis of "ethnicity, gender, sexual orientation, age, community, or medical condition". During mass-casualty incidents, biased profiling methods should be either avoided as

much as possible providing all people with an equal opportunity for real-time monitoring and urgent medical treatment, whenever their health is endangered.

High affordability of medical smart patches makes every person an eligible candidate for their use. If affected people are mentally able to make their own decisions, they should decide whether to get the smart patch or not, using their principle of autonomy (Zdravkova, 2017). Otherwise, the decision should be made by their relatives, if any, or by the medical staff, obeying the principle of beneficence and justice, which is guaranteed by the triage protocol of smart patches (Jaigirdar et al., 2019). It will guarantee the right to equality and non-discrimination.

4 Conclusions

In the coronavirus era, patient monitoring market has significantly increased. Smart medical devices have supported the health sector, preventing exceeding of hospital resources. They enabled real-time monitoring of various physiological signals.

Smart patches are affordable wearable sensor-based devices used in the medical industry. They are integrated into the wider concept of Internet of Medical Things (IoMT) and consequently, experience many technologies and user-related concerns and challenges, embracing privacy, security and reliability barriers (Loncar-Turukalo et al., 2019). So far, these challenges obstruct the massive adoption of smart wearables. Hopefully, several technological giants started investing in the creation of accurate, reliable and affordable smart medical devices. They are experienced to apply the four principles of biomedical ethics by design.

Privacy and security protection will be significantly improved by developing a more sophisticated data transfer protocol for remote communication. In such case, continuous monitoring of the essential health parameters will not harm privacy and security of collected and transmitted medical data.

Additionally, to protect the patients, a written informed consent should be signed by the patient or by the personal representative (Nakikj and Mamykina, 2017). The signed form is a legal document that will keep the physicians in the loop and let them go ahead with the treatment.

Standardisation of smart wearables will contribute to their higher reliability. It should be supported by innovative technologies, inventive materials and highly developed deep learning methods. The synergy between various technologies embedded in the creation of these devices will bypass the most common software bugs and hardware failures. They will gain more trust, which will contribute to faster approval by legal authorities. In such case, manufacturers will be obliged to obey the legal norms and reduce the risk of liability challenges.

By reducing multiple technological barriers, smart patches will become a reliable partner in the emergency triage systems. They will mitigate the inevitable panic level during mass-casualty incidents. Relying on their diagnostic power, health care system will be relieved without initiating any ethical challenges. Supported by new protocols, they will provide “the right care, at the right time, via the right medium” (Croymans et al., 2020).

Acknowledgement

This work was supported in part by grants from the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje and "Smart Patch for Life Support Systems" - NATO project G5825 SP4LIFE.

References

- Aacharya, R. P., Gastmans, C., & Denier, Y. (2011). Emergency department triage: an ethical analysis. *BMC emergency medicine*, 11(1), pp.1-pp.13. <https://doi.org/10.1186/1471227X-11-16>
- Bazyar, J., Farrokhi, M., Salari, A., & Khankeh, H. R. (2020). The principles of triage in emergencies and disasters: a systematic review. *Prehospital and disaster medicine*, Volume 35(3), pp. 305-313. <https://doi.org/10.1017/S1049023X20000291>
- Berlan, E. D., & Bravender, T. (2009). Confidentiality, consent, and caring for the adolescent patient. *Current opinion in pediatrics*, Volume 21(4), pp.450-pp.456. <https://doi.org/10.1097/MOP.0b013e32832ce009>
- Businesswire (2022). Global Wearable Healthcare Devices Market Share 2021- 2026. Retrieved from <https://www.businesswire.com/news/home/20220106005670/en/GlobalWearable-Healthcare-Devices-Market-Share-2021--2026---Trackers-Will-Continueto-Dominate-the-Market-in-2026---ResearchAndMarkets.com>
- Camporesi, S., & Mori, M. (2021). Ethicists, doctors and triage decisions: who should decide? And on what basis? *Journal of Medical Ethics*, 47(12), e18-e18. <https://doi.org/10.1136/medethics-2020-10649>
- Ching, K. W., & Singh, M. M. (2016). Wearable technology devices security and privacy vulnerability analysis. *International Journal of Network Security & Its Applications*, Volume 8(3), pp.19-pp.30. <https://doi.org/10.5121/ijnsa.2016.8302>
- Cilliers, L. (2020). Wearable devices in healthcare: Privacy and information security issues. *Health information management journal*, 49(2-3), pp.150-pp.156. <https://doi.org/10.1177/1833358319851684>
- CISA (2021). Medtronic Conexus Radio Frequency Telemetry Protocol (Update C), Retrieved from <https://www.cisa.gov/uscert/ics/advisories/ICSMA-19-080-01>
- Clay-Williams, R., Taylor, N., Ting, H. P., Winata, T., Arnolda, G., Austin, E., & Braithwaite, J. (2020). The relationships between quality management systems, safety culture and leadership and patient outcomes in Australian Emergency Departments. *International Journal for Quality in Health Care*, 32(Supplement_1), 43-51. <https://doi.org/10.1093/intqhc/mzz105>
- Coleman, C. N., Weinstock, D. M., Casagrande, R., Hick, J. L., Bader, J. L., Chang, F., ... & Knebel, A. R. (2011). Triage and treatment tools for use in a scarce resources-crisis standards of care setting after a nuclear detonation. *Disaster medicine and public health preparedness*, 5(S1), S111-S121. <https://doi.org/10.1001/dmp.2011.22>

- Combs, V. (2022). *Kaspersky: Many wearables and healthcare devices are open to attack due to vulnerable data transfer protocol*. Retrieved from <https://www.techrepublic.com/article/kaspersky-many-wearables-and-healthcare-devices-are-open-to-attack-due-to-vulnerable-data-transfer-protocol/>
- Cosoli, G., Spinsante, S., & Scalise, L. (2020). Wrist-worn and chest-strap wearable devices: Systematic review on accuracy and metrological characteristics. *Measurement*, 159, 107789. <https://doi.org/10.1016/j.measurement.2020.107789>
- Croymans, D., Hurst, I., & Han, M. (2020). Telehealth: The right care, at the right time, via the right medium. *NEJM Catalyst Innovations in Care Delivery*, 1(6).
- Etiwy, M., Akhrass, Z., Gillinov, L., Alashi, A., Wang, R., ... & Desai, M. Y. (2019). Accuracy of wearable heart rate monitors in cardiac rehabilitation. *Cardiovascular diagnosis and therapy*, 9(3), 262. <https://dx.doi.org/10.21037%2Fcdt.2019.04.08>
- EUR-Lex (2017). Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices. Retrieved from <https://eur-lex.europa.eu/legalcontent/EN/TXT/HTML/?uri=CELEX:32017R0745&from=EN>
- Fernandes, C. M., Tanabe, P., Gilboy, N., Johnson, L. A., McNair, R. S., Rosenau, A. M., ... & Suter, R. E. (2005). Five-level triage: a report from the ACEP/ENA Five-level Triage Task Force. *Journal of Emergency Nursing*, 31(1), 39-50. <https://doi.org/10.1016/j.jen.2004.11.002>
- Gebhart, M. E., & Pence, R. (2007). START triage: does it work?. *Disaster Management & Response*, 5(3), pp.68-pp.73. <https://doi.org/10.1016/j.dmr.2007.05.002>
- Gillon, R. (1994). Medical ethics: four principles plus attention to scope. *Bmj*, 309(6948), 184.
- GlobalData (2019). Wearable Technology in Healthcare – Thematic Research, Retrieved from <https://store.globaldata.com/report/wearable-technology-in-healthcare-thematicresearch/>
- Hall, H., & Smithard, D. G. (2021). A principlist justification of physical restraint in the emergency department. *The new bioethics*, 27(2), 176-184. <https://doi.org/0.1080/20502877.2021.1903152>
- Harford, M., Catherall, J., Gerry, S., Young, J. D., & Watkinson, P. (2019). Availability and performance of image-based, non-contact methods of monitoring heart rate, blood pressure, respiratory rate, and oxygen saturation: a systematic review. *Physiological measurement*, Volume 40(6), 06TR01. <https://doi.org/10.1088/1361-6579/ab1f1d>
- Hwang, I., Kim, H. N., Seong, M., Lee, S. H., Kang, M., Yi, H., ... & Jeong, H. E. (2018). Multifunctional smart skin adhesive patches for advanced health care. *Advanced healthcare materials*, 7(15), 1800275. <https://doi.org/10.1002/adhm.201800275>
- Ibrahim, B., & Jafari, R. (2019). Cuffless blood pressure monitoring from an array of wrist bioimpedance sensors using subject-specific regression models: Proof of concept. *IEEE transactions on biomedical circuits and systems*, 13(6), 1723-1735.
- IEEE Standard Association. (2014). IEEE standard for wearable cuffless blood pressure measuring devices. *IEEE Std*, 1708-2014.
- Jaigirdar, F. T., Rudolph, C., & Bain, C. (2019). Can I trust the data I see? A Physician's concern on medical data in IoT health architectures. *Proceedings of the Australasian*

- computer science week multiconference*. pp.1. pp.10.
<https://doi.org/10.1145/3290688.3290731>
- Khan, Y., Ostfeld, A. E., Lochner, C. M., Pierre, A., & Arias, A. C. (2016). Monitoring of vital signs with flexible and wearable medical devices. *Advanced materials*, 28(22), pp. 4373-pp.4395. <https://doi.org/10.1002/adma.201504366>
- Koydemir, H. C., & Ozcan, A. (2018). Wearable and implantable sensors for biomedical applications. *Annual Review of Analytical Chemistry*, 11, pp.127-pp.146. <https://doi.org/10.1146/annurev-anchem-061417-125956>
- Lee, C. H. (2010). Disaster and mass casualty triage. *AMA Journal of Ethics*, 12(6), pp.466pp.470. <https://doi.org/10.1001/virtualmentor.2010.12.6.cprl1-1006>
- Lee, T. (2021). Periodic Biometric Information Collection Interface Method for Wearable Vulnerable Users. *International journal of advanced smart convergence*, 10(3), pp.33pp.40. <https://doi.org/10.7236/IJASC.2021.10.3.33>
- Lehocki, F., Bogdanova, A. M., Tysler, M., Ondrusova, B., Simjanoska, M., Koteska, B., ... & Macura, M. (2021). SmartPatch for Victims Management in Emergency Telemedicine. In 2021 13th International Conference on Measurement (pp. 146-149). IEEE. <https://doi.org/10.23919/Measurement52780.2021.9446791>
- Loncar-Turukalo, T., Zdravovski, E., da Silva, J. M., Chouvarda, I., & Trajkovic, V. (2019). Literature on wearable technology for connected health: scoping review of research trends, advances, and barriers. *Journal of medical Internet research*, 21(9), e14017. <https://doi.org/10.2196/14017>
- Marlink, R. (2017). Urgently Creating the Better in Global Health. Hastings Center Report, 47(5), pp.25-pp.26. <https://doi.org/10.1002/hast.765>
- McKeon, J. (2021). 61M Fitbit, Apple Users Had Data Exposed in Wearable Device Data Breach. Retrieved from <https://healthitsecurity.com/news/61m-fitbit-apple-users-haddata-exposed-in-wearable-device-data-breach>
- Montgomery K, Chester J. & Kopp K. (2018). Health wearables: ensuring fairness, preventing discrimination, and promoting equity in an emerging IoT environment. *Journal of Information Policy* 8: pp.34–pp.77. <https://doi.org/10.5325/jinfopoli.8.2018.0034>
- Nakikj, D., & Mamykina, L. (2017, February). A park or a highway: Overcoming tensions in designing for socio-emotional and informational needs in online health communities. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (pp. 1304-1319). <https://doi.org/10.1145/2998181.2998339>
- Nash, I. (2021). Cybersecurity in a post-data environment: Considerations on the regulation of code and the role of producer and consumer liability in smart devices. *Computer Law & Security Review*, 40, 105529. <https://doi.org/10.1016/j.clsr.2021.105529>
- Ometov, A., Shubina, V., Klus, L., Skibińska, J., Saafi, S., Pascacio, P., ... & Lohan, E. S. (2021). A survey on wearable technology: History, state-of-the-art and current challenges. *Computer Networks*, 193, 108074. <https://doi.org/10.1016/j.comnet.2021.108074>

- Parimbelli, E., Bottalico, B., Losiouk, E., Tomasi, M., Santosuosso, A., Lanzola, G., ... & Bellazzi, R. (2018). Trusting telemedicine: a discussion on risks, safety, legal implications and liability of involved stakeholders. *International journal of medical informatics*, 112, pp.90-pp.98. <https://doi.org/10.1016/j.ijmedinf.2018.01.012>
- Phaneuf, A. (2022). Latest trends in medical monitoring devices and wearable health technology. Retrieved from <https://www.insiderintelligence.com/insights/wearable-technologyhealthcare-medical-devices/>
- Queiruga-Dios, A., Encinas, A. H., Martín-Vaquero, J., & Encinas, L. H. (2016). Malware propagation models in wireless sensor networks: a review. *International Joint Conference SOCO'16-CISIS'16-ICEUTE'16, Advances in Intelligent Systems and Computing, Vol 527*. Springer, Cham, pp. 648-pp.657. https://doi.org/10.1007/978-3319-47364-2_63
- Santos, L. F., Correia, I. J., Silva, A. S., & Mano, J. F. (2018). Biomaterials for drug delivery patches. *European Journal of Pharmaceutical Sciences*, 118, 49-66. <https://doi.org/10.1016/j.ejps.2018.03.020>
- Schall Jr, M. C., Seseck, R. F., & Cavuoto, L. A. (2018). Barriers to the adoption of wearable sensors in the workplace: A survey of occupational safety and health professionals. *Human factors*, 60(3), pp.351-pp.362. <https://doi.org/10.1177/0018720817753907>
- Schreiner C., Catherwood P., Anderson J. and McLaughlin J. (2010). Blood oxygen level measurement with a chest-based Pulse Oximetry prototype system, *Computing in Cardiology*, 2010, pp. 537-540.
- Shi, M., Wu, H., Zhang, J., Han, M., Meng, B., & Zhang, H. (2017). Self-powered wireless smart patch for healthcare monitoring. *Nano Energy*, 32, 479-487. <https://doi.org/10.1016/j.nanoen.2017.01.008>
- Spence, N., Paul III, D. P., & Coustasse, A. (2017). Ransomware in healthcare facilities: the future is now.
- Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., & Wang, G. (2018). Security and privacy in the medical internet of things: a review. *Security and Communication Networks*, 2018. <https://doi.org/10.1155/2018/5978636>
- Vijayalakshmi, K., Uma, S., Bhuvanya, R., & Suresh, A. (2018). A demand for wearable devices in health care. *Int. J. Eng. Technol*, 7(1-7), pp.1-pp.4.
- Williams, D. R., Lawrence, J. A., Davis, B. A., & Vu, C. (2019). Understanding how discrimination can affect health. *Health services research*, 54, pp.1374-pp.1388. <https://doi.org/10.1111/1475-6773.13222>
- Xue, Y. (2019). A review on intelligent wearables: Uses and risks. *Human Behavior and Emerging Technologies*, 1(4), pp.287-pp.294. <https://doi.org/10.1002/hbe2.173>
- Zdravkova, K. (2017). Who will rule the world in the future?: Can new technologies dramatically change humanity as we know it?. *ORBIT Journal*, 1(1), pp.1-pp.12. <https://doi.org/10.29297/orbit.v1i1.13>

Zhai, P., Ding, Y., Wu, X., Long, J., Zhong, Y., & Li, Y. (2020). The epidemiology, diagnosis and treatment of COVID-19. *International journal of antimicrobial agents*, 55(5), 105955. <https://doi.org/10.1016/j.ijantimicag.2020.105955>

Does Artificial Intelligence Evolve or Degenerate Sports?

Kiyoshi Murata¹[0000-0002-5632-8078], Yohko Orito²[0000-0002-4293-1722] and Yasunori Fukuta¹[0000-0002-2712-5538]

Meiji University, Tokyo, Japan
Ehime University, Matsuyama, Japan

kmurata@meiji.ac.jp

Abstract. This study deals with people’s attitudes towards the use of artificial intelligence (AI) technologies in the field of sports through a questionnaire survey, mainly focusing on the propriety or inappropriateness of AI systems’ substituting for human referees, umpires, or judges. Valid responses, 297 in total, were obtained from a questionnaire survey conducted online from May to July 2021. Statistical analyses of the responses indicate positive attitudes towards such AI use on the whole and that it could be socially acceptable for use in sports, as well as a broader range of fields. A large majority of respondents considered AI technologies would help evolve sports. On the other hand, however, the survey revealed nonnegligible objections and concerns in regard to some aspects of AI use, such as to replace human umpires, referees and judges.

Keywords: Artificial intelligence, sports, judgement, fairness, social acceptability

1 Introduction

This study deals with attitudes towards the use of artificial intelligence (AI) technologies in the field of sports through a questionnaire survey, mainly focusing on the substitution of human referees, umpires, and/or judges. Sports are believed to enrich the lives of people; they are cultural activities that allow people to refresh themselves through accomplishment, fulfilment, community inclusion, fun and joy. Various technologies have helped sports to evolve; as such, today’s sports could be referred to as technology-enhanced human activity. AI systems have been widely used throughout the sport scene for various purposes, including enhancing training performance and improving the quality of game analyses.

In his seminal book published in 1958, Caillois (2001: 9-10) defined “play”, including sports (*agôn*), as an activity with the following essential characteristics:

1. Free: in which playing is not obligatory; if it were, it would at once lose its attractive and joyous quality as diversion;
2. Separate: circumscribed within limits of space and time, defined and fixed in advance;

3. Uncertain: the course of which cannot be determined, nor the results attained beforehand, and some latitude for innovations being left to the player's initiative;
4. Unproductive: creating neither goods, nor wealth, nor new elements of any kind; and, except for the exchange of property among the players, ending in a situation identical to that prevailing at the beginning of the game;
5. Governed by rules: under conventions that suspend ordinary laws, and for the moment establish new legislation, which alone counts;
6. Make-believe: accompanied by a special awareness of a second reality or of a free unreality, as against real life.

However, the professionalisation of many sports and the great commercial success of sports events and businesses appear to have changed the nature of sports, especially as to unproductiveness. In 1930s, Huizinga (1971) already pointed out that sports had been away from the play-sphere. Today, sports provide players and relevant parties, including coaches and sports-related companies, with opportunities for financial success and social prestige.

In any sport, whether it is for professionals or amateurs, ensuring fairness in competition is the main priority. However, it can be difficult to draw a clear line between fairness and unfairness in sports. For example, reflecting the technology-enhanced nature of sports, access to financial, physical (training fields, training equipment, performance meters, etc.) and human (physical training instructors, mental health instructors, skill coaches, etc.) resources often becomes the decisive factor in determining who is a better athlete in a particular sport yet access is by no means equal among competitors. Specifically, the rich tend to have more opportunities to become better athletes. Is this fair?

Another element that affects fair competition is human judgement in sport games.

For example, a Washington Post analysis of home-plate umpires' calls showed that they were missing strike calls at an increasing rate (Greenberg, 2021), indicating that the baseball strike zone is subject to the umpire's whim. We have often witnessed questionable judgements made by referees in, say, professional football games. Many people may feel that the scores of gymnastics or figure skating performances are sometimes evaluated by judges in a partisan way. However, is bad umpiring or judgement part of sport games?

Technologies such as video equipment have been introduced to sports to aid human judges, with advantages such as instant replays: technologies cooperatively work with human judges such as replay officials in baseball, television match officials in rugby football and video assistant referees in association football. However, the use of such technologies is often criticised because it suspends the game, which should otherwise be played without interruption.

Given the tremendous advances in AI technologies, the introduction of AI systems to sports may provide a way to solve such problems. In fact, robot umpires are already being experimentally employed at minor league baseball games (Desmond, 2021). This leads to the following questions. What kind of AI intervention is permissible for sport games? To what extent can or should AI technologies intervene in sports? Under what conditions could AI systems act as a substitute for human referees, umpires, or

judges? Would this change the nature of sports? Should human factors be eliminated from refereeing or judgement in sport games? In short, would AI contribute to or detract from sports? Examining these questions will provide us with an opportunity to deliberate fair, justifiable and socially acceptable AI usage beyond the boundary of sports, e.g., in the ethical and social issues related to AI use in the criminal justice system (Partnership on AI, 2019; Simonite, 2020). Investigating the AI evaluation of artistic impressions of figure skating performers can lead to a re-examination of what art really is. Notably, AI can be found throughout the music industry (Marr, 2019; 2021; Goldstein, 2020).

2 Questionnaire survey

2.1 Overview of the questionnaire survey

The authors developed a questionnaire in Japanese. It was composed of a fact sheet, questions about AI-based systems replacing human umpires or referees and officiating professionals in top-level sport matches, questions regarding AI-based scoring of the performance of top-level players thus displacing human judges, and questions on whether AI would contribute to or detract from sports.

A survey utilising the questionnaire was conducted online using Google Forms from May to July 2021. Of the 300 responses, 297 were deemed valid by the authors. The age distribution and sport experiences of respondents are shown in Table 1. The sport experiences seemed to reflect the fact that many ordinary Japanese people participate in sports activities in earnest through school sporting clubs as part of their education, whereas local sporting clubs contributes to enhancing people's sport experiences from their childhood. For the sake of statistical analysis, level of experience was divided into two categories: advanced level (AL) and beginner or intermediate level (BIL). There were no statistically significant differences in

Table 1. Attributes of respondents

	Sport experiences					
Age (years)	Advanced level		Beginner or intermediate level		No experience	Total
	Professional or top athlete	Joined nationwide sporting contests	Played at school sporting clubs	Enjoyed sports as a hobby		
10s	2	17	20	1	2	42
20s	6	56	63	9	6	140
30s	0	8	7	2	0	17
40s	1	7	16	3	1	28
50s +	4	22	30	13	1	70
Total	13	110	136	28	10	297
	123		164			

experience level among age groups ($\chi^2(4) = 3.795$, $p = 0.434$). Only responses from respondents with sport experience were included in the statistical analyses.

Sports that had been played by more than 5% of respondents (multiple answers allowed) were as follows, in descending order: tennis (90/287), swimming (76), kendo (Japanese swordsmanship: 73), baseball (39), association football (31), basketball (29), track and field (23), rugby football (21), gymnastics (15), table tennis (15) and volleyball (15). Among the respondents, 89 had experienced one or more budō (武道) or Japanese martial arts (including aikido, judo, karate, kendo, kyudo, naginata, shorinji kempo and sumo, in this study), and 165 had played one or more sports in which human judgement or scoring was technically difficult and a decisive factor in evaluating the performance of players (hereafter, human-factor-sensitive [HFS] sports, which include association football, baseball, baton twirling, competitive dance, figure skating, gymnastics, judo, karate, kendo, rhythmic gymnastics, rugby football and softball, in this study).

Sports that more than 5% of respondents were interested in and often watched at venues or on TV (multiple answers allowed) were as follows, in descending order: baseball (137/287), association football (101), volleyball (71), tennis (66), kendo (65), rugby football (61), figure skating (55), basketball (43), table tennis (37), swimming (33), judo (32), track and field (31), badminton (27) and gymnastics (20). The two most popular sports in Japan are baseball and association football, for which domestic professional leagues are organised and operated.

Approximately 40% of respondents (41.8%; 120/287) knew that AI technologies were used in sports for some purpose such as tactical analysis, effective training of athletes, and assisting referees or judges. Only for respondents in their 40s, the number of those who knew such AI use exceeded one of those who did not. However, the results of a chi-square test demonstrate that age groups of respondents were independent from knowledge on AI use in sports ($\chi^2(4) = 4.313$, $p = 0.365$). Similar tendencies are found in terms of respondents' various aspects: there were no statistically significant differences in knowledge on AI use in sports between those who had AL experiences and those who had BIL experiences ($\chi^2(1) = 0.011$, $p = 1.000$), between those who had played budō and those who had not ($\chi^2(1) = 2.242$, $p = 0.155$), and between those who had played HFS sports and those who had not ($\chi^2(1) = 3.760$, $p = 0.054$).

2.2 Survey results and analysis

Replacement of human umpires or referees with AI systems

Overall, nearly 70% of respondents (68.3%; 196/287) agreed with using AI systems to substitute for human umpires or referees in professional or top-level matches in sports such as baseball or association football. Among those who played kendo, 54.8% (40/73) agreed. Such respondents tended to emphasise the objectivity of judgements (16/40) rather than their fairness (4/40). On the other hand, just a small number of respondents who had participated in that martial way (6/33) placed importance on gamesmanship with umpires or referees. These results may reflect the

culture of kendo: it is strictly prohibited to show dissatisfaction with referees during games, even considering that making a correct judgement is technically very difficult.

As shown in Table 2, the majority of respondents in each age group agreed to the substitution of AI systems for human umpires or referees at professional or top-level sport matches. However, the results of a chi-square test reveal that the pros and cons of the substitution statistically significantly differed among age groups at the 5% level ($\chi^2(4) = 10.983$, $p = 0.027$). A residual analysis indicates respondents in their 20s were significantly more likely to agree than others (the absolute value of the adjusted standardised residual in the row of “20s” exceeds 1.96).

Table 2. Replacement of human umpires or referees with AI systems (Age groups)

	Do you agree that AI systems, not humans, act as umpires or referees in professional or top-level sports matches?		
Age (years)	Agree	Disagree	Total
	Adjusted std. residual	Adjusted std. residual	
10s	28	12	40
	.3	-.3	
20s	103	31	134
	2.9	-2.9	
30s	9	8	17
	-1.4	1.4	
40s	15	12	27
	-1.5	1.5	
50s +	41	28	69
	-1.8	1.8	
Total	196	91	287

The same was true among respondents with AL as well as BIL experience in sports (Table 3), and the level of experience did not affect attitudes ($\chi^2(1) = 0.263$, $p = 0.611$).

Table 3. Replacement of human umpires or referees with AI systems (AL vs. BIL)

	Do you agree that AI systems, not humans, act as umpires or referees in professional or top-level sports matches?		
Sport experience level	Agree	Disagree	Total
AL	82	41	123
BIL	114	50	164
Total	196	91	287

As shown in Table 4, most respondents with budō experience (56.2%; 50/89) agreed with AI playing a role, however, far fewer of those respondents felt that way compared to respondents with no experience in budō (73.7%; 146/198). The results of a chi-square test demonstrate that those respondents with budō experience were less likely to agree with the replacement of human umpires or referees with AI-based systems statistically significantly at the 1% level ($\chi^2(1) = 8.741$, $p = 0.004$). The value of a phi coefficient ($\phi = -0.175$, $p = 0.003$) supports this. These findings suggest the usefulness of follow-up interviews with those who are well-versed in budō for this study.

Table 4. Replacement of human umpires or referees with AI systems (Respondents with budō experiences vs. others)

	Do you agree that AI systems, not humans, act as umpires or referees in professional or top-level sports matches?		
Budō experience	Agree	Disagree	Total
Yes	50	39	89
No	146	52	198
Total	196	91	287

Similar findings are found for those respondents who had played HFS sports (Table 5) and for those who were interested in HFS sports and often watched games of the sports at venues or on TV (Table 6). The former tended to be statistically significantly less likely to accept the replacement compared with those who had not

played such sports at the 0.1% level ($\chi^2(1) = 14.195$, $p < 0.001$; $\phi = -0.222$, $p < 0.001$), and the latter did compared with those who were not interested in those sports

Table 5. Replacement of human umpires or referees with AI systems (Respondents with HFS sports experiences vs. others)

	Do you agree that AI systems, not humans, act as umpires or referees in professional or top-level sports matches?		
HFS sports experience	Agree	Disagree	Total
Yes	98	67	165
No	98	24	122
Total	196	91	287

Table 6. Replacement of human umpires or referees with AI systems (Respondents who were interested in HFS sports vs. others)

	Do you agree that AI systems, not humans, act as umpire or referee at professional or top-level sport matches?		
Interest in HFS sports	Agree	Disagree	Total
Yes	147	78	225
No	49	13	62
Total	196	91	287

at the 5% level ($\chi^2(1) = 4.212$, $p = 0.045$; $\phi = -0.121$, $p \neq 0.040$).

Figures 1 and 2 summarise the reasons for agreeing and disagreeing with the idea of using AI systems to substitute for human umpires/referees, respectively. Those who agree tend to emphasise the correctness and fairness of judgements during a sport game, and those who do not agree tend to worry about changing the nature of sports and losing important characteristics.

Further investigations on the nature of sports need to be carried out. We should examine, for example, the dramatic aspects of sports, how dramatic scenes are brought about during a sport game, whether the existence of a human umpire or

referee is an essential element of sports, and whether players' gamesmanship with umpires or referees is an indispensable element.

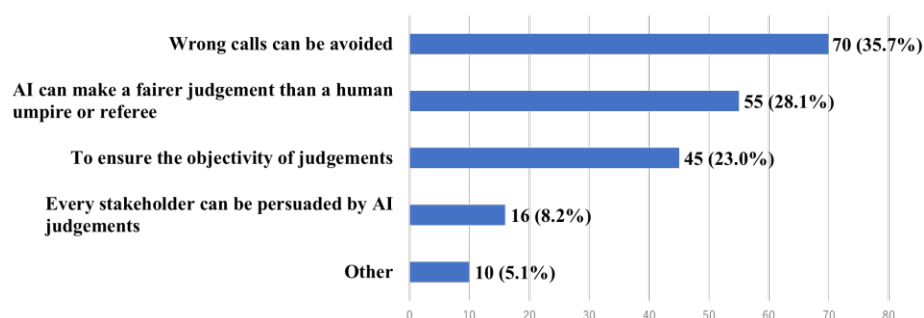


Fig. 1. Reasons for agreeing with AI-based umpires/referees (n=196)

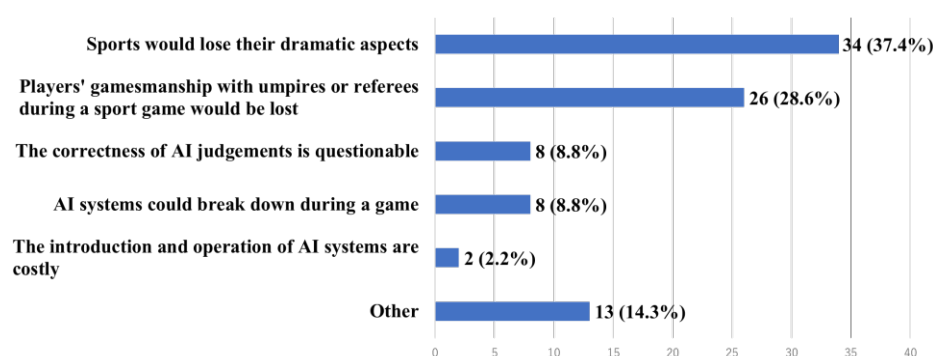


Fig. 2. Reasons for disagreeing with AI-based umpires/referees (n=91)

Slightly more than half of respondents (52.6%; 151/287) considered it acceptable for an AI umpire or referee to eject a player (Table 7). Only among respondents in their 50s was the majority opposed to this. However, the results of a chi-square test show that there was no statistically significant difference among age groups in this regard ($\chi^2(4) = 3.313$, $p = 0.507$).

Table 7. Ejection by an AI-based umpire or referee (Age groups)

	Is it acceptable for an AI umpire or referee to eject a player during a professional or other toplevel sport match?		
Age	Acceptable	Unacceptable	Total

(years)	Adjusted std. residual	Adjusted std. residual	
10s	22	18	40
	.3	-.3	
20s	76	58	134
	1.3	-1.3	
30s	9	8	17
	.0	.0	
40s	14	13	27
	-.1	.1	
50s +	30	39	69
	-1.7	1.7	
Total	151	136	287

There is a statistically significant difference in opinions on this topic between those who had played HFS sports and those who had not (Table 8), where the majority of those in the former group found it unacceptable and the majority of those in the latter group thought the opposite. The outcomes of a chi-square test ($\chi^2(1) = 6.685$, $p = 0.012$) indicate respondents with HFS sport experiences were significantly more likely to dispute the ejection at the 5% level. The value of a phi coefficient ($\phi = 0.153$, $p \neq 0.010$) supports this.

There was a similar difference between respondents who were interested in and often watched such sports and those who were/did not (Table 9) at the 5% significant level ($\chi^2(1) = 4.494$, $p = 0.044$; $\phi = 0.125$, $p \neq 0.034$).

Table 8. Ejection by an AI-based umpire or referee (Respondents with HFS sports experiences vs. others)

	Is it acceptable for an AI umpire or referee to eject a player during a professional or other top-level sport match?		
HFS sports experience	Acceptable	Unacceptable	Total
Yes	76	89	165
No	75	47	122
Total	151	136	287

Table 9. Ejection by an AI-based umpire or referee (Respondents who were interested in HFS sports vs. others)

	Is it acceptable for an AI umpire or referee to eject a player during a professional or other top-level sport match?		
Interest in HFS sports	Acceptable	Unacceptable	Total
Yes	111	114	225
No	40	22	62
Total	151	136	287

The reasons for agreeing (Fig. 3) and disagreeing (Fig. 4) with this idea seem to reveal that AI systems are expected to make an objective, fair and impartial decision, but that the inability of AI systems to read human emotions is considered a drawback. This is interesting, considering that a human umpire or referee is not necessarily required to read a player's emotions: they don't necessarily need to judge whether the act triggering ejection was intentional or not, and also need not consider whether the player is shocked at being ejected.

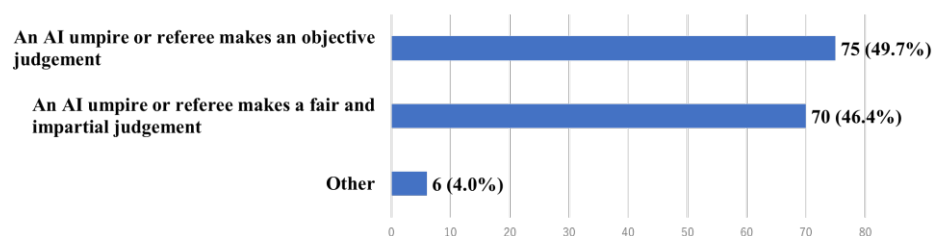


Fig. 3. Reasons for agreeing with an AI-based umpire's or referee's ejection of a human player (n=151)

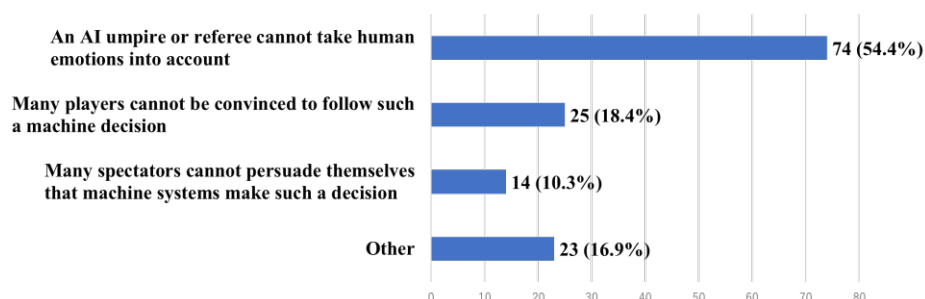


Fig. 4. Reasons for disagreeing with an AI-based umpire's or referee's ejection of a human player (n=136)

Table 10. Replacement of human judges with AI systems (Age groups)

	Do you agree that AI systems, instead of human judges, score the performances of toplevel players in gymnastics, figure skating and other sports?		
Age (years)	Agree	Disagree	Total
	Adjusted std. residual	Adjusted std. residual	
10s	15	25	40
	-2.8	2.8	
20s	77	57	134
	.0	.0	
30s	11	6	17
	.6	-.6	
40s	20	7	27
	1.8	-1.8	
50s +	42	27	69
	.7	-.7	
Total	165	122	287

Scoring the performance of top-level players by AI systems

Nearly 60% of respondents (57.5%; 165/287) agreed that AI systems replaced humans for scoring the performance of top-level players in sports, such as gymnastics and figure skating. As shown in Table 10, only among teenage respondents did the majority disagree with that idea. Through a chi-square test, it is demonstrated that there were statistically significant differences in respondents' attitudes towards the replacement among age groups at the 5% level ($\chi^2(4) = 10.264$, $p = 0.036$). In

particular, teenage respondents significantly tended to oppose to it at the 5% level, according to the outcomes of a residual analysis.

Contrary to expectations, among respondents who had played HFS sports, the majority considered such replacement acceptable, whereas those who had not played such sports were evenly split over it (Table 11). The results of statistical tests demonstrate the significant difference in the attitudes towards the replacement between the two respondent groups at the 5% level ($\chi^2(1) = 4.873$, $p = 0.03$; $\phi = 0.130$, $p = 0.027$). These, in addition to the findings in this study described above, suggest the usefulness of follow-up interviews with those who have played HFS sports for this study. On the other hand, between those respondents who were interested in HFS sports and often watched games of them at venues or on TV and other respondents, there was no statistically significant difference in this regard ($\chi^2(1) = 0.035$, $p = 0.885$).

Table 11. Replacement of human judges with AI systems (Respondents with HFS sports experiences vs. others)

	Do you agree that AI systems, instead of human judges, score the performances of top-level players in gymnastics, figure skating and other sports?		
HFS sports experience	Acceptable	Unacceptable	Total
Yes	104	61	165
No	61	61	122
Total	165	122	287

While those who agreed with this idea tended to place importance on fairness and accuracy in scoring (Fig. 5), it is impressive that 82.0% of respondents who disagreed (100/122) felt that it was impossible for AI systems to quantify the beauty and resonance of a performance (Fig. 6). We may need to reconsider what beauty and resonance in sports are, and the meaning of a human's, rather than a machine's, scoring them at sport competitions.

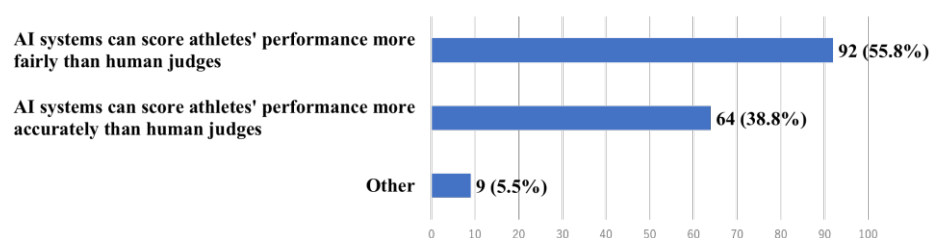


Fig. 5. Reasons for agreeing with the replacement of human judges with AI systems (n=165)

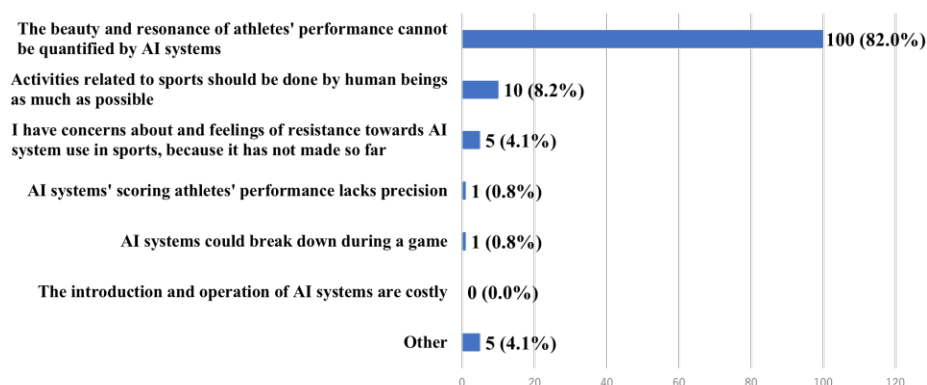


Fig. 6. Reasons for disagreeing with the replacement of human judges with AI systems (n=122)

Table 12. The evolution of sports owing to AI technologies (Age groups)

	Do you think sports will evolve through the application of AI technologies?		
Age (years)	Yes, I think so	No, I don't think so	Total
	Adjusted std. residual	Adjusted std. residual	
10s	35	5	40
	.8	-.8	
20s	113	21	134
	.6	-.6	
30s	14	3	17
	-.1	.1	
40s	24	3	27
	.9	-.9	
50s +	52	17	69
	-1.9	1.9	
Total	238	49	287

AI-driven evolution or degeneration of sports

More than 80% of respondents (82.9%; 238/287) felt that AI technologies could contribute to the evolution of sports (Table 12). The outcomes of a chi-square test indicate that there was no statistically significant difference in this feeling among age groups ($\chi^2(4) = 4.247$, $p = 0.374$).

The responses to this concept did not vary by respondent attribute, except in regard to HFS sports. Those interested in (and who watched) HFS sports were statistically

significantly less likely to endorse this idea than those who were not interested in HFS sports at the 5% level (Table 13; $\chi^2(1) = 4.533$, $p = 0.036$; $\phi = -0.126$, $p = 0.033$).

Table 13. The evolution of sports owing to AI technologies (Respondents who were interested in HFS sports vs. others)

	Do you think sports will evolve through the application of AI technologies?		
Interest in HFS sports	Yes, I think so	No, I don't think so	Total
Yes	181	44	225
No	57	5	62
Total	238	49	287

The reasons for agreeing (Fig. 7) or disagreeing (Fig. 8) with this concept suggest the need for further investigations of various issues, including the nature of sports. This is because it is key to establish socially acceptable ways of using AI technologies in sports while also promoting the healthy development of this indispensable component of human culture.

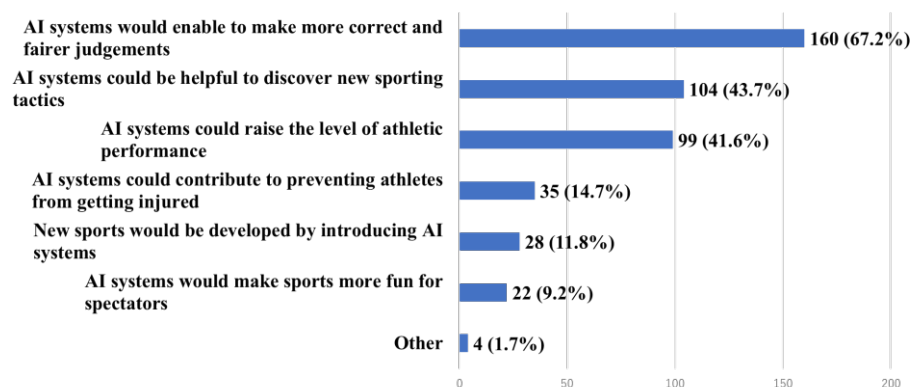


Fig. 7. Reasons for agreeing with the evolution of sports owing to AI technologies (n=238; multiple answers allowed)

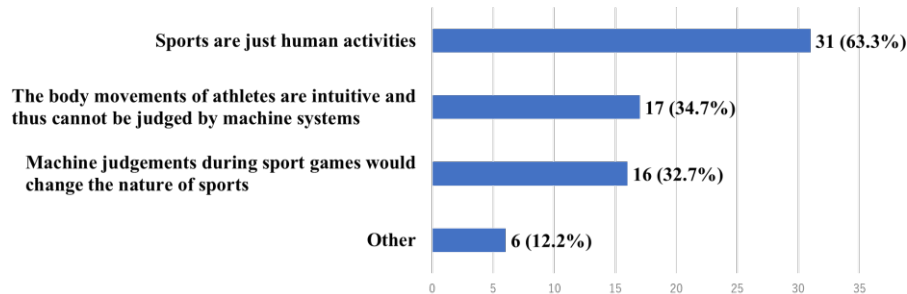


Fig. 8. Reasons for disagreeing with the evolution of sports owing to AI technologies (n=49; multiple answers allowed)

3 Concluding remarks

Overall, respondents had positive attitudes towards the use of AI technology in sports. Most felt that it would help evolve sports. On the other hand, however, the survey revealed nonnegligible objections and concerns in regard to some aspects of AI use, such as to replace human umpires, referees and judges. Given that sports are an important part of human culture, it is critical to achieve social acceptability in AI use in sports for the sound development of human culture and society. Above all, the nature of sports should be examined. Follow-up interviews with knowledgeable people may be helpful in this regard. In future work, the authors plan to conduct interviews with an expert in kendo and a world-class figure skater to this end.

Acknowledgments

This work was supported by the JSPS Grand-in-Aid for Scientific Research (C) 20K01920 and (C) 22K02063.

References

- Caillois, R. (2001). *Man, play and games*. Urbana and Chicago, IL: University of Illinois Press (translated by Barash, M.).
- Desmond, J. P. (2021, September 23). Robot umpires invade baseball; AI that makes mistakes on purpose could help. Retrieved from <https://www.aitrends.com/robotics/robotumpires-invade-baseball-ai-that-makes-mistakes-on-purpose-could-help/>.
- Goldstein, L. (2020, June 3). *How is AI transforming the music industry?* Retrieved from <https://medium.com/@liangoldstein/how-is-ai-transforming-the-music-industry5b46087eb589>.

- Greenberg, N. (2021, August 16). *MLB umpires are squeezing the strike zone, and it's hurting some teams more than others*. Retrieved from <https://www.washingtonpost.com/sports/2021/08/16/worst-strike-zones-data/>.
- Huizinga, J. (1971). *Homo ludens: a study of the play-element in culture*. Boston, MA: Beacon Press (translated by Hull, R. F. C.).
- Marr, B. (2019, July 5). *The amazing ways artificial intelligence is transforming the music industry*. Retrieved from <https://www.forbes.com/sites/bernardmarr/2019/07/05/theamazing-ways-artificial-intelligence-is-transforming-the-musicindustry/?sh=3ae6525f5072>.
- Marr, B. (2021, May 14). *How artificial intelligence (AI) is helping musicians unlock their creativity*. Retrieved from <https://www.forbes.com/sites/bernardmarr/2021/05/14/how-artificial-intelligence-aiis-helping-musicians-unlock-their-creativity/?sh=7b8f1c847004>.
- Partnership on AI (2019, April 23). *Report on algorithmic risk assessment tools in the U.S. criminal justice system*. Retrieved from <https://partnershiponai.org/paper/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/2/>.
- Simonite, T. (2020, February 19). *Algorithms were supposed to fix the bail system. They haven't*. Retrieved from <https://www.wired.com/story/algorithms-supposed-fix-bail-system-they-havent/>.

The social implications of brain machine interfaces for people with disabilities: Experimental and semistructured interview surveys

Yohko Orito ¹[0000-0002-4293-1722], Tomonori Yamamoto ¹[0000-0002-9600-3876], Hidenobu Sai ¹[0000-0002-6010-4747], Kiyoshi Murata ²[0000-0002-5632-8078], Yasunori Fukuta ²[0000-0002-2712-5538], Taichi Isobe ³[0000-0002-2477-0830] and Masahi Hori ⁴[0000-0002-0651-1544]

¹Ehime University, Matsuyama, Japan

²Meiji University, Tokyo, Japan

³Health Sciences University of Hokkaido, Tobetsu, Japan

⁴Waseda University, Tokyo, Japan

orito.yohko.mm@ehime-u.ac.jp

Abstract. A brain-machine interface (BMI) is a cyborg technology that enables the communication between the human brain and an external device by interpreting brain signals. This cyborg technology is expected to develop further to be utilised for supporting people with disabilities, in particular, those who cannot move their limbs by themselves. However, the social implications of BMI, even the negative impacts or social influences on people with disabilities who are expected to use this interface, have not been discussed thoroughly.

In this study, we invited two people with disabilities to participate in our experiment with a non-invasive BMI device (EEG input device). The device was placed on their heads so that the robotic arm can be operated remotely by their brain signals. Semi-structured interviews were conducted with them before, during, and after the experiment to assess their reactions and views to the BMI device and their ethical awareness of a BMI device and to estimate the benefits and risks of using such devices for people with disabilities.

The result of this experimental survey provided some meaningful insights: the importance of understanding the differences in the perception of people with and without congenital disabilities; the effect of such technologies on the individual identities and the social relationships of people with disabilities; risks caused by the gaps between the expectations from such technologies and their actual use; and social importance of communication among engineers, policymakers, researchers and people with disabilities in order to develop and utilise cyborg technologies to assist people with disabilities

Keywords: brain machine interface, people with disabilities, cyborgisation, negative social impacts

1 Introduction

A brain machine interface (BMI), or a brain computer interface (BCI), enables the communication between the human brain and an external device and vice versa using dedicated hardware and software and brain signalling (Orito et al., 2020a). Many researchers and practitioners from various fields such as medicine, marketing and gaming have identified and actively discussed the positive impacts and potential utilities of this interface. In terms of its social application, in particular, this cyborg technology is expected to support people with disabilities, such as amyotrophic lateral sclerosis (ALS), muscular dystrophy, and cerebral palsy. The utilisation of a BMI in medical facilities and rehabilitation hospitals is gaining attention in society.

However, at present, BMI devices are relatively expensive and not widely used by people with disabilities in society. Therefore, it may be difficult to imagine the risks or negative social impacts of the devices. Although some studies have examined the ethical issues related to the usage and social applications of BMI (e.g., Gilbert et al., 2019; Wahlstrom, 2018; Gröbler & Hildt, 2014; Fukushi & Sakura, 2007), the social implications of BMI usage for people with disabilities have not been thoroughly discussed in the literature. Thus, the ethical aspects related to the application of BMI should be examined in a proactive manner in order to avoid serious risks or social harms of this application for people with disabilities.

Therefore, we conducted experiments using BMI systems and semi-structured interview surveys to investigate the ethical and social issues related to BMI usage based on the findings of our previous studies (Murata et al., 2019, 2018, 2017; Isobe, 2013). In our previous studies, we involved people without disabilities (students, professors in social security law, and professional social workers) as the participants (Orito et al., 2021a, 2021b, 2020a, 2020b). In our previous experiments, we asked the participants to place a non-invasive BMI device (EEG input device) on their heads to operate a robotic arm remotely or without touching it, as if through psychokinesis. We conducted interview surveys with the participants before, during and after the experiments to assess their reactions to the BMI device, feelings about the operation of a robotic arm and ethical awareness of brain signal collection using a BMI device and to estimate the benefits and risks of utilising such devices. In this study, however, we invited two people with disabilities as participants of the experiment and our semistructured interview survey.

2 Overview of the experiment and the semi-structured interview survey

2.1 An experiment environment

In the experiment environment of this study, the participants controlled a robotic arm using their brain signals via a non-invasive wearable BMI/EEG (electroencephalography) device placed on their heads. At the training stage of the experiment, the participants' brain signals were collected by the system through

BMI/EEG devices. At that time, a ‘relaxed state’ and an ‘in-operation state’ (the participants were asked to imagine that they were pushing the small box shown on the computer screen) of the brain signal data were recorded by the systems, and the data were transmitted to and stored in the application software that enables the manipulation of the robotic arm.

After the training stage, each participant was asked to move to the in-operation state by imagining that they were pushing the robotic arm backward. If their brain signal patterns matched with the in-operation state registered in the software system, the robotic arm turned toward the back (for details on the procedures of the experiment and survey, see Orito et al., 2021a; 2020a; 2020b).

2.2 Survey participants and questions

This experiment that used non-invasive EEG/BMI devices and the semi-structured interviews were conducted between September and November 2021 at Ehime University in Matsuyama, Ehime Prefecture, Japan. All procedures performed for this experiment and survey were in accordance with the ethical standards of the Research Ethics Committee at the Faculty of Collaborative Regional Innovation, Ehime University. The attributes of the two participants (one with impaired hearing and the other with physical paralysis caused by cervical spinal cord injury) are listed in Table 1. The interview survey was conducted in Japanese and the original transcripts were translated into English for the purpose of this study and the responses of this survey described here were confirmed by the interviewees.

Table 1. The participants of the survey (n = 2)

Subject ID	Age	Gender	Types of disabilities	Have you ever heard about a BMI device or are you familiar with a BMI device?	Expectation/anxiety about the experiment (Weak 0 – Strong 7)
1	30s	Male	Deaf (uses hearing aids and cochlear implants)	Yes, I knew about it because of my work that supports people with disabilities.	1/5
2	40s	Male	Physical paralysis (uses a wheelchair)	No, I have never seen a BMI device before this. I have come across some information on social media about disability and I know that it exists, but this is basically my first time.	1/7

As an important consideration in understanding the contribution of this study, the two participants have a common attribute—both of them are committed to supporting people with disabilities in their jobs. Subject 1 is a deaf person who has gained

advanced teaching and research experience in assisting people with disabilities at the undergraduate and graduate levels and currently assists students with reasonable accommodation needs at a higher education institution. Subject 2 became disabled during his university years after a traffic accident. Later, he established an organisation that aimed to support the independent living of people with disabilities. Thus, both of them play central roles in helping people with disabilities. Before the survey, we confirmed the health conditions of the participants and ensured coordination between the experimental facility and the environment, especially for Subject 2 who uses a wheelchair.

The interview questions were categorised as follows: (a) privacy and personal data protection; (b) human autonomy and dignity; (c) identity development and personal transformation; (d) the acceptance of a body extension in an individual and organisational context; (e) workplace cyborgisation; and (f) social responsibility and informed consent, same as in our previous studies (Orito et al., 2021a; 2021b; 2020a; 2020b). However, this was our first survey of people with disabilities, so we added some questions based on their disabilities and circumstances.

3 Results of the experiment and the semi-structured interview survey

The two participants successfully controlled the robotic arm using their brain signals via the BMI device. They suggested that the device can be used for people with disabilities or conditions such as ALS, muscular dystrophy, cerebral palsy and cervical spine injuries, and noted that the BMI device and systems have a great potential to assist people with disabilities in navigating through daily life and work activities, expressing their wishes and communicating with others, and realising selfactualisation. Regarding the privacy and personal data (EEG data) protection issues, the two participants stated that they would be willing to provide their brain signal data depending on the purpose of use. However, they were somewhat afraid that when the BMI technology would be further developed and adapted for social applications, the data could be utilised and analysed in more complicated and different means compared to the current usage. These points are almost similar to those noted by the social welfare professionals of the participants invited in previous studies (Orito et al., 2021a; 2021b).

By contrast, the people with disabilities, who participated in this study, offered unique opinions regarding the use of BMI and its ethical concerns for people with disabilities. The following subsections describe some of the particularly insightful views and comments expressed as responses to the survey questions.

3.1 Differences between people with and without disabilities / between congenital and acquired disabilities

Subject 1 described the possible influence of EEG and imaginary sense differences between people with and without disabilities on the experiment's results, indicating people without limbs.

“Perhaps, the [BMI] system was developed assuming it works with the brain signals of a person without disability. So, I feel that it is impossible to know beforehand whether the robot arm moves as expected or how it reacts to brain signals when a person without limbs operates the system. In addition, this experiment system uses a visual image, when collecting a participant's brain signals, to control the robot arm. But, I'm afraid the way of seeing the visual image may differ from person to person.” (Subject 1)

Moreover, both of the participants indicated that the results of this kind of experimental survey using BMI may depend on whether participants have congenital or acquired disabilities. For example, Subject 1 suggested that the results may vary depending on whether they are habilitated or rehabilitated (reverting back to an original state, as indicated by the prefix “re-”). They expressed doubts regarding participants of the experiment being asked to imagine pushing a box. They pointed out that if participants did not have an experience of pushing or could not visualise pushing something by their hands or any other body part, the results of the experimental survey may differ.

“If a participant [an individual with a disability] has no sense of pushing, they may show a different way of pushing a box than our image of pushing it hard with our hands. For example, they may try to push keeping their body down. Maybe, their image of pushing differs from a person without disabilities has. I feel like this.” (Subject 1)

“If an individual was born with a disability, it may be difficult for them to imagine body movement because they have never experienced it. For example, there are people who cannot visualise themselves moving. If a person who has never walked or held an object before sees another person holding an object, that person would feel completely differently from this person. I wonder if such a person [who has never walked or held an object before] would be able to imagine moving things. Of course, this is a case of a person with very severe disabilities, though.

In my own organisation, some people were born with disabilities. Through communicating with them, I have often realised that I don't understand at all what they cannot understand or recognise.” (Subject 2)

Furthermore, Subject 1 noted that for people who have always lived in a soundless world and use sign language as their first language, there is a possibility that they might have unique senses, modalities and community.

“It seems that in some cases where people cannot hear, certain behaviours and modalities were exhibited by them. In such cases, while they often lived on the premise that they cannot hear each other, no serious problems were caused in their community.

For example, there are some cities or villages where many deaf people live, such that there seem to be no communication problems between those who could hear and those who could not. In such places, people who can hear speak sign language. In such places, [hearing-impaired] people can live without [cochlear] implants or supporting devices because many of them cannot hear, so they do not need such devices. However, this is not the case in an urbanised area. Thus, they may still need cochlear implants or cyborg devices, and the level of their desire changes depending on the various ways and situations.

People using only sign language may feel that listening to sounds is being forced to put a foreign object into their bodies. In such a case, they might wonder why they have to go through such a trouble even though there is nothing wrong with them.” (Subject 1)

This statement implies that the implementation of cyborg technology, not only BMI, into the daily lives or social activities of people with congenital disabilities can transform, or force to change, the physical and communication sensitivities and perceptions of them, even if they are not consciously aware of it.

Subject 1 also told the following about whether cyborg technology allows people with disabilities to have more physical power and/or intellectual abilities than people without disabilities.

“I believe that using such a cyborg device, people with disabilities should not be allowed to do more than people without disabilities. I don’t know whether ethics on this idea have been established or not, but when such [cyborg or emergent] technologies are developed, I think that ethical guidelines should restrict people with disabilities to be able to do more than people without disabilities.” (Subject 1)

“As with the concept of ‘reasonable accommodation’, it is our job now to support people with disabilities in having the same level of capacity as people without disabilities. We cannot go beyond that. I don’t think that [anyone exceeding standard physical and mental capabilities] should be allowed; no matter who that is. So, unless a proper boundary is established, there is a risk of reverse discrimination.” (Subject 1)

This point will also be important for future considerations regarding the development and application of cyborg technologies for people with and without disabilities.

3.2 Influences on the self-identity and relationship development of people with disabilities

Subject 2 became disabled in his adulthood after a serious traffic accident, which resulted in physical paralysis. He expressed that for some time after the accident, he spent his days holed up in his own home, hesitating to engage in active communication with anyone, except for his family. However, now, he plays a leading role in his organisation that supports the independent living of people with disabilities. Sometimes, he gives his opinions and lectures on their activities and the significance of independent living for people with disabilities in public or academic events. He remarked that, for him, being a person with a disability is now his identity and that if technology eliminates the characteristics of disabilities, he will lose a part of his identity.

“I have been [living with a disability] for 20 years, and I have become not to hate this body due to my disabilities and inability to do various things such as moving on my own terms. I will profess this if necessary. I’m able to engage in satisfactory work thanks to my disabled body, and I do not want my disability to disappear because it is an integral part of my identity.

Thus, if some robots or cyborg devices can do everything [for people with disabilities], what I can talk in my lectures will decrease by half. I would even lose my job! Those who will lose their jobs [by the usage of robots, BMI, and other technologies] are people like me, not caregivers.” (Subject 2)

Besides speaking on self-recognition and self-identity, the two participants also highlighted the issues pertaining to the relationship development between people with disabilities and their caregivers and other people. In particular, they pointed out that some people with disabilities have become so accustomed to being helped or assisted that they are unable to express their gratitude to those who help or assist them and that this makes it difficult for them to form favourable relationships with others.

“Even if the actual cyborg technologies such as BMI become very convenient for some people who have disabilities, or even if we think that they should buy and use it, they may not wish to use such technologies. For example, one person with a disability may believe he/she does not need it, because he/she is already looked after by the caregiver and leave almost every necessary thing to the caregivers. However, there are also people who do not want to do any tasks themselves, because it is taken for granted

that someone else will help them; some people are also satisfied with their current lives.

As a task of a person in my position, the first thing to do, when some people with disabilities come in [educational facilities], is to teach them to ask for things for themselves, or ask others what they want in their 'own words'. This is the principle behind reasonable accommodation.” (Subject 1)

“Although they don’t have any bad intentions, they have become accustomed to asking people for help in their daily lives; however, they may not be able to say ‘Thank you’ for something someone has done for them. For example, when people with disabilities ask someone to do something that is harder than usual things, they really don’t know how to ask that of someone, and they can’t express their appreciation. But if they don’t thank people, they will lose opportunities to do something together with others more and more.” (Subject 2)

If some people with disabilities have problems in communicating with other people or their caregivers, the application of cyborg technology may aggravate this problem. Is it conceivable that the promotion of cyborg technology usage may make it more difficult for such people to ask others for help and further compromise their ability to build favourable relationships with others? Or is it that such communications between people with disabilities and their caregivers are unnecessary in the future in which cyborg technologies are utilised ubiquitously? Conversely, of course, the application of those technologies could improve their relationships with caregivers.

3.3 Risks associated with malfunctions and uncontrollability of BMI systems

Subject 1 who uses hearing aids and cochlear implants presented the following concerns about the responsibilities and problems arising from operational errors in using BMI devices.

“For example, it is possible that the robotic arm could accidentally go off and suddenly touch someone and injure him or her due to an unintended or different action from what the user intended. I’m a bit concerned about such undetermined and unimaginable situations occurring beyond the scope of my consideration.” (Subject 1)

Subject 2, who operated a wheelchair by himself in his daily life, expressed the following suspicions about malfunctions of BMI devices.

“I thought it would be great to be able to move the wheelchair using my brain signals, but meanwhile, I would think a lot of things while controlling the wheelchair movements. Then, what would happen? When I move my

wheelchair, I don't always think of moving it forward. I wonder whether I can stop it if I thought of something else, say my unfinished job, while I was on it. Is it dangerous? When I'm moving in a wheelchair along the pavement and feel hungry, can I control it not to hit a guardrail?" (Subject 2)

These concerns about malfunctions or uncontrollability of BMI systems have been also pointed out by people without disabilities and social welfare professionals who participated in our previous studies. However, if more opportunities to conduct surveys of people with disabilities who use wheelchairs, assistive technologies or, in future years, cyborg devices arise, these participants may highlight more practical concerns and often overlooked precautions regarding how to operate them in more specific situations.

3.4 Economic disparities in accessing cyborg devices

Subject 2 expressed his concern about the disparity or inequality between the haves and have-nots in terms of access to cyborg devices and its influences on society and the workplace.

"Everyone with a disability may want to use a wheelchair or a support machine, but it is financially impossible for some of them. People who have a lot of money from workers' compensation or accident insurance can afford to use those devices and have more options; however, if they do not, they need to finance the cost on their own. For example, people who live on public funds or welfare benefits, or those who live off their savings, need to make ends meet to use a wheelchair. Therefore, those who are willing to live within the social security systems will consider how much they will be compensated for to access those devices." (Subject 2)

"I think it's no problem that people using various cyborg devices join an organisation and consequently it becomes more diverse, although this may make organisational management more complicated. However, I'm still afraid I'll be very embarrassed if such diversification leads to the idea that using such technologies is good and those who use the technologies are people of ability. This implies that money can buy any human ability and one who cannot afford to buy the technologies is a person with an inferior." (Subject 2)

As a practical matter, it would be necessary to consider the issue of how to provide fairness in access and opportunities for cyborg technologies for people with and without disabilities.

3.5 Gaps between expectation and the reality regarding the use of cyborg technology

Even if the cyborg technology is to some extent accessible to a wide range of people with disabilities, there are several concerns about the gaps between expectation and the reality of using cyborg devices. Subject 1 who used a cochlear implant represented himself as already a cyborgised individual; however, he stated that even when cochlear implants are surgically implanted, sometimes, there is a gap between actual hearing and expectation among users. In this regard, he highlighted some concerns.

“I think there are many actual cases in which a great expectation for a chip [cyborg device] patients have before it is implanted turns to their deep despair because the reality they know after surgery is quite different [from the expectation]. I know some people’s expectations were so high that they actually ended up not using the [implanted] cyborg devices. When people with disabilities have a high expectation of safety and needs, they decide to use the devices or programs, as they believe that the devices will fulfil their expectations, but the reality turns out to be different, and their expectations gradually decrease.

However, if they start with the ideas such that it would be good if it could do something at a minimum level, they will find that they can do more things than they expected, and they will be able to realise things that they have never done before, and it is possible to occur that their desire for selfrealisation will become higher. This case, I think, would be a good one. So it is often found in the case of cochlear implant users that the greater the initial expectations, the greater the disappointment.” (Subject 1)

3.6 Policymaking regarding the development and use of cyborg technologies for people with disabilities

Finally, both of the participants emphasised the importance of communication among engineers, policymakers, researchers and people with disabilities in order to develop and utilise the information systems with cyborg technologies to assist people with disabilities.

“When researching this kind of assistive or cyborg technology, I think it is important not to force people with disabilities to adopt this kind of technology [after the development of it], but to examine [beforehand] their imagination of what they would be able to achieve using such technology. That is, first we need to investigate what they desire to do or achieve using this kind of technology, and, after that, we have to consider what interface we need to develop to realise their desire.

Therefore, I think it is essential to develop and introduce these technologies based on what people with disabilities would like to realise, taking the diversity of them and their tendencies and major needs into account.” (Subject 1)

“Human beings are flexible, so I think I will adjust to it [cyborg technologies such as BMI]. If someone asks me in the future, ‘What about this BMI or cyborg technology?’ I don’t know how I would respond. Perhaps, at that time and place, if it works well, I would think that it’s fine. However, if I experienced some kind of harmful effects as a result of the use of such technology, I would put them into adequate words and communicate to others.

To be honest, cyborg technology is somewhat scary for me, but simultaneously somewhat exciting. So, I don’t think I would wish that this kind of research didn’t go ahead, although the development of the technology may affect my vocational life. Therefore, I think it is absolutely important that people with disabilities are allowed to participate in this sort of research as well as in the processes of cyborg technology development and that we work together and exchange opinions.” (Subject 2)

4 Discussion

The experiment and the semi-structured interview surveys conducted in this study with two people with disabilities provided several eye-opening findings. In particular, the participants’ comments described in Sections 3.1 and 3.2 show that their fundamental recognition of their bodies and that their feelings for and relationship development with others are definitely different from those of people without disabilities. This finding makes us aware of the underlying ethical issues and questions related to the dignity of people with disabilities and the nature of human beings in the modern society where cyborg technologies are widely available, as follows.

- Should physical disabilities be corrected?
 - If yes, why and to what extent?
 - What is a disability?
- How the dignity of people with disabilities be respected?
 - Should people with disabilities be accepted by society as people with the individual characteristics of disabilities under the conceptions of diversity and inclusion?

On the other hand, in cases where cyborg technology is used for people with disabilities—most notably, the use of BMI technology for an individual who is unable to move on their own—it appears to be more readily expected and perceived in

society, with obvious benefits for people with disabilities and their caregivers. However, the two participants agreed that, when it comes to the use of such technology, what people with disabilities want to do, or their life goals, should be considered first and then mutually discussed among developers, policymakers, and users. In other words, the results of this survey may imply that there are social risks or ethical concerns that the use of such powerful technology will become the goal, making it difficult to recognise that the means should be defined by the goals, and the means and the goals have been replaced, as the benefits and usefulness of BMI devices are very easy to imagine and understand for the general public. As Imamichi (1990) pointed out, if powerful means such as electronic technologies are available, then an achievable goal can be attained using those means, and thus, the means tend to rule the goal.

Furthermore, based on the results of the surveys in this study, it appears that several issues concerning the use of BMI and cyborg technology for people with disabilities have been implied to be important, including the operational, ethical and institutional issues listed below.

1. What is the BMI system's actual usefulness, possibilities, and risks for people who were born without any physical sensation?
2. Do people with disabilities, both congenital and acquired, cause differences in the results of this type of experimental survey using BMI devices, and if so, what are the reasons and backgrounds for such differences?
3. How can it be recognised and assessed whether or not the technical assistance provided by cyborg technologies to people with disabilities exceeds the ability of people without disabilities?
4. How does the use of cyborg technology, particularly BMI systems, for people with disabilities positively or negatively impact the development of their identities and relationships with others?
5. How can such technology use change people's perceptions of themselves or their relationships with caregivers or others? Which characteristic of people with disabilities is considered to make them more vulnerable to such influences? Is it possible to prevent such influences on people with disabilities by implementing some kind of policy on the use of cyborg devices?
6. How far should the performance of cyborg technologies and/or BMI systems be certified before they are actually implemented and applied to society, taking the gaps between expectations of the function of the BMI systems and their actual usage, as well as ethical and legal concerns about the responsibility for system errors or malfunctions, into account?
7. How can access to cyborg and emerging technologies be maintained in a fair and equitable manner for people with abilities?
8. How can we ensure the opportunities for and improve the conditions of this type of research which is conducted in collaboration with

professionals in BMI technology and technology ethics, involving people with disabilities?

Before cyborg and BMI devices are used in daily life of people with disabilities and in broader society, it is necessary to deliberate these issues and develop appropriate policies, and countermeasures when necessary, while also improving the availability of BMI devices.

5 Concluding remarks

In this study, the experimental and the semi-structured interview surveys were conducted with two people with disabilities who were engaged in supporting other people with disabilities, in a similar manner to our previous studies (Orito et al., 2021a; 2021b; 2020a; 2020b). Both of the participants provided valuable comments and insightful suggestions for the development of appropriate experimental environments and the application of BMI devices or cyborg technologies for people with disabilities. Their perspectives are very unique and may be inconceivable to people without disabilities, which may also indicate the importance of involving participants with disabilities in academic studies and policy-making processes to address the ethical issues and risks of using cyborg technologies such as the BMI system. In addition, some points they suggested in the surveys are insightful to explore the ethical dilemmas which would be caused by BMI usage for non-medical purposes.

To enrich this research, more experimental surveys inviting people with variety kinds of disabilities should be conducted with due considerations to their health conditions and with establishing an experimental setting that is appropriate for them. Moreover, it is necessary to cultivate a better understanding of the human physiological nature and the distinctive sensitivities of people with disabilities.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers 22K02063, 20K01920, 19K12528, the Kurata Grants subsidised by the Hitachi Global Foundation, and the Meiji University Grant-in-Aid for the international collaborative research project “Cyborg Ethics.” We also appreciate Professor Shizuka Suzuki of Ehime University, Mr. Yukio Takeda, Dr. Yoshitaka Moritsugu and all the participants in the experiments and researchers for their support of our study.

We certify that all procedures of the experiments performed in this study were in accordance with the ethical standards of the research ethics committee established at the Faculty of Collaborative Regional Innovations, Ehime University (issued June 2021, No. 2021-01).

References

- Fukushi, T. & Sakura, O. (2007). Ethical implementation of research and development on brain-machine interface. *Keisoku to Seigyo*, 46(10), 772-777.
<https://doi.org/10.11499/sicejl1962.46.772> (in Japanese) .
- Gilbert, F., Cook, M., O'Brien, T. & Illes, J. (2019). Embodiment and estrangement: Results from a first-in-human 'Intelligent BCI' trial. *Science and Engineering Ethics*, 25(1), 8396.
- Grübler, G. & Hildt, E. (eds.) (2014). *Brain-computer interfaces in their ethical, social and cultural context*. Dordrecht: Springer.
- Imamichi, T. (1990). *Eco-Ethica*, Tokyo: Kodansha.
- Isobe, T. (2013). The perceptions of ELSI researchers to Brain-Machine Interface: Ethical and social issues and the relationship with society. *Journal of Information Studies*, 84, 4763.
- Murata, K., Adams, A. A., Fukuta, Y., Orito, Y., Arias-Oliva, M. & Pelegrín-Borondo, J. (2017). From a science fiction to reality: Cyborg ethics in Japan. *Computers and Society*, 47(3), 72-85.
- Murata, K., Fukuta, Y., Orito, Y., Adams, A. A., Arias-Oliva, M. & Pelegrín-Borondo, J. (2018). Cyborg athletes or technodoping: How far can people become cyborgs to play sports? Retrieved from
https://www.researchgate.net/publication/327904976_Cyborg_Athletes_or_Technodoping_How_Far_Can_People_Become_Cyborgs_to_Play_Sports.
- Murata, K., Arias-Oliva, M. & Pelegrín-Borondo, J. (2019). Cross-cultural study about cyborg market acceptance: Japan versus Spain. *European Research on Management and Business Economics*, 25(3), 129-137.
- Orito, Y., Yamamoto, T., Sai, H., Murata, K., Fukuta, Y., Isobe, T. & Hori, M. (2021a). How a brain-machine interface can be helpful for people with disabilities? Views from social welfare professionals. In Mario Arias Oliva, Jorge Pelegrín Borondo, Kiyoshi Murata & Ana María Lara Palma (eds.), *Moving technology ethics at the forefront of society, organisations and governments* (pp. 103-115). Lgroño: Universidad de La Rioja.
- Orito, Y., Yamamoto, T., Sai, H., Murata, K., Fukuta, Y., Isobe, T. & Hori, M. (2021b). The ethical issues of the use of BMI in social welfare: An experimental and semi-structured interview study with professionals. *The Proceedings of National Conference of Japanese Society for Management Information 2021 Spring* (in Japanese).
- Orito, Y., Yamamoto, T., Sai, H., Murata, K., Fukuta, Y., Isobe, T. & Hori, M. (2020a). The ethical aspects of a "psychokinesis machine": An experimental survey on the use of a brain-machine interface. In Arias-Oliva, M. et al. (Eds.), *Societal Challenges in the smart society Ethicomp book series* (pp. 81-91). Lgroño: Universidad de La Rioja.
- Orito, Y., Murata, K. & Suzuki, S. (2020b). Possibilities and ethical issues surrounding brainmachine interfaces in the realm of social welfare: Potential for use by people

with disabilities based on results from psychokinesis experiments. *E-poster at the 68th Academic Conference of Japanese Society for Social Welfare* (in Japanese).

Wahlstrom, K. (2018). *Privacy and brain-computer interfaces*, Doctoral Thesis, De Monfort University, UK.

IT Ethics Protocol and Usability for Enhancing Culture in the Spanish Museums

Raquel García-Martín¹[0000-0003-1657-9558], Ana María Lara-Palma²[0000-0002-0127-7963],
Bruno Baroque Zanón³[0000-0002-4993-204X] and Rodrigo Alonso Alcalde⁴[0000-0002-8853-1150]

¹ Universidad de Burgos, Burgos, Spain

² Universidad de Burgos, Burgos, Spain

³ Universidad de Burgos, Burgos, Spain

⁴ Museo de la Evolución Humana, Burgos, Spain

rgm0000@alu.ubu.es

Abstract. New technologies have changed the way of communicating and disseminating science. These new inclusions raise new possibilities, new challenges, but also new ethical conflicts. Museums interact with digital media in different ways, due to the particular characteristics of each one of them, whether by theme, number of staff members, ownership, location or target audience. The objective of this work is to analyze the current situation of museums with respect new technologies and IT ethics protocol and its usability for enhancing culture in Spain. For this, a study methodology has been followed by questionnaires, carried out on experts in 18 Spanish museums. The results obtained allow us to infer that the adaptation of museums with respect to new technologies must be reconsidered at the present time. Currently, digital media are an essential tool for identifying, generating, disseminating and updating the necessary knowledge to improve the development of the objectives that sustain a museum.

Keywords: Knowledge Management, Dissemination of Knowledge, IT Ethics, New Technologies, Museum, Digital Bases.

1 Introduction

Museums have traditionally been avant-garde sites where societies have been applying new technological discoveries with the aim of displaying, preserving and disseminating their cultural legacy. Therefore, it is not surprising that, during the 21st century, many museums have functioned as laboratories where the new technological resources and communication channels that have been developed in recent times have been systematically applied. Virtual visits, 360o tours, artificial intelligence museographic resources or social networks, are part of the daily life of museums today. But has the implementation of these new advances been common to all museums? What types of museums have applied greater technological development? Do they use new

technologies to communicate with their visitors? Do they train their employees in the mastery of new technologies? Do they develop ethical protocols in the implementation of new technologies? The present work shows the statistical results of the survey carried out on Spanish museums of different types (national, regional, etc.) and themes (fine arts, archaeology, etc.). The comparative analysis of these results allows us to identify 1) to what extent new technologies are used, 2) in what type of Spanish museums they have a greater implementation and 3) IT ethical protocols to enhance culture at a museum.

2 Theoretical Framework

The discipline of Knowledge Management (KM) is booming. The importance of the knowledge available in an organization and its management, is increasingly taken into account. However, KM in museums is a discipline that has not been so widespread as has been in private companies, more recent than that which has been applied in companies. The transmission of knowledge is carried out by different means such as books, documents, digital platforms and interpersonal contact for learning, where digital media acquire a greater role. Information management and KM, as well as the people involved in it, have existed since the dawn of humanity, when people, in their eagerness to communicate, used different techniques to record, exchange, transmit and share events and insights of their work and way of life, through cave paintings, papyri and other media (Ledo Vidal and Pérez Araña, 2012). This process continues today, using other means. The knowledge acquisition and use of KM practices, positively influence strategic and operational improvement in organizations (Jayasingam, et al., 2013). In addition, knowledge and information acquire more and more value (Sánchez López, 2011).

The drawback found is, the difficulty of quantifying the benefits and usability, since knowledge, as such, is an intangible value. Currently, new ICT technologies must be considered in this management, since they have become a first-rate informative medium in this century.

Dimensions of creation, transfer and storage, as well as their application and use, are part of the concept of organizational knowledge (Tarí and Fernandez García, 2009). So much so, that many experts talk about the need to standardize information, documentation and knowledge through rules and procedures (Smit, 1986), as well as, the option of evaluating them (Kim, 2006). How can knowledge be measured? Knowledge represents an intangible and strategic asset that generates competitive advantages. Results of KM, favor to safeguard, intellectual assets of institutions, as a characterization factor and source of value (Espinoza, 2013), hence, the importance of KM audit and questionnaires as tool used for its development. Limitations are observed to be applied to any organization, because it's not easy diagnose the development of KM processes and check the status of significant variables for an audit process (Medina Nogueira, et al. 2019) (Medina Nogueira, et al., 2019). There are many organizations and companies that have adapted to the new ISO 9001 Quality Management systems, where the most innovate section of ISO 9001:2015 is 7.1.6. On Organizational

Knowledge Management. This section was not considered in previous versions, but highlights the importance of prioritizing the sharing of available knowledge is taken into account, linking information, technological means and people, so that a multidisciplinary team, may be the best option (Gómez Martínez, 2015).

The absence of the right processes for knowledge retrieval and reuse can lead to organizational knowledge retention and loss (Levallet and Chan, 2019). Negative impact of knowledge loss can be addressed with proper KM. And measurement concepts, help us understand the nature of the impact of knowledge loss (Massingham, 2018).

Digital and Technological Environments are one of the emerging categories within KM (Ayala, Cuenca-Amigo and Cuenca, 2019). As time passes, new forms and technologies of exchange appear. As examples, Internet, which provides the necessary infrastructure for communication, exchange and development bases; or the intensive use of technology in mobile devices, where numerous services that were previously independent are integrated. These technologies modify the criteria of space and time, and thereby, the processes in the different spheres of society are globalized and streamlined, providing greater potential, not only to the improvement of said processes, but also to the exchange of information and the generation of knowledge in people (Ledo Vidal and Pérez Araña 2012).

More and more organizations are making available to their employees, technological platforms that allows them to manage their knowledge, with the aim of transferring it to workplace and generating authentic organizational learning (Redondo Duarte, et al., 2017). The exchange of knowledge with electronic networks within a corporate group can provide numerous benefits that are reflected in time or costs (Sedighi, et al., 2017). Learning communities or discussion, and on-line sharing and evaluation, they are an important part of improving creativity and knowledge. The observation and peer evaluation of group assignments and creativity related feedback enhance the learning of knowledge and dispositions. A KM model, is effective in improving creativity knowledge, dispositions and skills. (Yeh, Yeh and Chen, 2012).

Degree of knowledge and use of scientific digital social networks, are valued positively. They present a growing tendency, in management and usefulness (Rodríguez Fernández, Sánchez-Amboage and Martínez Fernández, 2018). Its role in knowledge creation, shows a complex relationship between various facets of social networks and the ability of their members to create knowledge (Nieves and Osorio, 2013). Knowledge sharing benefits both knowledge sharers and knowledge recipients, from a learning theory perspective (Zhu, Chiu and Infante Holguin-Veras, 2018); as it is proven in cases where the use of technology has been involved to improve teaching-learning experiences (Harncharnchai and Saeheaw, 2018). Today, organizations take advantage of social media tools to improve performance, since they allow people to connect, facilitating a broader and more distributed communication, which favors innovation and the creation of knowledge. Could also provide and effective solution to help museum fill the gap of lacking experts (Sarka and Ipsen, 2017).

As a result of the rise of social networks, platforms or App`s, new challenges arise, such as structural holes that can lead to a greater spread of risks on networks, being important to detect and reduce or prevent the spread of the same (Song, et al., 2020).

These systems are attractive and encourage exchange of intergenerational knowledge, therefore, they have the potential to help in cultural preservation of the associated communities (Sieck and Zaman, 2017).

The value of digitalization adds to this information and its dissemination through cultural connectivity networks based on research, creation, innovation and shared services in the field of culture, making it a sustainable resource both economically and socially (Carrasco Garrido, 2012).

Does the same thing happen in museums?

The increasing demand of technological facilities for museums, galleries and archives has led to the need for designing practical and effective solutions for managing the digital life cycle of cultural heritage collections (Dragoni, Tonelli and Moretti, 2017). We find an increasing tendency in the musealization / patrimonialization in electronic site or use of the Virtual Museum, where the tangible and intangible aspects are integrated (Farjalla and Lima, 2019), since as a communication tool, it is an excellent language to general and specialized public. The research community tends to be a closed sphere sometimes, especially for non-specialized public, which contrasts with the social duty that knowledge should have. An interesting initiative to revert this situation is, virtual archaeology, which provides a broadcasting medium to get information to the public, whether it is specialized or not, showing itself as a very suitable language to transmit ideas to that society in a much more intelligible and less abstract way than traditional formulas used in archeology (García Carpintero López de Mota and Gallego Valle, 2018).

Use of prototypes in digital media in museums also favors the KM framework of Socialization-Externalization-Combination-Internalization. Where a better understanding of how and why team members can take advantage of prototypes to create, transfer, combine and embody knowledge is reached (Mason, 2015). The media competence of a museum, can be developed from the point of view of didactic design and its characteristics. Currently, development and production of multimedia products (magazine articles, video and audio podcast) are very valued in area of knowledge transfer (Vogt and Maschwitz, 2014).

In today's interconnected pluralistic society, communications management through public relations policies is vital to the visibility. The viewing public participates actively in the knowledge society, and the museum staff need to adopt to this demand. (Martínez Peláez, Oliva Marañón and Rodríguez Rivas 2012).

3 Work Methodology and Empirical Study

This section shows assessments collected from questions in the questionnaire about the use or lack of the presence of new technologies in the workplace. In order to assess current status of KM in museums with respect to new technologies, with the final objective of identify, generate, disseminate and update knowledge. And we carry out the comparative analysis of usability parameters and implementation of new technologies, in national museums in Spain. To do this a questionnaire has been carried out that includes the variables of Identification, Generation, Dissemination and

Updating of knowledge (Lara Palma, 2006), associated with new technologies that can influence job performance.

The questionnaire is self-made, taking into account numerous articles on questionnaires based on the literature of Medina Nogueira, et al. 2019, Handzic, Lagumdzija and Celjo, 2008, Liebowitz and Suen, 2000 and Pulido and Quintana, 2013 among others. Below are the museums that collaborate in the study, including their theme, location and ownership:

Table 1. Museums participating in the research and their characteristics.

N ^o	Museum	Thematic	Subtheme	Location/ Town	Region	Territorial Scope
1	ARQVA	Archaeology	Subaquatic	Cartagena	Murcia	National
2	CAB	Art	Contemporary	Burgos	Castilla y León	National
3	MEH	Archaeology	Human Evolution	Burgos	Castilla y León	Regional
4	Museo Altamira	Archaeology	Prehistory	Santillana del Mar	Cantabria	National
5	Museo Cerralbo	Art	Legacy of the Marquis of Cerralbo	Madrid	Madrid	National
6	Museo de Ávila	Archaeology and Fine Arts	Ávila's Heritage	Ávila	Castilla y León	Provincia 1
7	Museo de Burgos	Archaeology and Fine Arts	Burgos's Heritage	Burgos	Castilla y León	Provincia 1
8	Museo de Palencia	Archaeology and Fine Arts	Palencia's Heritage	Palencia	Castilla y León	Provincia 1
9	Museo de Valladolid	Archaeology and Fine Arts	Valladolid's Heritage	Valladolid	Castilla y León	Provincia 1
10	Museo de Salamanca	Archaeology and Fine Arts	Salamanca's Heritage	Salamanca	Castilla y León	Provincia 1
11	MECYL	Ethnographic	Heritage Ethnographic CYL	Zamora	Castilla y León	Provincia 1
12	Museo del Greco	Art	Greco Artist	Toledo	Castilla la Mancha	Nacional
13	MUSAC	Art	Contemporary	León	Castilla y León	Regional
14	Museo Nacional de Antropología	Anthropology	Spanish Anthropology and Ethnographic	Madrid	Madrid	National

15	Museo Nacional de Arte Romano	Archaeology	Roman	Mérida	Extremadura	National
16	Museo Romanticismo	Art	Roman	Madrid	Madrid	National
17	Museo Sefardí	Art	Hispano-Jewish and Sephardic	Toledo	Castilla la Mancha	National
18	Museo Sorolla	Art	Joaquín Sorolla Artist	Madrid	Madrid	National

The allowed responses to the questionnaire range from 1 to 5 (1 being the most unfavorable value and 5 being the most favorable). Values are grouped and named by category: 1 and 2, corresponding to “Bad” category; 3, corresponding to “Regular” category; 4 and 5, to the “Good” category and unanswered questions, “n/a” (No Answer). The application of this methodology through questionnaires to expert workers in museums, has offered the chance to grasp the reality about the insufficient training in new technologies of workers. That is why it seems necessary to articulate a model that would enable museums that are inclined to, to implement an action protocol in this area.

Each organization has particular characteristics in many aspects, like size, area of work, technical and human capacity, number of workers, subcontracted services, activities or organization (Medina Nogueira, et al., 2019) (Liebowitz and Suen, 2000), (Handzic, Lagumdžija and Celjo, 2008), (España Pulido, 2013). But thanks to this questionnaire, a result is obtained for each characteristic of a museum, as well as a global qualification of all or a group of museums. Taking this data into account, the protocol must be customized for each museum according to its particular characteristics, for a greater effectiveness of KM. This protocol will facilitate, among other aspects, the dissemination processes in order to offer knowledge to a greater share of the public. In the case of museums, they can also bring their works of art, pieces or fossils that are non-fungible materials, since these non-fungible items represent, the culture of the digital age.

4 Results

New technologies are here to stay, and in the field of museums they are currently an important tool and support for KM. Each museum should adapt to its own situation based on its characteristics.

Regarding the answers from the questionnaires, we have divided the answers in seven groups, according to each museum’s characteristics: general museums (100% surveyed), national museums (50% of the total surveyed), regional museums (16’6% of the total surveyed), provincial museums (27’7% of the total surveyed), archaeological museums (50% of the total surveyed), art museums (38’8% of the total surveyed) and

fine arts museums (27.7% of the total surveyed). It can be observed that answers vary according to the characteristics of the museum that divides the sample into groups.

From the questionnaires, it is possible to analyze and find points for improvement and deficiencies depending on the characteristics of the museum. Regarding the sense of culture through expositions or other activities by using new technologies, findings allow us to conclude the need of IT ethical protocols for enhancing culture, because public intervenes with their feedback with different opinions.

Two questions are taken from the questionnaire on use of new technologies in museums:

- Perception about the training offer respondents receive in new technologies for the performance of the job in Spanish museums.
- Perception of the respondents about the Technological means (database, web, mail or social networks) that the museum has to update knowledge.

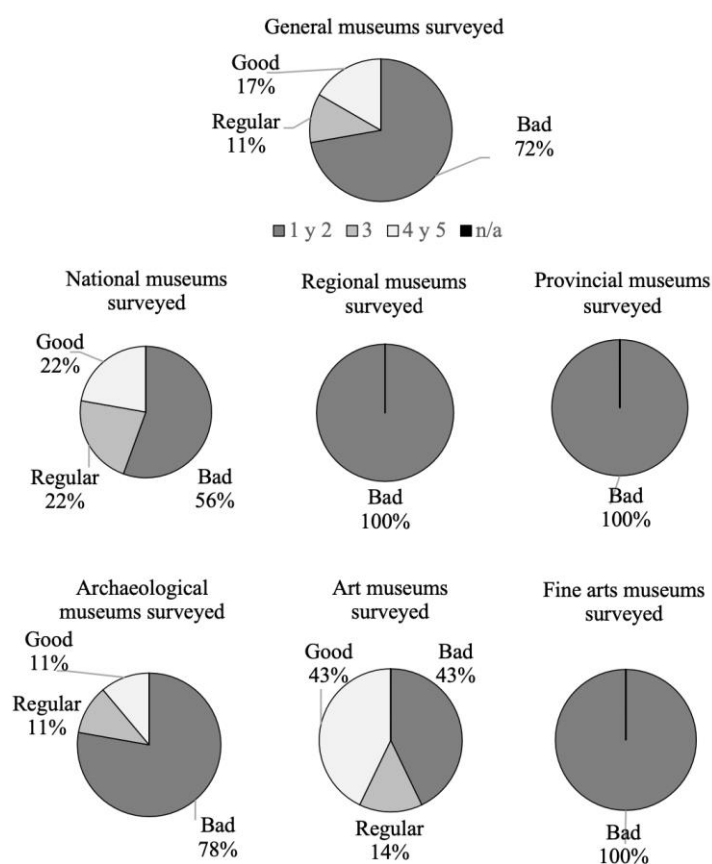


Fig. 1. Perception about the training offer respondents receive in new technologies for the performance of the job in Spanish museums.

Regarding the training offer received by the employees of each type of museum, 72% of the participants in general museums, perceive it as “bad”, to this first question. But it is not the same in all fields. It is 78% in archaeological museums, 56% in national and 100% in regional, provincial and fine arts museums. (See Fig.1.). There is a lack of courses for workers, in new technologies.

Regarding to the perception of the respondents about the technological means (database, web, mail or social networks) that museum has to update knowledge, results can be found on Fig.2.

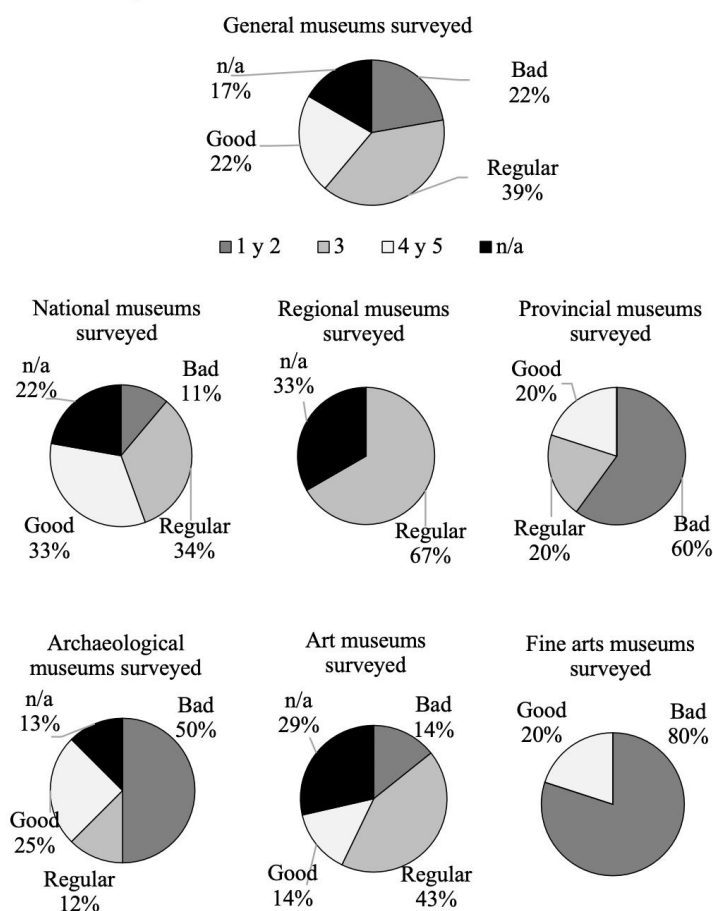


Fig. 2. Perception of the respondents about the technological means (database, web, mail or social networks) that museum has to update knowledge.

Finally, regarding the updating of knowledge, only 22% of those general museums surveyed, perceive it as “good”. It also varies according to areas. It is 33% in national museums, 25% in archaeological museums, 20% in provincial and fine arts museums,

14% in art and 0% in regional museums (See Fig.2.). Knowledge updating is highly improvable in museums.

5 Conclusions and Future Work

This study allows parameterizing KM processes with percentage values. It allows to open a field of study to know situation in KM in which a museum or organization is located.

Using the methodology of questionnaires, the lack of new technology courses and lack of knowledge updating for workers in museums from different areas, has been detected. Currently, new technologies are essential for the functions carried out in the museum, to spread knowledge and bring museum pieces closer to a greater number of the public. All this information gives rise to a model to generate a protocol. Model and protocol, must take into account characteristics of each museum.

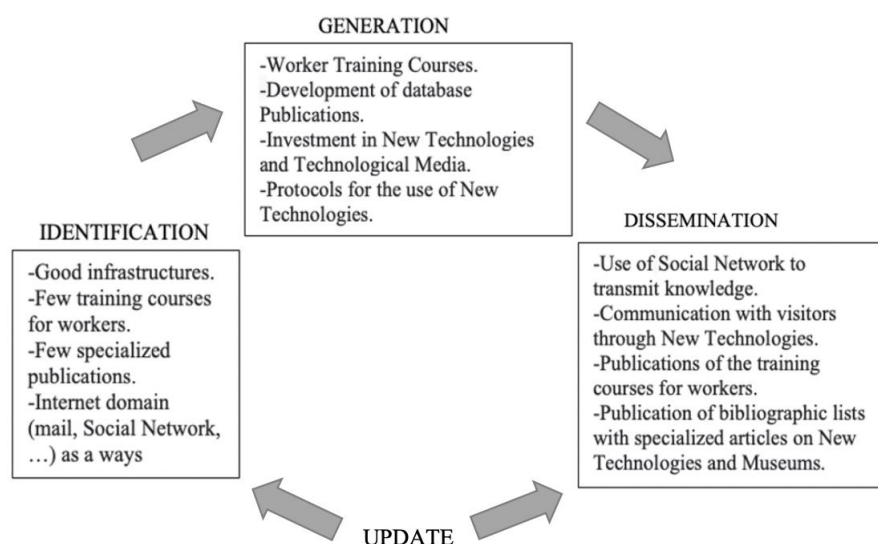


Fig. 3. Protocol for the implementation of New Technologies in KM in museums.

In a general model, processes of identifying, generating, disseminating and updating knowledge, must be taken into account. Identification is influenced by, good infrastructures, worker training courses or internet domain. To favor generation of knowledge, it is good to develop an adequate database, invest in new technologies and technological media, or have a protocol for use of new technologies. For improve dissemination of knowledge, tools such as social networks or communication with visitors through new technologies can be used. Update of knowledge, present in the previous processes, also improves when applying the mentioned tools. (See Fig. 3)

Acknowledgements

We greatly appreciate collaboration of experts who have participated in completing the questionnaires. Without them, this study would not have been possible.

References

- Ayala, I., Cuenca-Amigo M. and Cuenca J. (2019). Examining the state of the art of audience development in museums and heritage organizations: a Systematic Literature review. *Museum Management and Curatorship*, 1–22.
- Carrasco Garrido, R. (2012). Documentar el patrimonio: cuando la información se transforma en un recurso sostenible. *Museos.es: Revista de la Subdirección General de Museos Estatales* 7 (8), 120–125.
- Dragonì, M., Tonelli S. and Moretti G. (2017). A knowledge management architecture for digital cultural heritage. *Journal on Computing and Cultural Heritage* 10 (3), 1–18.
- Espinoza, L. (2013). Knowledge Management through Institutional Memory. Case Study: FUNDAMETAL. *Strategos: Artículos y Ensayos. Institucional, Memoria* 5 (10), 11–17.
- Farjalla D. and Correia L. (2019). Musealization/Patrimonialization on electronic site Virtual Museum: integration of tangible and intangible aspects. *Memória e Informação* 3 (2), 86–105.
- García Carpintero López de Mota, J. and Gallego Valle, D. (2018). Archaeology of the military orders in Castilla-La Mancha and the virtual reconstruction of its heritage. *Virtual Archaeology Review* 8 (19), 76–88.
- Gómez Martínez, J. A. (2015). *Guía para la aplicación de UNE-EN ISO 9001:2015*. Editado por ISBN: 978-84-8143-911-3.
- Handzic, M., Lagumdžija, A. and Celjo, A. (2008). Auditing knowledge management practices: model and application. *Knowledge Management Research & Practice* (6), 90–99. https://www.researchgate.net/publication/263022587_Auditing_knowledge_management_practices_Model_and_application
- Harncharnchai, A. and Saeheaw, T. (2018). Context-aware learning using augmented reality and WebQuest to improve students' learning outcomes in history. *International Journal of Innovation and Learning* 23 (3), 283–303.
- Jayasingam, S., Mahfooz, A., Ramayah T., and Jantan M. (2013). Knowledge management practices and performance: Are they truly linked?. *Knowledge Management Research and Practice*, 11 (3), 255–264. https://www.researchgate.net/publication/258052756_Knowledge_management_practices_and_performance_are_they_truly_linked
- Kim, J. (2006). Measuring the Impact of Knowledge Management. *IFLA Journal* 32 (4), 362–367.
- Lara Palma, A.M. (2006). Modelo de los seis estadios de rentabilidad del conocimiento aplicación de una herramienta de gestión de intangibles para pequeñas y medianas

- empresas del sector servicios. Doctoral Thesis, Universidad de Burgos, Dpto. Ingeniería Civil, Spain.
<https://www.educacion.gob.es/teseo/imprimirFichaConsulta.do?idFicha=134884>
- Ledo Vidal, M.J. and Pérez Araña, A.B. (2012). Information management and knowledge / Gestión de la información y el conocimiento. *Revista Cubana de Educación Médica Superior*, (1), 474–484.
<http://scielo.sld.cu/pdf/ems/v26n3/ems13312.pdf>
- Levallet, N. and Chan, Y.E. (2019). Organizational knowledge retention and knowledge loss. *Journal of Knowledge Management* 23 (1), 176–199.
- Liebowitz, J. and Suen, C.Y. (2000). Developing knowledge management metrics for measuring intellectual capital. *Journal of Intellectual Capital*, (1), 54–67.
https://www.researchgate.net/publication/242021931_Developing_knowledge_management_metrics_for_measuring_intellectual_capital
- Martínez Peláez, A., Oliva Marañón, C. and Rodríguez Rivas, A. (2012). Comunicación interna y externa en el Museo Reina Sofía: interacción del público en un entorno virtual. *Telos: Cuadernos de comunicación e innovación* (90), 71–78.
- Mason, M. (2015). Prototyping practices supporting interdisciplinary collaboration in digital media design for museums. *Museum Management and Curatorship* 30 (5), 394–426.
- Massingham, P.R. (2018). Measuring the impact of knowledge loss: a longitudinal study. *Journal of Knowledge Management*, 22 (4), 721–758.
<https://ro.uow.edu.au/buspapers/1417/>
- Medina Nogueira, E., Nogueira Rivera, D., El Assafiri, Y. and Medina León, A. (2019). Proposal of a questionnaire for the development of the knowledge management audit. ResearchGate. *Universidad y Sociedad*. 11 (4), 61–71.
https://www.researchgate.net/publication/338002812_Design_and_Application_of_A_Questionnaire_for_the_Development_of_the_Knowledge_Management_Audit_Using_Neutrosophic_Iadov_Technique
- Nieves, J. and Osorio J. (2013). The role of social networks in knowledge creation. *Knowledge Management Research and Practice* 11 (1), 62–77.
- España Pulido, F.A. (2015). Encuesta sobre la aplicación de los conceptos y técnicas de auditoría a la Gestión del Conocimiento. Thesis. Universidad del Valle, Santiago de Cali, Colombia.
<https://bibliotecadigital.univalle.edu.co/bitstream/handle/10893/9027/CB-0516274.pdf?sequence=1&isAllowed=y>
- Redondo Duarte, S., Navarro Asencio, E., Gutiérrez Vega, S. and Iglesias Ortega I.I. (2017). Improvement of learning in organizations through virtual communities. *Revista Complutense de Educación* 28 (1), 101–123.
- Rodríguez Fernández, María Magdalena, Eva Sánchez-Amboage, and Valentín Alejandro Martínez Fernández. (2018). Use, knowledge and assessment of the scientific digital social networks in the Galician universities. *El Profesional de la Información* 27 (5), 1386–6710.

- Sánchez López, M.L. (2011). La relevancia de la gestión del conocimiento en las empresas. *Apuntes del CENES*, 30 (51), 223–237.
<https://revistas.uptc.edu.co/index.php/cenes/article/view/40>
- Sarka, P., and Ipsen C. (2017). Knowledge sharing via social media in software development: A systematic literature review. *Knowledge Management Research and Practice* 15 (4), 594–609.
- Sedighi, M., Lukosch S., Van Splunter S., Brazier F.M.T., Hamed M., and Van Beers C. (2017). Employees' participation in electronic networks of practice within a corporate group: Perceived benefits and costs. *Knowledge Management Research and Practice* 15 (3), 460–470.
- Sieck, J., and Zaman T. (2017). Closing the distance: Mixed and augmented reality, tangibles and indigenous culture preservation. *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*.
- Smit, J.W. (1986). Vocabulário controlado e controle de vocabulário em arquivos. *Acervo.Arquivo Nacional (Brazil)* 31 (3), 46–56.
- Song, Y., Yang N., Zhang Y., and Wang J. (2020). Do more structural holes lead to more risk propagation in R&D networks? *Management Decision* 58 (1), 39–57.
- Tarí, J.J., and Fernández García M. (2009). ¿Cómo se puede medir el conocimiento y la calidad? Administrando en entornos inciertos. *XXIII Congreso Anual AEDEM (2009) Universidad de Alicante*. 1–14.
- Vogt, S., and Maschwitz A. (2014). The non-cartesian way: Developing media competence through media production. 16 (2), 13–25.
- Yeh, Y., Yeh Y., and Chen Y. (2012). From knowledge sharing to knowledge creation: A blended knowledge-management model for improving university students' creativity. *Thinking Skills and Creativity* 7 (3), 245–257.
- Zhu, Y.Q., Chiu H., and Infante Holguin-Veras E.J. (2018). It is more blessed to give than to receive: examining the impact of knowledge sharing on sharers and recipients. *Journal of Knowledge Management* 22 (1), 76–91.

Rationalizing Emotion through Technology: Or Addressing the Political and Liberal Narrative in Emotion Technology

Eugenia Stamboliev^[0000-0003-3839-1636] and Mark Coeckelbergh^[0000-0001-9576-1002]

University of Vienna, Vienna, Austria

eugenia.stamboliev@univie.ac.at

Abstract. The new emergence of algorithmic emotion technology can supposedly measure people's emotions and steer public moods and sentiments, which raises red flags for political theory but also for ethical debates. Not surprisingly, applications, like 'sentiment analysis' or 'emotional AI', are critiqued from several sides; from being seen as reductionist in regard to detecting a fluidity and diversity within human emotions and cultural expressions, to being technologies that foster political manipulation, to also growing into new industries that invisibly exploit and target moods and sentiments. Most critical foci being on the latter, ours will be on the political lineage between liberal views on emotions as liabilities and the way in which emotion technology controls and measures emotions. We call this the liberal narrative within emotion technology, which we associate with the particular elements in the work of Habermas or Rawls. We develop two critiques around the power of the liberal narrative: First, we show that a liberal view on emotions as liabilities has led to controlling and measuring technologies in the first place, which illustrates how political narratives shape technology. Second, we utilize the political theory of Mouffe to counter this liberal narrative but also its technological manifestation. By revisiting the antagonistic value of emotions and or dissensus that she proposes, we resume that neither liberal concepts nor emotion technologies leave any conceptual space for the empowering force of emotions like found in movements of resistance. By informing a critique on emotion technology through political theory, we show that technology design is continuously shaped by, and shaping, political narratives on different levels and that technology can play a role in extending, or breaking up, conceptual, economic, political narratives or hegemonies once identified.

Keywords: Affect technology, politics of emotion, computational politics, liberal and antagonistic political theory, cognitivism

1 Introduction

Politics is increasingly shaped by the growing power of computation (Sudmann, 2019; Coeckelbergh, 2020), but politics is also shaped by a fear of uncontrollable emotions (Miller, 2011; Marcus, 2000). Recently, both have been united in the algorithmic sphere. Through new technologies, like 'emotional AI' (McStay, 2018) or 'sentiment analysis' (Chakriswaran et al., 2019), we see new computational software being developed that analyses, extracts or targets emotional states (or is *claimed* to do so) via speech or visual analysis. While these applications are partially unregulated and invisible within digital platforms, certain emotion technologies might be limited due to technical shortcomings, but they have controlling and politically powerful tendencies. Their manipulative potential lies in those technologies merging with 'computational politics' (Tufekci, 2014) and with cognitivist norms, political ambitions with unregulated business interests (Bösel and Wiemer, 2020). Overall, emotion technology poses a risk for democracy. While Rhodes (2022) speaks of filter bubbles leading to homogeneous information streams, Levendusky (2013) points to the growing force of digital architectures that polarize people through keeping them on 'informational diets'. At the same time, researchers like König and Wenzelsburger (2020) emphasize that technologies like AI will even increase the power of computational politics 'because it affects public opinion formation' (2) and its influence over regulating voters' and citizens' affects. According to Howard (2020), if combined with AI, then emotion technology could 'probably be used to charm and provoke voters, spread misinformation, and take advantage of rich personal data to make powerful emotional appeals. For countries that hold elections, AI-driven fake users will become a new, serious, and internal threat to their democracy' (146).

Having mentioned certain critical angles around emotion technology, we emphasize an underdeveloped angle in this discussion. We will examine emotion technology as a wider techno-political trend, not as a specific application, to trace what we call its *liberal narrative*. We will illustrate how political theory, scientific norms and computational technology co-narrative the value of emotions with consequences for us either seeing emotions as liabilities or as empowering forces. The technification of emotions, we will argue, is not only an technological challenge nor only a scientific goal, but it has political power and embeds a political narrative in its design ambition to control emotions; we call this the *liberal narrative*. Due to our very narrow focus as philosophers, not political scientists, we only emphasize one specific moment of liberal thinking; the devaluation and fear of emotions referring which we revisit through the work of Habermas and Rawls. We will not provide an overview on liberal or antagonistic theory, but only pick moments to contextualize our work. One of which would be to say that liberal theories do not consider emotions as empowering but rather as problematic political forces (Papacharissi, 2015; Marcus, 2000), and see emotions as liabilities to political rationality and to the rational discourse (Habermas, 1987). Our argument is that this way of thinking is embedded in

how emotion technology is designed. This does not overlook that these technologies show traces of Cartesian traditions or cognitivist and reductionist approaches too (Gates, 2011), but we mainly focus on the influence of political narratives on technology so that we can establish a new critique on emotion technology and its political aura.

We approach a critique on emotion technology and its political narratives in two ways: First, we discuss how and why emotions became an interesting topic for technology and for software design, how digital applications shape and steer political moods even though recognition systems are not as accurate as expected. The first part will mainly pay attention to the design of emotion technology, to its political power like on manipulation, and to it being a lucrative business (Bösel and Wiemer, 2020; Zuboff, 2019). This part revisits known concerns around emotion technology and its limits, which brought together, act as a stepping stone for the second part. Second, we then link emotion technology, with all its shortcomings, to what we call the liberal narrative. Here, we debate the political lineage of political narratives on emotions and state that emotion technology is the effect and result of a liberal political tradition that problematizes emotions (Marcus, 2000; Miller, 2011; Habermas, 1987). Seeing this as a negative approach to emotions, we point to consequences and to an alternative that allows to break it up. Our goal is not to advertise an unconditional value of emotions as good forces, but to present concerns about the liberal influence as a way to devalue emotions while being an authority and hegemony over the normativity of emotions. This seems like a political intervention and while it is, our contribution is on how this way of thinking shapes the design of applications supposed to measure and control emotions, namely, *emotion or affect technology*. Then we go one step further: By including the antagonistic political work of Chantal Mouffe (1993, 2000, 2005), we open up towards alternatives in narrating the power of emotions that are not embedded in present technology. Our critique on emotion technology is one on narrative, but we will show that this is not the only narrative to follow. Hence, challenging the narrative, should challenge what technology we want to design.

We are motivated to equally show that the liberal narrative ignores a positive value of emotions as forms of resistance or to broaden access to discourse in democracy. Hence, pushing for new narratives on emotion as a critique on how emotion technology occupies the discussion on emotion, is not a matter of *nice to have*, but about political and technological hegemony. Acknowledging social movements like *Black Lives Matter* or *Friday For Future*, we see that politicized emotions or passions, as Mouffe would say, are not simply sidetracked rational discourses, but lead to legitimate upheavals against power structures and discrimination. Hence, we need to be careful in what narratives and biases we support through technology (Noble, 2019) or if we want the discourse around emotions to be on them being steered as physiological triggers or being liabilities. By employing two opposing political theories on emotions (liberal and antagonistic) within a critique on emotion technology, we employ political theory as one way to inform techno-philosophy and vice versa (Coeckelbergh, 2022).

2 Emotion technology: On cognitivist reductionism, affect industries and computational politics

Emotion technology, also referred to as affect technology, emerged in the 90's as a result of cognitive sciences and programming unifying their ambition to make emotional technology (Picard, 1997; Freed, 2020). Emotion technologies differ from each other but we address one critique they have in common. These applications are as reductionist and inaccurate (Pennebaker et al., 2015) as they are powerful (Stark, 2018). Let us unpack this briefly. The 'turn to emotion in AI' (Wilson, 2010), also known as the cognitivist turn to emotions, changes how science and computation deals with emotions; shifting away from experience towards emotion as physiological or cognitive (Freed, 2020; Barrett et al., 2019; Wilson, 2010; Bösel and Wiemer, 2020). McStay (2018) writes that this implied that machines can count our 'heart rate, body temperature, respiration and the electrical properties of our skin, among other bodily behaviours' (3) and that now, 'bodies and emotions have become machine-readable' (3). Yet, this does not imply that technology can 'experience emotions' (3, emphasis in original). With methods on measuring emotions advancing, new possibilities emerge in how to extract information and how to guide affective behavior. This relates to recognition software but also to affective architectures and nudging software.

Emotion *recognition* technology (as one branch of emotion technology) aims to scale or read emotions, but with partially successful results (McStay, 2018; Stark, 2018; Telford, 2019; Crawford, 2021). At the same time, there are certain affect triggering architectures that raise lots of critique due to being politically and economically powerful and potent. Let us first pay attention to the idea of measuring or scaling emotions. We face new 'ways of algorithmically filtering content' that manifest a rationalistic view on emotion as detectable signs and not as culturally or psychologically diverse or fluid practices. In fact, we do not even speak of measuring *emotions* as much as we speak of tracing physiological markers that are said to be emotional expressions (Freed, 2020), since recognition software cannot *read* emotions. It tracks and detects visual or textual cues or patterns. This is highly culturally biased as Gates (2011) points out in saying that 'humans see the faces of others change, of necessity, along with changing cultural practices, social conventions, and forms of social and technological organization' (11). Interestingly, this critique is growing as rapidly as their value as an industry. The latter being a major concerns around this topic. Thinking about the profitable side, then these technologies are still lucrative, despite not being necessarily accurate or ethical. For researchers like Zuboff (2019), the main purpose of affect technologies, like Picard's *Affectiva*, is for economic gain and to implement new behavioral triggers to manage public moods. Even if Microsoft claims that is emotion technology *Deep Empathy* can detect emotion and 'interpret' what individuals are feeling from reading video images of a person's face (Barrett et al., 2019), this is more of a prolific claim than the reality.

Despite its technical shortcomings, emotion technology seems politically powerful in fostering manipulation, making nudging invisible, or in keeping informational bubbles segregated. This has transformed political campaigning radically (Anstead,

2017; Tufekci, 2014). Hence, the value of affective technologies is equally political as it is lucrative even if this refers more to how algorithmic architectures steer public debates rather than to emotion recognition software tracing individual expressions. For instance, Nielson (2012) states that because of computational politics, we face a new 'ground war' in engaging directly and more efficiently with the electorate that can be easier mobilized and manipulated. A central concern here is that emotion software opens doors to easier manipulation of affective behavior. There are various tools to foster this concern; micro-targeting (Zuiderveen Borgesius et al., 2018: 83), testing (Baldwin-Philippi, 2017: 628) or nudging (Thaler and Sunstein, 2008; Sætra, 2019) – which are different forms of behavioral design, but each is influential while not transparent to those influenced by it. Despite the fact that nudging is not necessarily linked to the exploiting emotions triggers, but rather to multiplying cognitive bias (Sætra, 2019). Next to the question if this new software is mobilizing emotions in particular, we can say that algorithmic technologies raise questions on citizen autonomy overall and throughout. This has consequences for campaigning and for democratic participation beside how we define empowerment in discourse, or the lack thereof. We already mentioned Rhodes' (2022) concern over filter bubbles as a way to keep political opinions in homogeneous spheres (which is not a new phenomena), or Levendusky's concern (2013) about how citizens are polarize via algorithmic architectures keeping them on an 'informational diet'. In this context, we also see the growing power of AI as concerning, which is pointed out by Coeckelbergh (2021), Sudmann (2019) and Howard (2020). Its rise can have negative effects by making it more possible to 'charm and provoke voters, spread misinformation, and take advantage of rich personal data to make powerful emotional appeals. For countries that hold elections, AI-driven fake users will become a new, serious, and internal threat to their democracy' (Howard, 2020: 146). While the political value of mobilizing emotion via technological means cannot be underestimated, new concerns are situated within the overlap of technogloical capacities, economic gain and political goals. We leave this discussion for now. Next, we will unpack a different angle from political theory to explore the narrative of emotions we see embedded in technology.

3 Tracing the liberal narrative in emotion technology

What does it mean to follow a rationalistic view on emotions and how does this tell a liberal narrative? Let us sum up what we discussed in the previous section. First, we said that emotion technology is reductive and inaccurate to identify emotions as states, and yet, some applications trace and shape political moods and sentiments increasingly well. Keeping this paradox in mind, we cannot resolve it, but it also does not conflict with our next discussion on how emotion technology extends a liberal narrative since we are more concerned with the factors that go into the making of technology than its accuracy. Besides the addressed shortcomings we revisited so far, our argument is therefore that these kind of technologies support or extend a liberal narrative, which means that not only do they classify emotions to signs and causalities

and as steerable moods, but they favor a rational and predictive behavior that tries to bring emotions under control, not to empower citizens. Our main point is that emotion technology is designed for a problem inherently framed from a liberal angle on emotions. This is not representative for the political value of emotions per se, but it has become a dominant view for how science, politics and ethics discuss emotions as liabilities or uncontrollable (and as distinct from reason).

Political emotions are historically devalued as an irrational driver behind decisions, as easy to manipulate because out of affect, and anti-dotes to reason (Papacharissi, 2015, Mouffe, 2000). In political terms, the aim to control emotions is not a new approach, but in technical terms, it is newly expressed. The idea of controlling emotions can be traced back to how emotions are defined; as irrational, as unrulable, as unstable and manipulative. For instance, Habermasian or Rawlsian 'liberal theories are dominated by an individualistic, universalistic, and rationalistic framework that assumes a pre-given, pre-political rational subjectivity' (Morrison, 2017: 534). While Rawls (1996) considers a minimum level of rationality 'and a minimum level of morality as natural and non- or pre-political rather than 'as articulating particular political values' (Biesta, 2011: 143), Habermas (1987) sees the value of universalist forms of morality and law as expressions of a collective process of learning. Overall, the value of emotions is mostly limited to nationalist sentiment and is considered as pre-discursive by Habermas (1987). While there is a theoretical lineage in emotions being politically discredited resulting in a democracy of feeling (Davies, 2018), we see traditionally how emotional and subjective assessments and evaluations are often described as clouding the truth (Boler and Davis, 2018), which implies a 'willful blindness to evidence, a mistrust of authority, and an appeal to emotionally based arguments often rooted in fears or anxieties' (Laybats and Tredinnick, 2016: 204).

However, we might face (at least) two political trends intersected within emotion technology. First, affect and emotion technology might be able to mobilize moods, anger or sentiments of frustration by making them *irrational*, but second, affect technology does not fabricate the reasons for these emotions to emerge *as justified*. Emotions are embedded in a socioeconomic and political framework that tends to be overlooked when fearing *how very manipulative* emotion technology has become (Abreu and Öner, 2020). At the same time, it is valid to fear how anger and rage are amplified or steered through technologies and exploited by industry. For instance, Bösel (2020) discusses how the 2016 US presidential election and the Brexit vote were used as strategic approaches to mobilize voters' and 'instrumentalized for the advancement of authoritarian and populist policies' (10). Overall, we need to be careful what conclusion we draw from these two trends, one on emotion technology not producing unjustified emotions, and the other, emotion technology *still* exploiting emotional and affective dynamics. Both embed political power and traditions that can be challenged or at least, outlined.

Concluding, we do not place *all the blame* on the liberal narrative as the main driver behind emotion technology, but we can surely consider that there is a concerning lineage between controlling, rationalizing and devaluating. And moreover, this is not a new one. This thinking represents a wider normative hegemony and narrative of how '(p)hilosophy, political theory, and common sense tend to view

emotion and reason as two opposite forces that must somehow be reconciled so that people can function as informed citizens' (Papacharissi, 2015: 10). Now, if we argue that political narrative matters for the norming, not only applying, of technology, then we can go further and say that classification, political and economic value of emotion technology fits into a liberal narrative to control something quite powerful like emotions. Doing this, we then need to think of emotion technology employed by 'Google, Facebook, and the other tech giants themselves' as expanding on politically liberal trends by also fostering 'serious threats to democracy and/or acting in contrast to democratic values, in terms of their business strategies, data practices, and enormous economic and socio-cultural power' (Sudmann, 2019: 10). This leads to more worries; not only should worry about emotions as exploited through technology in hidden, unregulated ways, but also about how technology exploits emotions by extending *liberal imaginaries of control*. To be clear, we do not oppose liberal ideas as such. We do not argue against liberal societies, but as we mentioned, we have picked a specific moment within liberal thinking that problematizes emotions. Problematizing emotions has established new ways to control emotions, new hegemonies, which are spaces for more control and profit, *and this* where we see the problem for democracy (Nemitz, 2018). Let us shift our attention to another perspective that reflects critically on the liberal narrative. To counter the liberal narrative we portrayed so far, we move on to alternatives which have dealt with the limits of liberal thinking of emotions, not with emotion technology. Our aim is to show a way in which we can approach the value of emotions differently to then urge for new technologies of empowerment and a re-evaluation of emotions as political forces. We saw that the value of political emotions is entangled with that of emotion technology in and for politics, let us therefore look at why emotions can also matter as empowering forces.

4 The antagonistic view on emotions and why changing the political narrative matters

What should have emerged from the previous discussions is that emotions are problematized in political narratives and standardized via cognitive measures into technological design; all influencing each other. What might not be clear yet is that there are different ways to norm emotions politically and maybe also technically. While emotion technology assigns itself to the measuring tradition of cognitivist programming, the work of Mouffe (1993, 2000, 2005) as an antagonistic approach to political emotions offers us an empowering view on emotions (or passions). Having said that the value of political emotions is entangled within emotion technology, let us untangle that by leaving aside technology and focus on an empowering and constructive narrative on emotions. Mouffe's work urges us to move beyond the liberal narrative by taking into account that emotions are empowering forces, which could make us resist the techno-liberal narrative. We clarify again that it is not about us *thinking positive* and advertising a 'rule of emotions', it is about bringing together emotional narratives and political power to challenge new and problematic

technological applications. This allows for two discussions to unite: First, to present alternatives saying that emotions are powerful forces *within* democracy and discourse not outside of them. Second, to establish a critique on emotion technology as a controlling, but not empowering technology and to demand for new technologies that can foster better narratives.

Presenting an empowering narrative on emotions, we turn to Mouffe as one of the major voices of an antagonistic model of democracy. In her work, she emphasizes the value of passions and dissensus (or conflict) arguing that collective passions are 'moving forces of human conduct' (Mouffe, 1993: 140). Still, we have to make sure this is not misunderstood: Mouffe is equally concerned about a rise of undemocratic or right-wing surges on the grounds of political mobilization of emotions (2005). We do not employ Mouffe's work to say; *the more emotions, the better*. But we question the power narrative in trying to control emotions or to deny their political legitimacy. While she does not engage with emotion technology, she critically responds to the liberal narrative. Mouffe does not endorse a discursive chaos or a rule of emotions, even if she is not very clear about how to unite and organize collective passions once conflict arises, or what kind of conflict to allow and what kind of conflict to condemn. However, what she is clear about, is that liberal values and norms are hegemonic and exclusive and those in power decide whom to let into the discursive space via problematizing emotions/passions. This is a crucial critique we share in understanding that the value of emotions can be strategically devalued to dismantle a legitimate social or political critique through labelling those who raise it as too emotional. For Mouffe, the aggregative and deliberate political models proposed by Habermas and others are not sufficiently incorporating that the political has to be inherently antagonistic (2005, 2016). She hereby distinguished between what is *political* and what is *politics*. The *political* is the ontological space in which antagonism unfolds while *politics* refers to the practices and institutions that operate and manage the democratic sphere (2005: 8,9). What she critiques is that there is no *real* antagonistic space given by the liberal center to those who want to protest its power. Instead, undemocratic (and for her, also right-wing) forces are mobilizing emotions in problematic ways by using this very human need to express anger or fear, but through pseudo-conflicts that end up manifesting unequal systems. Real antagonism, for her, is necessary to counter these systems, or what she defines as 'hegemonic practices' (2016, 1) which are 'practices of articulation through which a given order is created, and the meaning of social institutions fixed therein' (1). Tambakaki (2014) sums up Mouffe's approach to fight the liberal hegemony as one that encourages 'acts of 'doing' to construct counter-hegemonic projects' and, therefore, 'of being agonistic by virtue of such acts of construction'. This is why passions are 'a way of identifying (with liberty and equality) that is affective enough to constitute political subjects and strong enough to unite, ...' (6). Mouffe is not alone with this critique. For instance, Martha Nussbaum (2013) discusses the role of emotion as part of liberal societies and from a philosophical view. For Nussbaum, emotions are as much a positive force and a crucial pile of societal justice, but these have to be cultivated politically within liberal society (3). Having pointed out that there are issues in how the liberal narrative

devalues emotions, we will integrate this discussion into a critique on emotion technology as a control mechanism instead of an empowering one.

5 Conclusion and summary

The *liberal narrative* in emotion technology has various consequences for how we critique emotion technology and allows to identify how political theory or normativity *shapes* technology and emotions. We will discuss two consequences of emotion technology we see emerging through the lens of Mouffe's work; first, in thinking dissensus; and second, in understanding the *taming* versus the *controlling* of emotions. These two allow us to elaborate short coming with emotion technology and their liberal influence.

First, Mouffe endorses the role of passions in dissensus as political forces which is not compatible with emotion technology as a hegemony of cognitivist classification, political aims, and economic value. Since, it is not only about enabling conflict, but about *how* conflict is enabled, by *whom*, and through what form of *controlled* space. Allowing for a *good* conflict to emerge means to allow for a space in which resistance (in whatever form) can be expressed and counter narrative to discrimination can exist, but the implementation of emotion technology does the opposite: It is a controlled, not a *political* space. This means it controls the norms and standards and the economic value of emotions without asking for consent and without much regulation. This creates a psychometric loop of cognitivist ambitions, economic interests, and emotion narratives which is political too, but not empowering for citizens. Enabling real antagonism as Mouffe suggests, would mean to allow for resistance and a counter hegemony to emerge on the system in place. In this case, the *political* should offer a participatory space for passions to be expressed, but not in form of access to platform communication or to social media, but in the sense of being able to escape the calculation and control of one's moods via hidden means. It is not about joining the communicative space to shape political conflict as noise, but about escaping the psychometric hegemony via resistance.

Second, Mouffe sees political power in passions or emotions, but also in taming these. Would this not even fit into what emotion technology does? In both cases, conflict is endorsed and emerges. However, there is a major difference between the liberal rationality in technology and Mouffe's approach. Emotion technology does not *tame* emotions in the sense of Mouffe, but it rationalizes emotions through a cognitive, and eventually, liberal tradition, which we argued opens doors to manipulation and to industry, not to the empowerment of citizens. These steps are charged with political value and embed capitalist and classifying structures into democratic spaces. In both cases, there could be some agreement over saying that emotions are ok, but we need to manage these somehow to avoid political chaos. While the regulation and rationalization of emotions through technology follows a liberal fear of losing control over too emotional citizens that ironically and equally shows in the ambition to manipulate these, Mouffe advertises for emotions to be *tamed* differently. She promotes the possibility of an antagonistic and empowering

space to avoid chaos, not to measure and control. The difference being that the first feeds a liberal hegemony and the other urges for a counter hegemony thereof. Hence, the first is about control, the second about avoiding chaos and enabling a space for dissensus to foster counter hegemonies. Considering there are different consequences from taming versus rationalizing emotions, the latter is highly problematic since emotion technology is misused by a few powerful and unregulated industries while increasing cultural bias and reductionism, it also does not legitimize movements like Black Lives Matter or Fridays for Future. A rationalistic approach to emotions not see much value in political rage or frustration. While this seems one-dimensional considering the critique of this liberal and somewhat Cartesian thinking of discourse, it also overlooks that certain social movements are not the result nor distortion of reason or affect of manipulation, but reasonable, organized and justified responses to systematic discrimination or ignorance. However, how to organize resistance and where, fell short in this paper, but our optimism is shaky on the question if we can escape the computational hegemony of psychometric scaling and observation to which emotion technology seems to be only one tentacle.

To conclude our exploration, while not placing too much attention to how emotions are steered politically, we did not underestimate that emotions can be exploited for political manipulation or reactionary conflict. Instead, we preferred to emphasize another angle; how scientific norms, political theories, algorithmic architectures and economic interests are intersected in such a way to discredit the value of emotions as empowering forces. Hence, we paid more attention to normative lineages and less to the economic or manipulative power of emotion technology. While we were brief on, we point to the crucial value of research on manipulation, nudging, information bubbles or algorithmic power. Despite being rightly concerned about this trend, the mobilization of emotion, for the sake of polarizing society, is not new and it was not our main focus. We see that the manipulation of emotions (or what counts as such) is less about tracing the weak spots of individual emotionality or rationality, but about nudging or micro-targeting or persuasive design architectures. In addition to these issues, the coding and reducing of emotions to signs seemed to have fostered a new industry that exists unregulated and hidden. Neither endorsing a 'democracy of feeling' nor a Cartesian revival of reason, we highlighted that the wider strategy to problematize emotions or affects is met by the goal to measure and regulate these, and that this fits into a wider control narrative that liberal theories provided indirectly. Our task was to highlight that *liberal narrative* in emotion technology and its issues. One of which would that it becomes a dominant narrative of how we frame emotions politically and to which technology then seems a solution to set-up problem. However, emotion technology might only be an extension of the very political hegemony it serves, and this matters as much as its other short comings. Ultimately, we urge for more normative diversity and empowering narratives regarding emotions in politics, and in how we translate political power into technology. Would this mean to have *less* or just *different* technologies? In any case, we need to scrutinize the political technologies and institutions at place so that neither can manipulate nor ban emotions from discourse, but enables constructive and inclusive angles on democratic and political thinking instead.

References

- Abreu, M. and Öner, Ö. (2020). Disentangling the Brexit vote: The role of economic, social and cultural contexts in explaining the UK's EU referendum vote. *Environment and Planning A: Economy and Space*, 52(7), p.0308518X2091075. doi:10.1177/0308518x20910752.
- Anstead, N. (2017). Data-Driven Campaigning in the 2015 United Kingdom General Election. *The International Journal of Press/Politics*, 22(3), pp.294–313.
- Baldwin-Philippi, J. (2017). The Myths of Data-Driven Campaigning. *Political Communication*, 34(4).
- Barrett, L.F., Adolphs, R., Marsella, S., Martinez, A.M. and Pollak, S.D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest*, 20(1), pp.1–68. doi:10.1177/1529100619832930.
- Biesta, G. (2011). The Ignorant Citizen: Mouffe, Rancière, and the Subject of Democratic Education. *Studies in Philosophy and Education*, 30(2), pp.141–153. doi:10.1007/s11217-011-9220-4.
- Boler, M. and Davis, E. (2018). The affective politics of the 'post-truth' era: Feeling rules and networked subjectivity. *Emotion, Space and Society*, 27, pp.75–85. doi:10.1016/j.emospa.2018.03.002.
- Bösel, B. (2020). Affective Transformations: An Introduction. In: B. Bösel and S. Wiemer, eds., *Affective Transformations: Politics—Algorithms—Media*. Lüneburg: Meson Press.
- Bösel, B. and Wiemer, S. (2020). *Affective Transformations: Politics - Algorithms - Media*. Lüneburg: Lüneburg Meson Press.
- Chakriswaran, P., Vincent, D.R., Srinivasan, K., Sharma, V., Chang, C.-Y. and Reina, D.G. (2019). Emotion AI-Driven Sentiment Analysis: A Survey, Future Research Directions, and Open Issues. *Applied Sciences*, 9(24), pp.1–27. doi:10.3390/app9245462.
- Coeckelbergh, M. (2020). *AI Ethics*. Cambridge, Massachusetts: The MIT Press.
- Coeckelbergh, M. (2022). *The Political Philosophy of AI*. Cambridge: Polity Press.
- Crawford, K. (2021). *Atlas Of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.
- Davies, W. (2018). *Nervous States: How Feeling Took Over the World*. London: Jonathan Cape.
- Freed, S. (2020). *AI and Human Thought and Emotion*. Boca Raton: CRC Press, Taylor & Francis Group.

- Gates, K. (2011). *Our Biometric Future: Facial Recognition Technology and the Culture of Surveillance*. New York: New York University Press.
- Habermas, J. (1987). *Theorie des kommunikativen Handelns*. Frankfurt A.M.: Suhrkamp.
- Howard, P.N. (2020). *Lie Machines: How to Save Democracy From Troll Armies, Deceitful Robots, Junk News Operations, and Political Operatives*. New Haven: Yale University Press.
- König, P.D. and Wenzelburger, G. (2020). Opportunity for renewal or disruptive force? How artificial intelligence alters democratic politics. *Government Information Quarterly*, 37(3), pp.1–11. doi:10.1016/j.giq.2020.101489.
- Laybats, C. and Tredinnick, L. (2016). Post truth, information, and emotion. *Business Information Review*, 33(4), pp.204–206. doi:10.1177/0266382116680741.
- Levendusky, M.S. (2013). Why Do Partisan Media Polarize Viewers? *American Journal of Political Science*, 57(3), pp.611–623. doi:10.1111/ajps.12008.
- Marcus, G.E. (2000). Emotions in Politics. *Annual Review of Political Science*, 3(1), pp.221–250. doi:10.1146/annurev.polisci.3.1.221.
- McStay, A. (2018). *Emotional AI: The Rise of Empathic Media*. London: Sage Publications.
- Miller, P.R. (2011). The Emotional Citizen: Emotion as a Function of Political Sophistication. *Political Psychology*, 32(4), pp.575–600. doi:10.1111/j.1467-9221.2011.00824.x.
- Morrison, A. (2017). Rescuing politics from liberalism: Butler and Mouffe on affectivity and the place of ethics. *Philosophy & Social Criticism*, 44(5), pp.528–549. doi:10.1177/0191453717730875.
- Mouffe, C. (1993). *Return Of The Political*. London: Verso Books.
- Mouffe, C. (2000). *The Democratic Paradox*. London: Verso.
- Mouffe, C. (2005). *On the Political*. London: Routledge.
- Mouffe, C. (2016). Democratic Politics and Conflict: An Agonistic Approach. *Política Común*, 9, pp.1–8. doi:10.3998/pc.12322227.0009.011.
- Nemitz, P.F. (2018). Constitutional Democracy and Technology in the Age of Artificial Intelligence. *Phil. Trans. R. Soc.*, 376(20180089), pp.1–14. doi:10.2139/ssrn.3234336.
- Nielsen, R.K. (2012). *Ground Wars: Personalized Communication in Political Campaigns*. Princeton University Press.
- Noble, S.U. (2018). *Algorithms of Oppression How Search Engines Reinforce Racism*. New York: NYU Press.
- Nussbaum, M.C. (2013). *Political Emotions: Why Love Matters For Justice*. Cambridge: Harvard University Press.

- Papacharissi, Z. (2015). *Affective Publics: Sentiment, Technology, and Politics*. Oxford: Oxford University Press.
- Pennebaker, J.W., Boyd, R.L., Jordan, K. and Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.
- Picard, R. (1997). *Affective Computing*. Cambridge: MIT Press.
- Rawls, J. (1996). *Political Liberalism*. New York: Columbia University Press.
- Rhodes, S.C. (2022). Filter Bubbles, Echo Chambers, and Fake News: How Social Media Conditions Individuals to Be Less Critical of Political Misinformation. *Political Communication*, 39(1), pp.1–22. doi:10.1080/10584609.2021.1910887.
- Sætra, H.S. (2019). When nudge comes to shove: Liberty and nudging in the era of big data. *Technology in Society*, 59, p.101130. doi:10.1016/j.techsoc.2019.04.006.
- Stark, L. (2018). Algorithmic psychometrics and the scalable subject. *Social Studies of Science*, 48(2), pp.204–231. doi:10.1177/0306312718772094.
- Sudmann, A. (2019). *The Democratization of Artificial Intelligence: Net Politics in the Era of Learning Algorithms*. Bielefeld: Transcript.
- Tambakaki, P. (2014). The Tasks of Agonism and Agonism to the Task: Introducing 'Chantal Mouffe: Agonism and the Politics of Passion'. *Parallax*, 20(2), pp.1–13. doi:10.1080/13534645.2014.896543.
- Telford, T. (2019). 'Emotion detection' AI is a \$20 billion industry. *New research says it can't do what it claims*. [online] Washington Post. Available at: <https://www.washingtonpost.com/business/2019/07/31/emotion-detection-ai-is-billion-industry-new-research-says-it-cant-do-what-it-claims/>.
- Thaler, R.H. and Sunstein, C.R. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven: Yale University Press.
- Tufekci, Z. (2014). Engineering the Public: Big data, Surveillance and Computational Politics. *First Monday*, 19(7). doi:10.5210/fm.v19i7.4901.
- Wilson, E.A. (2010). *Affect and Artificial Intelligence*. Seattle: University Of Washington Press.
- Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs.
- Zuiderveen Borgesius, F.J., Möller, J., Kruijemeier, S., Ó Fathaigh, R., Irion, K., Dobber, T., Bodo, B. and De Vreese, C. (2018). Online Political Microtargeting: Promises and Threats for Democracy. *Utrecht Law Review*, [online] 14(1), pp.82–96. doi:10.18352/ulr.420.

Extended Abstracts and Short Papers

Water Use as Remote Monitoring Technology: An Ethical Analysis (extended abstract)

Tania Moerenhout¹[MD, PhD] and Katleen Gabriels²[PhD]

¹ Bioethics Center, University of Otago, Dunedin, Aotearoa New Zealand

² Department of Philosophy, Maastricht University, Netherlands
k.gabriels@maastrichtuniversity.nl

Abstract. Remote monitoring technology ensures that we can remain connected with our intimates from a distance. In this way, we can keep an eye on, for example, our parent who needs care but who prefers to age in place. This extended abstract focuses on an ethical analysis of remote monitoring via water use, a seemingly harmless and everyday activity. Nevertheless, water use reveals a lot of information about a person's lifestyle, for example, cooking, washing, and going to the toilet. Although remote monitoring via water use is less intrusive, it raises compelling ethical questions. What if the data show that someone is not washing themselves enough (hygiene) or is using too much water (sustainability)? Can we know 'too much'? Overall, this extended abstract has two guiding concerns: What are the ethical implications or risks of monitoring through the use of water? And, more generally speaking, how can we conduct an ethical technology assessment that succeeds in identifying ethical tensions?

Keywords: Remote monitoring technology, Water monitoring, Technology assessment, eTA, eCTA

1 Introduction

Today, global healthcare is confronted with the challenges of a rapidly ageing population – not necessarily healthier – who live longer with chronic illness and disabilities. Most older adults, even when facing declining health and increasing impairment, prefer to age in place, if possible (van Hoof et al., 2011; den Ouden et al., 2021). Despite culture change in nursing homes taking place over the last two to three decades, most nursing homes are still institutional hospital-like environments, with smaller, person-centred homes being the exception rather than the norm (Zimmerman, Shier, & Saliba, 2014).

The Covid-19 pandemic has exposed the vulnerabilities of this traditional nursing home system once more, in a most harrowing way. Not only was the virus in itself particularly deadly among older adults, the architecture and organisational structure of nursing homes, and the lack of a proper management plan in the first months of the outbreak all contributed to the dramatic mortality numbers (Moerenhout, 2020).

These current events may contribute to an increase in demand for ageing in place. Smart homes are expected to provide the solution to the question how older adults

facing chronic illness and impairment can continue living in their own home. Wearables, apps, sensors, cameras, and other devices are integrated through the Internet of Things (IoT) “to offer continuous, objective, and holistic monitoring, alleviating the burden of human caregiver effort and supporting clinical decision making” (Stavropoulos et al., 2020, web; see also Offermann-van Heek & Ziefle, 2019). But there are also concerns that these remote monitoring technologies may negatively affect trust, privacy, autonomy, and self-confidence, and could alter older adults’ perception of home (Ben Mortenson, Sixsmith, & Beringer, 2016; Birchley et al., 2017).

Most of these barriers or objections are related to the invasiveness of this type of technology. So, what if we could offer non-invasive monitoring that is to a large extent invisible, or moves to the background (Van Den Eede, 2011)?

2 Case: Water monitoring technology

2.1 Introduction of case

Interestingly, some developers, in their search for monitoring technology that is cost effective, fades into the background, does not require frequent interaction, and does not store sensitive information, have turned their focus to water usage (Tsukiyama, 2015). An early example is that of the Japanese i-pot, tracking the use of an electric kettle to keep an eye on older adults living alone (Doi, 2005). If the kettle is not used, this can point at an alarming situation and can be a reason to reach out. The kettle is easy to use, low cost, non-invasive, and embedded in daily routines.

The second example will be central in our final paper and presentation. It proposes a sensor-based monitoring system aimed at tracking the use of tap water (Tsukiyama, 2015). Through water usage, a lot of information about daily activities is revealed: drinking, cooking, cleaning, toilet use, washing, and so on. The regular use of water, often at set times (e.g., mealtime), can provide insight into the health of older people. For example, frequent trips to the toilet at night can point at a health problem.

2.2 Objectives and aims

Could this type of unobtrusive health monitoring solve the ethical challenges that usually accompany tracking technology? Our final paper will seek to answer two questions:

1. What are the ethical implications or risks of monitoring through the use of water?
2. More generally speaking, how can we conduct an ethical technology assessment that succeeds in identifying ethical tensions?

We are obviously not the first ones to consider the latter question. There is a vast literature on ethical technology assessment (Palm & Hansson, 2006; Wright, 2010; Kiran, Oudshoorn, & Verbeek, 2015; van de Poel, 2016). However, the ‘how’ question remains a challenge. Technology development and ethical assessment take place in largely different worlds that do not often interact with each other, even speak the same language. Moreover, different ethical frameworks for technology assessment (TA) that

have been developed on the theoretical side have not always found their way to practical application and testing in real-life circumstances.

3 Case Analysis: eTA and beyond

In the presentation, we will first apply the existing theoretical framework of ethical technology assessment (eTA) to the specific case of water monitoring technology. We will start by using the checklist as laid out by Palm and Hansson, which “refers to nine crucial ethical aspects of technology; (1) Dissemination and use of information, (2) Control, influence and power, (3) Impact on social contact patterns, (4) Privacy, (5) Sustainability, (6) Human reproduction, (7) Gender, minorities and justice, (8) International relations, and (9) Impact on human values” (Palm & Hansson, 2006, p. 543).

For example, when we analyse water monitoring’s impact on social contact patterns (cf. aspect#3), a complex effect emerges. As Palm and Hansson (p. 552) state: “There is a tendency for electronically mediated contacts to substitute face-to-face contacts”. This means that once the sensors are installed and an automated system is in place to warn caretakers when something looks to be out of place, frequent visits or phone calls to check in on the older person may decline. On the other hand, such a system could also alleviate care burdens and worries, and by doing so create a space for more meaningful interactions.

When comparing the use of a water monitoring system with other forms of remote monitoring (e.g., cameras), the former seems to perform better on the privacy aspect (cf. aspect#4) as such system is less intrusive and invasive. However, the information that is being collected is very intimate information exposing personal hygiene routines, toilet use, and cooking habits. One may wonder which conclusions are drawn when someone does not often take a shower or does not regularly use the washing machine? To what extent is personal freedom (cf. aspect#9) respected in this regard?

The checklist allows us to identify ethical challenges of water monitoring, but there are several shortcomings to using a checklist-based system. In order to address these, we will extend our analysis to other forms of TA. In our final paper and presentation, we will include insights from an ethical-constructive TA approach (eCTA), and from an anticipatory technology ethics (ATE) perspective to complete our analysis of water monitoring (Brey, 2012; Kiran, Oudshoorn, & Verbeek, 2015). For instance, to move away from a set of ‘given’ ethical principles, eCTA is a more dynamic framework, that also includes questions concerning how much room there is to tinker the technology.

4 Conclusion

Overall, the final paper and presentation will present an eTA of water monitoring systems that goes beyond the checklist-based approach, providing an in-depth analysis of ethical issues that need to be addressed when using this form of technology in ageing in place. This will also bring us to a conceptual reflection on eTA and a meta-perspective on the application of current eTA models.

References

- Ben Mortenson, W., Sixsmith, A., & Beringer, R. (2016). No place like home? Surveillance and what home means in old age. *Canadian Journal on Aging*, 35(1), pp. 103–114. doi: 10.1017/S0714980815000549.
- Birchley, G., Huxtable, R., Murtagh, M., Ter Meulen, R., Flach, P., & Goberman-Hill, R. (2017). Smart homes, private homes? An empirical study of technology researchers' perceptions of ethical issues in developing smart-home health technologies. *BMC Medical Ethics*, 18(23). <https://doi.org/10.1186/s12910-017-0183-z>
- Brey, P. A. E. (2012). Anticipatory ethics for emerging technologies. *Nanoethics*, 6, pp. 1–13. <https://doi.org/10.1007/s11569-012-0141-7>
- den Ouden, W. V., van Boekel, L., Janssen, M., Leenders, R., & Luijkx, K. (2021). The impact of social network change and health decline: a qualitative study on experiences of older adults who are ageing in place. *BMC Geriatrics*, 21(480). <https://doi.org/10.1186/s12877-021-02385-6>
- Doi, E. (2005, 11 April). *Technology in a teapot keeps watch on elderly*. *The Seattle Times*. Retrieved from <https://www.seattletimes.com/business/technology-in-a-teapot-keeps-watch-on-elderly/>
- Kiran, A. H., Oudshoorn, N., & Verbeek, P.P. (2015). Beyond checklists: toward an ethical-constructive technology assessment. *Journal of Responsible Innovation*, 2(1), pp. 5–19. <https://doi.org/10.1080/23299460.2014.992769>
- Moerenhout, T. (2020, 11 September). *The problem in nursing homes is not Covid-19 – it is nursing homes*. Journal of Medical Ethics blog. Retrieved from <https://blogs.bmj.com/medical-ethics/2020/09/11/the-problem-in-nursing-homes-is-not-covid-19-it-is-nursing-homes/>
- Offermann-van Heek, J., & Ziefle, M. (2019). Nothing else matters! Trade-offs between perceived benefits and barriers of AAL technology usage. *Frontiers in Public Health*, 7, pp. 1–16. <https://doi.org/10.3389/fpubh.2019.00134>
- Palm, E., & Hansson, S. O. (2006). The case for ethical technology assessment (eTA). *Technological Forecasting and Social Change*, 73(5), pp. 543–558. <https://doi.org/10.1016/j.techfore.2005.06.002>
- Stavropoulos, T.G., Papastergiou, A., Mpaltadoros, L., Nikolopoulos, S., & Kompatsiaris, I. (2020). IoT wearable sensors and devices in elderly care: A literature review. *Sensors*, 20(10). <https://doi.org/10.3390/s20102826>
- Tsukiyama, T. (2015). In-home health monitoring system for solitary elderly. *Procedia Computer Science*, 63, pp. 229–235. <https://doi.org/10.1016/j.procs.2015.08.338>
- Van Den Eede, Y. (2011). In between us. On the transparency and opacity of technological mediation. *Foundations of Science*, 16(2), pp. 139–159. <https://doi.org/10.1007/s10699-010-9190-y>
- van de Poel I. (2016). An ethical framework for evaluating experimental technology. *Science and Engineering Ethics*, 22, pp. 667–686. <https://doi.org/10.1007/s11948-015-9724-3>

- van Hoof, J., Kort, H. S. M., Rutten, P. G. S., & Duijnste, M. S. H. (2011). Ageing-in-place with the use of ambient intelligence technology: Perspectives of older users. *International Journal of Medical Informatics*, 80(5), pp. 310–331. <https://doi.org/10.1016/j.ijmedinf.2011.02.010>.
- Wright, D. (2010). A framework for the ethical impact assessment of information technology. *Ethics and Information Technology*, 13, pp. 199–226. <https://doi.org/10.1007/s10676-010-9242-6>
- Zimmerman, S., Shier, V., & Saliba, D. (2014). Transforming nursing home culture: Evidence for practice and policy. *Gerontologist*, 54. <https://doi.org/10.1093/geront/gnt161>.

Sustainability and Technology: Digital Marketplaces in Voluntary Market Offsetting

Rafael Canorea-García¹ [0000-0002-6637-4369] and Mario Arias-Oliva² [0000-0002-6874-4036]

¹Pegaso International University, Madrid, Spain

²Complutense University of Madrid, Madrid, Spain

rcanoreag@gmail.com, mario.arias@ucm.es

Abstract. Achieving the environmental objectives determined by United Nations in the 2030 GDS goals will be difficult. CO2 emissions are the leading causes of the greenhouse effect on the planet. Nowadays, public and private institutions are applying innovative methodologies to bring together profitability and sustainability, reducing CO2 emissions. Companies are aligning their processes to reach net-zero emissions in 2050. Each year increase number of this. Carbon credits can help companies to meet their climate-change goals. The involvement of all citizens will be an essential factor to get this challenge. For this, the Voluntary offsetting market is created. Companies and citizens will pay for their emissions based on their carbon footprint. Investment in climate-related activities to combat climate change should reach more than 5 trillion by 2030. Based on the framework described above, digital technologies should play an important role. Digital applications and digital environments have begun to emerge. That can be defined as a voluntary digital market where people and organizations can calculate and offset or sell their CO2 footprint according to their needs. There is a lot of work yet. The voluntary offsetting market needs to be transparent and credible. Compliance standards, which should be a keystone of this market, are scarce or nonexistent. The purpose is not easy. Technology can help to reach this goal. Appear a lot of digital marketplaces trading CO2 offsetting every day. We search into these websites, comparing their most essential characteristics.

Keywords: Technology and Sustainability, CO2 offsetting, marketplaces

1 Introduction

It is widely accepted that society is increasingly aware of environmental challenges (Thompson & Harris, 2021). Public and private institutions are researching and applying innovative methodologies to create an ecosystem that can bring together profitability and sustainability (Valls-Val & Bovea, 2021; Thompson, 2021). Another keystone to cope with the environmental challenge is the citizen's role. According to the United Nations (UNFCC, 2021), it will also be essential to have the involvement

of all citizens to achieve the environmental objectives determined by the United Nations in their 2030 GDS goals. The ecological challenge is complex and has many facets, but one of the most important is CO₂ emissions. CO₂ emissions are the leading causes of the greenhouse effect on the planet. Today, the results of these emissions are highly relevant to all world economies. Global warming of the earth is mainly due to these emissions that provoke devastating climatic events with an enormous economic impact (Bjelle et al., 2021). Extreme weather won't be the only climate-related threat to supply chains in the years ahead. One consequence worries many: as demand increases for materials with low emissions intensity, such as green steel, production capacity may not expand quickly enough to keep pace, at least in the short term. For example, McKinsey analysis suggests that shortages of high-quality iron ore could constrain the production of zero-emissions steel (Bowcott et al., 2021). Environmental concerns are affecting both the planet and each person. It is urgent to limit CO₂ emissions to overcome the climate imperative. In the last Conference of the Parties Held, COP-26 in Glasgow (Growing Concerns for COP26, 2021), there has been a significant advance exploring the mechanisms that can generate and favour a Voluntary Emissions Offset trade. For instance, large industrial companies must seek efficiencies in their operational processes to reduce their carbon footprint. A growing number of companies are pledging to help stop climate change by reducing their greenhouse gas emissions as much as they can. But with the current development of technology, it is almost impossible for many businesses to eliminate their emissions or even lessen them as quickly as they might like. The challenge is especially tough for organisations that aim to achieve net-zero emissions, which means removing as much greenhouse gas from the air as they put in. This situation pushes the creation of carbon credits to offset emissions they can't get rid of by other means, compensating in that way their CO₂ emissions (Blaufelder et al., 2021). The COP26 has been treated in-depth on establishing a fair trade of rights for their compensation. But it is not enough only with the contribution of large companies. It is necessary that each individual, at a particular level, can offset their carbon footprint. The first step should be calculating each person's Carbon Footprint (Geneidy et al., 2021).

When persons and companies know their emissions and the amount that should be compensated, a second step should create a voluntary offset market that allows companies and individuals who wish to offset their emissions to directly pay for their greenhouse gas emissions. A voluntary offset market has been developed independently of the international Kyoto Protocol. In this market, NGOs, businesses, and individuals can produce and consume voluntary offsets. According to Lovell (2010), a significant problem exists in this market: the lack of widely used international standards or regulations. Under the 2015 Paris Agreement, nearly 200 countries have endorsed the global goal of limiting the rise in average temperatures to 2.0 degrees Celsius above preindustrial levels and ideally 1.5 degrees. Reaching the 1.5-degree target would require that global greenhouse-gas emissions are cut by 50 per cent of current levels by 2030 and reduced to net-zero by 2050. More companies are aligning themselves with this agenda: in less than a year, the number of companies with net-zero pledges doubled, from 500 in 2019 to more than 1,000 in 2020 (Blaufelder et al., 2021). Carbon credits can help companies to meet their climate-

change goals. Therefore, it will be essential in the coming years to find the means that generate this voluntary market credibly and efficiently to attract as many users as possible. (Kreibich & Hermwille, 2021). However, finding the footprint calculation requires a methodology, and the allocation of compensation rights will not be easy to determine (Sheather, 2021). The analysis and allocation are fundamental initial steps for compensation markets, becoming a top (Sheather, 2021). priority for the development of these markets.

2 Methods

Based on the framework described above, digital technologies should play an important role. Digital applications and digital environments have begun to emerge with two main objectives: (1) to calculate the carbon footprint by applying a world standard method; and (2) based on the previous calculation of the carbon footprint, to create mechanisms to offset this footprint by choosing a specific compensation reliable project inside the same digital environment that (Warburg et al., 2021). That can be defined as a voluntary digital market where people and organisations can calculate and offset or sell their CO₂ footprint according to their needs. In recent years, studies on consumer behaviour when choosing and participating in these markets have allowed consumers to compensate for emissions voluntarily, improving their CO₂ challenges (Warburg et al., 2021).

In these markets, measurement and disclosure are unavoidable; forcing digital technologies to create exchanges defining prices transparently can have benefits (Bowcott et al., 2021). Even using digital technologies to calculate and compensate carbon emissions, monetizing it in a safe and trusted environment is not easy for both individuals and organisations (Bowcott et al., 2021). Carbon offsets are produced and sold under the international climate change regime (the United Nations Kyoto Protocol) and within an expanding voluntary offset market. Companies and individuals can voluntarily trade to balance their greenhouse gas emissions. The volume of carbon produced and consumed within compliance and voluntary markets has grown dramatically in the last five years, raising several governance challenges (Lovell, 2010).

Digital Marketplaces: E-commerce of Voluntary Carbon Offsetting.

A global carbon market has evolved in recent years after the United Nations Kyoto Protocol. It has significant growth potential serving countries, organisations and individual customers. However, this market has been characterized by an absence of publicly available market information and a lack of transparency (Harris, n.d.). Compliance standards, which should be a keystone of this market, are scarce or nonexistent.

According to Reuters, investment in climate-related activities to combat climate change should reach more than 5 trillion by 2030.

Given the small and fragmented nature of the retail market and the lack of centralized registration for non-CDM¹ projects, it isn't easy to estimate the size of the market. The World Bank maintain databases of non-CDM project transactions, but they are primarily incomplete due to the above reasons (World Bank, 2021). The retail market for carbon offsets is relatively small and fragmented. Many consumers and organisations are unaware of what makes complex voluntary compensation for both people and organisations. In this context, it isn't easy to achieve ambitious objectives. In recent years, taking advantage of technological advances, many marketplaces are appearing. Their primary purpose is to show potential clients the projects to offset their carbon footprint. Our study analyses the current marketplaces with a qualitative methodology to determine their main features, what motivates consumers' participation, and their ethical priorities of users' influence. We are going to review 14 websites. These sites choose the most important in the market. We search into the websites the number of users, the total of carbon offsetting, the main characteristics, their first application in this market, and other specific situations that we find out while we do the study. Then, we will make a ranking that recognizes the essential qualities of users.

References

- Atwoli, L., Baqui, A. H., Benfield, T., Bosurgi, R., Godlee, F., Hancocks, S., Horton, R., Laybourn-Langton, L., Monteiro, C. A., Norman, I., Patrick, K., Praities, N., Rikkert, M. G. O., Rubin, E. J., Sahni, P., Smith, R., Talley, N., Turale, S., & Vázquez, D. (2021). Call for emergency action to limit global temperature increases, restore biodiversity, and protect health. *Tidsskrift for Den Norske Laegeforening*, 141(12). <https://doi.org/10.26633/rpsp.2021.122>
- Bjelle, E. L., Wiebe, K. S., Többen, J., Tisserant, A., Ivanova, D., Vita, G., & Wood, R. (2021). Future changes in consumption: The income effect on greenhouse gas emissions. *Energy Economics*, 95. <https://doi.org/10.1016/j.eneco.2021.105114>
- Blaufelder, C., Levy, C., Mannion, P., & Pinner, D. (2021). A blueprint for scaling voluntary carbon markets to meet the climate challenge.
- Growing concerns for COP26. (2021). *Nature Plants*, 7, 1323, Issue 10). <https://doi.org/10.1038/s41477-021-01012-x>
- Harris, E. (n.d.). The voluntary carbon offsets market An analysis of market characteristics and opportunities for sustainable development. www.iiied.org
- Lovell, H. C. (2010). Governing the carbon offset market. *Wiley Interdisciplinary Reviews: Climate Change*, 1(3), 353–362. <https://doi.org/10.1002/wcc.43>

¹ CDM: Clean Development Mechanism

- Thompson, B. S. (2021). Corporate payments for ecosystem services in theory and practice: links to economics, business, and sustainability. *Sustainability* (Switzerland), 13(15). <https://doi.org/10.3390/su13158307>
- Thompson, B. S., & Harris, J. L. (2021). Changing environment and development institutions to enable payments for ecosystem services: The role of institutional work. *Global Environmental Change*, 67. <https://doi.org/10.1016/j.gloenvcha.2021.102227>
- UNFCC. (2021). COP26 Explained. In COP26, Explained.
- Valls-Val, K., & Bovea, M. D. (2021). Carbon footprint in Higher Education Institutions: a literature review and prospects for future research. In *Clean Technologies and Environmental Policy*, 23(9). <https://doi.org/10.1007/s10098-021-02180-2>
- Warburg, J., Frommeyer, B., Koch, J., Gerdt, S. O., & Schewe, G. (2021). Voluntary carbon offsetting and consumer choices for environmentally critical products—An experimental study. *Business Strategy and the Environment*, 30(7). <https://doi.org/10.1002/bse.2785>
- World Bank. (2021). World Bank Indicators Database. In Data.

Using VLE Engagement Tracking to Offer Leniency to Under Performing Students

Brian Keegan¹, Dympna O'Sullivan¹, Brian Gillespie¹, and Paul Doyle¹

ASCNet Research Group, Technological University Dublin, Dublin, Ireland
brian.x.keegan@TUDublin.ie

Abstract. Data available in VLEs offer insights into student engagement and participation on a course. Where there is no formal requirement for engagement, the question arises if this information can be used as a form of discrimination where a lecturer can decide if additional support or leniency can be given based on access to this data. In the absence of policy, an inconsistent use of this data could be considered an ethical issue, allowing use of the data to justify the introduction of forms of discrimination based on pseudoscientific interpretations of partial data. Current mechanisms for leniency involve informal discussions between lecturer and student, mentor and staff, review of personal circumstances etc. Areas of discretion are late submissions, suspicion of plagiarism, and leniency. Risks include using the data at later stages to offer leniency to the same student in future assessments. The open book exam process removes the blind grading which may previously have utilised a number to identify candidates.

Keywords: VLE, bias, engagement, leniency, policy, ethics

1 Introduction

As student's progress through their education, their path is not always linear. For some student's, this pathway includes missed deadlines, missed assessments, failed exams and in some cases repeating or withdrawal from the course. Leniency can be offered when exceptional circumstances arise which can affect the group or for personal circumstances which affect the individual. These scenarios are covered by policy which act as enablers to provide students with a fair opportunity to successfully progress. Discretionary effort is well documented (Lloyd, 2008) in the workplace and is defined as an employee's willingness to go above minimal job responsibilities. This is linked directly to their job performance. A similar a behaviour can also be observed in academic studies. However, it is not as apparent since students who underperform do not progress. To mitigate underperformance, students can be offered extended deadlines or supplemental material to allow them to progress. Virtual Learning Environments (VLE) offer a source of detailed historical information on the student. The lecturer may simply review this information to quickly ascertain the perceived level of engagement the student has in the course. By doing so, they may be offering

leniency for no other reason than the student making an effort or turning up. We ask if this form of unconscious bias is a fair assessment or means to offer leniency. Such a student may in fact be gaining an unfair advantage over other students.

2 Background

2.1 Literature

The issue of bias, whether it is conscious or not, has been reported in the literature across all fields and all levels of education (Dee, 2017) (Backhus, 2019). While the validity of different bias is often debated, such as unconscious racial bias (URB) (Frisby, 2021), the ability to act, either consciously or unconsciously on a bias is based on the availability of data that potentially feeds that bias. Within education, specifically with the use of VLEs, educators have additional data available when dealing with students. While the capability exists for creating anonymous information about students, our VLEs often do the opposite, and provide further information about student engagement, and activity within a module. While the argument could be made that lack of engagement could be used to help identify students that are struggling, it can also be argued that that information can be used to reinforce a bias regarding students' performance. Indeed, AI systems using such data could identify generalised correlations resulting in a bias either for or against an individual (Silberg, 2019). While we adopt increased technology into our assessment and learning environments, we should be cognisant that the provision of such data may create additional engagement biases into our assessment. While some studies report on increased dissatisfaction by students in the increased use of VLEs (Martín, 2021), the move to assessment using such technology eliminates the anonymity of the student in the grading process where there is grading ranges and discretion by the lecturer for example in the use of Open Book Exams. Indeed, there is literature to support the concept that the halo effect as outline by Malouff et al (Malouff, 2013) requires that our VLE systems should assist in keeping students anonymous to minimise the risk of biased grading.

2.2 Virtual Learning Environments

VLEs form a vital part in the delivery and communication of material and assessments for students. Usage has increased significantly as a direct result of Covid measures brought in to facilitate remote learning (QQI, 2020). The increased usage of these platforms has now created a rich source of data in relation to engagement and progression. The VLEs provide resources and tools to actively monitor students on a course of study in a way that was not easily accessible or previously available (UCL, 2021). Furthermore, the use of open book exams delivered via the VLE coupled with a change in policy in the exams process provides a more complete profile on the students' academic achievement over each semester. The availability of this

information has presented the possibility for unconscious bias by lecturing staff when making decisions on assessments.

3 Approach

3.1 Source of Data

As part of our study, we use the VLE Brightspace as a use case with the tools and resources accessible to all staff. While we have seen in the past VLEs used simply as place holders for links to staff hosted content, the recent pivot to online delivery has meant that a consistent approach be taken to the student learning experience. Weekly assessments provide an insight into the level of student engagement in addition to their actual performance. This information can also be viewed alongside their online logins and access to the VLE. When online exams are hosted in the VLE it provides an additional source of performance information available in the same place. The VLE submission system removes the anonymity and clearly identifies the individual. The VLE also offers the ability for students to engage in forums for peer interaction which can also be used to measure their engagement.

Our use case contains a qualitative staff survey to gauge the levels of leniency offered to students on an individual basis. The survey gathers data on the perceived level of bias and unfairness which may be introduced. We also ask questions relating to awareness and compliance with existing policies and how they would regard the introduction of future policy.

3.2 Ethical Issues

In the absence of a defined process and regulation, the availability of this information can allow an assessor to view the students' performance in the context of their engagement without a consistent approach. For example, when leniency is being requested for a deadline extension the individual student's engagement could be used to determine the decision. This is a form of bias as the student has now been given an unfair advantage over the student who has not fully engaged, as no such requirement has been enforced in policy. The risks from the scenarios outlined here include potential for harm at an individual level and an institutional level. It is also possible that such scenarios could be in breach of dignity and respect policies regarding the lack of anonymity around student data.

4 Summary

The paper describes a scenario which can lead to unfairness and unconscious bias when determining the level of leniency offered to students based on their level of engagement. In the absence of regulation and clear policy we query the ethics behind the use of such sources of information as influencers in decision making. We

recommend that any such decisions should at a minimum, be included in clear school level policy to maintain a consistent approach. The policy should be communicated clearly to students in advance.

References

- Backhus, Leah M., et al. (2019) "Unconscious bias: addressing the hidden impact on surgical education." *Thoracic surgery clinics*, 29.(3), 259-267.
- Dee, T, and Seth Gershenson. (2017) "Unconscious Bias in the Classroom: Evidence and Opportunities, 2017." *Stanford Center for Education Policy Analysis* (2017).
- Frisby, Craig L. (2021) "Science versus the Unconscious Bias Paradigm in Education: A Review Essay." *Journal of School Choice*, 15(1), 139-156.
- Lloyd, R. (2008). Discretionary effort and the performance domain. *The Australasian Journal of Organisational Psychology*, 1, 22-34.
- Malouff, John M., Ashley J. Emmerton, and Nicola S. Schutte. (2013) "The risk of a halo bias as a reason to keep students anonymous during grading." *Teaching of Psychology* 40(3), 233-237.
- Martín, Torres, César, et al. (2021), "Impact on the virtual learning environment due to COVID19." *Sustainability* 13(2) 582.
- QQI (2020). The Impact of COVID-19 Modifications to Teaching, Learning and Assessment in Irish Further Education and Training and Higher Education. <https://www.qqi.ie/>
- Silberg, Jake, and James Manyika. (2019) "Notes from the AI frontier: Tackling bias in AI (and in humans)." *McKinsey Global Institute* (June 2019)
- UCL (2021). Monitoring Student Engagement and Progress in Moodle. <https://www.ucl.ac.uk/teaching-learning/publications/2021/nov/monitoring-studentengagement-and-progress-moodle>

Towards A Theory of Artificial Justice: Rawlsian Ethics Guidelines for Fair AI

Salla Ponkala¹[0000-0002-1441-8673]

¹ University of Turku, Turku, Finland
salla.k.ponkala@utu.fi

Keywords: Artificial Intelligence, AI ethics guidelines, Rawls, justice as fairness, democracy

Extended abstract

The expansion of Artificial Intelligence (AI) technologies in the 21st century has sparked numerous discussions on ethicality of AI. Several authors have speculated over AI's potential to threaten societies, democracies, and the entire humanity (see e.g., Bostrom 2014; Bostrom 2016; Russell 2019; Zuboff 2019). This development has forced governmental organizations, NGOs, as well as private companies to react and draft ethics guidelines for future development of ethical AI technologies (for extensive reviews, see Jobin et al. 2019; Hangendorff 2020). To give a few examples, The Organisation for Economic Co-operation and Development introduced *The OECD Principles on Artificial Intelligence* in 2019, and the European Union published the first draft of *Ethics Guidelines for Trustworthy AI*¹ in 2019 and a *White Paper on Artificial Intelligence*² in 2020. UNESCO adopted a set of guidelines in 2021.³ Even Vatican state has introduced its own guidelines to encourage ethical development of these technologies, prepared in cooperation with tech giants, such as Microsoft and IBM.⁴

Jobin et al. (2019) identified five principles that are dominating ethics guidelines: more than half of the 84 documents reviewed addressed **transparency, justice and fairness, non-maleficence, responsibility, and privacy**. They found, however, that these principles are not always interpreted in a similar way. Furthermore, guidelines appear to differ in emphasis and justification of these principles as well as how they should be interpreted, and they do not necessarily address the question of priority of principles in case of conflicting values (Jobin et al. 13–14; Hangendorff 2020; Vakkuri et al. 2021). It thus seems that, although the abovementioned principles seem to dominate AI guidelines, they still remain ambiguous.

¹ The Ethics Guidelines for Trustworthy AI by the High-Level Expert Group on Artificial Intelligence, set by the European Commission: <https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1>.

² European Commission's White Paper on Artificial Intelligence: https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

³ UNESCO AI ethics guidelines: <https://en.unesco.org/artificial-intelligence/ethics#drafttext>.

⁴ Vatican backs AI regulations to 'protect people': <https://www.politico.eu/article/vatican-calls-for-ai-ethics-with-backing-of-ibm-microsoft/>.

Moreover, according to Hagendorff (2020, 105), the majority of the most important AI ethics guidelines tend to neglect the potential impacts of AI on democracy, governance, and political deliberation. Meanwhile, several studies imply that the introduction of new AI technologies might threaten the existence of Western democracies by distorting political opinion formation and elections (Nemitz 2018; Chesney & Citron 2019; Kilovaty 2019; Manheim & Kaplan 2019; Brkan 2019; Alnemr 2020; Feezell et al. 2020; König & Wenzelburger 2020; Paterson & Hanley 2020; Cohen & Fung 2021), eroding trust towards democratic institutions (Manheim & Kaplan 2019; Paterson & Hanley 2020; Chesney & Citron 2020), and violating fundamental democratic values, such as equality and justice (Hacker 2018; Tolan 2018; König & Wenzelburger 2020; Janssen et al. 2020).

If AI technologies indeed threaten Western democracies, it seems alarming that the existing guidelines neglect this question. Although the five most prevalent principles extracted by Jobin et al. (2019) – transparency, justice and fairness, non-maleficence, responsibility, and privacy – do overlap with the democratic values found in Western constitutions, not explicitly addressing the relationship between AI and democracy seems alarming: it risks arriving at ethics guidelines that do not take into consideration the potential effects of AI on the very fundamental institutional and political structure of Western societies.

The common principle – or value – that is strongly highlighted in the majority of AI ethics guidelines, democratic theories and democratic constitutions is **justice**. According to Jobin et al. (2019), 68 out of the 84 ethics guidelines they reviewed took justice and fairness as an ethical principle to align AI with, entailing concepts such as inclusion, equality, equity, (non-)bias, (non-)discrimination, diversity, accessibility, and distribution. Numerous scholars handle questions concerning issues with biased algorithms and discrimination (see e.g., Hacker 2018; Tolan 2018; König & Wenzelburger 2020; Janssen et al. 2020), and the number of media outlets on the matter is overwhelming^{5, 6, 7}, which justifies its relevance as a key concept for ethics guidelines.

Putting justice in the centre of AI ethics turns the focus towards one of the most important modern theorists on the matter, John Rawls. Rawls developed his theory of justice mainly in his works *A Theory of Justice* (1971, revised edition published in 1999) and *Political Liberalism* (1993). Rawls is often described as one of the most influential philosophers of the 20th century, which is why his contractarian theory on justice as fairness and its principles of justice give a fruitful starting point for defining principles for ethical AI in a democratic setting. Rawls intended his theory to build “the most appropriate moral basis for a democratic society” (Rawls 1971, viii), which in

⁵ “What Do We Do About the Biases in AI?” Publication in Harvard Business Review on 25 October 2019: <https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai>.

⁶ “Algorithms Are Running Foul of Anti-Discrimination Law” Publication in Brink News on 7 December 2020: <https://www.brinknews.com/algorithms-are-running-foul-of-anti-discrimination-law/>.

⁷ “Discrimination algorithms: 5 times AI showed prejudice” Publication in New Scientist on 12 April 2018: <https://www.newscientist.com/article/2166207-discriminating-algorithms-5-times-ai-showed-prejudice/>.

theory makes it suitable for creating AI ethics guidelines for democratic societies. This is something that we will consider critically in the course of this paper.

In this paper, Rawls's theory of justice is applied to draft a set of AI ethics guidelines. The goal is to contribute to broadening of the discussion on AI ethics by exploring the possibility of constructing AI ethics guidelines with a specific – although vast – goal of aligning AI development with democracy. This is, however, an endeavour with obvious difficulties. Both AI and democracy are concepts with no shared definition. For this reason, this experiment might seem like an oversimplification of both AI and democracy. Regardless, I argue that bold experiments are necessary to open ever deeper discussion and debate, and thus finally, perhaps, end up with a set of ethical guidelines for AI that is democratically sustainable.

First, the key concepts of Rawls's theory are applied to the context of AI ethics, which serves as a theoretical starting point for defining ethics guidelines. Institutions that belong to the basic structure of society and thus are targeted by the AI ethics guidelines are discussed; Rawls's concept of the original position as key basis and a tool for further tuning of the principles is introduced; the principles of justice – including the basic liberties – and intended interpretations are discussed. Then, on this foundation, a suggestion for a set of ethics guidelines, titled *Rawlsian Ethics Guidelines for Fair AI* are drafted, which is in line with Rawls's theory of justice as fairness. Finally, conclusions are drawn and the academic potential of such an experiment and possibilities for future research are discussed.

It needs to be highlighted that this paper is an experiment, and it is by no means suggested that a guideline based on only one theory and one societal aspect can offer an exhaustive solution to the numerous problems concerning the current state of AI ethics guidelines. In addition, we hardly need a higher number of ethics guidelines to accompany the existing ones. Furthermore, as any ambitious theory, Rawls's theory contains several challenges (see e.g., Anderson 2003; Sen 2009) that need to be addressed when developing AI ethics guidelines. We believe, however, that deliberation around existing ethics theories and reflecting them with the current AI development is necessary in order for us to align AI development with our values of choice. We hope this paper gives insights and inspiration for everyone working with AI technologies and inspires others to explore AI ethics from other theorists' viewpoints.

References

- Anderson, B. C. (2003). The antipolitical philosophy of John Rawls. *Public Interest*, 151(2003), 30–51.
- Alnemr, N. (2020). Emancipation cannot be programmed: blind spots of algorithmic facilitation in online deliberation. *Contemporary Politics*, 26(5), 531–552.
- Bostrom, N. (2016). The control problem. Excerpts from superintelligence: Paths, dangers, strategies. *Science Fiction and Philosophy: From Time Travel to Superintelligence*, 308–330.
- Bonstrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
- Brkan, M. (2019). Artificial Intelligence and Democracy: The Impact of Disinformation, Social Bots and Political Targeting. *Delphi – Interdisciplinary Review of Emerging Technologies*, 2(2), 66–71.

- Chesney, B. – Citron, D. (2019). Deep fakes: looming challenge for privacy, democracy, and national security. *California Law Review* 107(6), 1753–1820.
- Cohen, J. – Fung, A. (2021). Democracy and the Public Sphere. IN *Digital Technology and Democratic Theory*, Bernholz, L., Landemore, H. & Reich, R. (eds.). Chicago: The University of Chicago Press. 23–61.
- Freezell, J. T. – Wagner J. K. – Conroy, M. (2020). Exploring the effects of algorithm-driven news sources on political behavior and polarization. *Computers in Human Behavior*, 116(2021), 1–11.
- Hacker, P. (2018). Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review*, 55(2018), 1143–1186.
- Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30(2020), 99–120.
- Janssen, M. – Hartog, M. – Matheus, R. – Yi Ding, A. – Kuk, G. (2020). Will Algorithms Blind People? The Effect of Explainable AI and Decision-Makers' Experience on AI-supported Decision-Making in Government. *Social Science Computer Review*, 2020, 1–16.
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399.
- Kilovaty, I. (2019). Legally cognizable manipulation. *Berkeley Technology Law Journal*, 34(2), 449–501.
- König, P. D. – Wenzelburger G. (2020). Opportunity for renewal or disruptive force? How artificial intelligence alters democratic politics. *Government Information Quarterly*, 37(2020), 1–11.
- Manheim, K. – Kaplan, L. (2019). Artificial Intelligence: Risks to Privacy and Democracy. *Yale Journal of Law and Technology*. Vol. 21. 107–188.
- Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions A*, 376(20180089), 1–14.
- Paterson, T. – Hanley, L. (2020). Political Warfare in the digital age: cyber subversion, information operations and 'deep fakes'. *Australian Journal of International Affairs*, 74(4), 439–454.
- Rawls, J. (1971). *A Theory of Justice*. Harvard.
- Rawls, J. (1999). *A Theory of Justice. Revised edition*. Cambridge: Belknap press of Harvard University Press.
- Rawls, J. (2005) [1993]. *Political Liberalism*. New York: Columbia University Press.
- Russell, S. (2019). *Human Compatible. Artificial Intelligence and the Problem of Control*. New York: Penguin.
- Sen, A. (2009). *The Idea of Justice*. London: Penguin.
- Tolan, S. (2018). Fair and Unbiased Algorithmic Decision-Making: Current State and Future Challenges. *JRC Digital Economy Working Paper 2018-10*. European Union.
- Vakkuri, V. – Jantunen, M. – Halme, E. – Kemell, K. K. – Nguyen-Duc, A. – Mikkonen, T. & – Abrahamsson, P. (2021). Time for AI (Ethics) Maturity Model Is Now. *arXiv preprint: arXiv:2101.12701*.
- Zuboff, S. (2020). *The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs.

Minding the Gap:

Computing Ethics And the Political Economy of Big Tech

Ioannis Stavrakakis¹, Damian Gordon, J¹. Paul Gibson², Dympna O’Sullivan¹ and
Anna Becevel¹

¹Technological University of Dublin, Ireland

²TELECOM SudParis, France

Ioannis.Stavrakakis@TUDublin.ie, Damian.X.Gordon@TUDublin.ie,
Dympna.OSullivan@TUDublin.ie, Anna.Becevel@TUDublin.ie,
Paul.Gibson@telecomsudparis.eu

Extended abstract

In 1988 Michael Mahoney wrote that “[w]hat is truly revolutionary about the computer will become clear only when computing acquires a proper history, one that ties it to other technologies and thus uncovers the precedents that make its innovations significant” (Mahoney, 1988). Today, over thirty years after this quote was written, we are living right in the middle of the information age and computing technology is constantly transforming modern living in revolutionary ways and in such a high degree that is giving rise to many ethical considerations, dilemmas, and social disruption. To explore the myriad of issues associated with the ethical challenges of computers using the lens of political economy it is important to explore the history and development of computer technology. A significant turning point was in the late 1930s with the work of Alan Turing in conjunction with the work of Alonzo Church and his Lambda Calculus. The 1940s and 50s were integral for the researching and development of computing technologies. According to Edwards (1995) the technical demands of weaponry in WWII created huge needs for a large number of fast computations and since then the US armed forces have been one of the most important agents for advanced research in computing and this research along with the massive investments in military projects during the 1950s brought in a new entrepreneurial and interdisciplinary way of collaborating between scientists, administrators and military personnel while breaking down bureaucratic barriers (Turner, 2006). All that, produced concepts such as cybernetic systems, webs of information and the importance of gathering and interpreting of information (or data) as a way of making sense of the physical and social worlds (ibid.).

It comes as no surprise then that the field of Computer Ethics was founded in the 1940s by Norbert Wiener, almost simultaneously with the development of electronic computers, in what he called “Cybernetics” (Bynum, 2000). Wiener predicted

computing technology's potential social and ethical implications while working on developing WWII anti-aircraft weaponry. Turner (2006) writes that after WWII two cultures emerged in the US around technological progress. The first was the military – industrial research that was established in the 1940s and came into full power during the Cold War. The second emerged some years later as a reaction to the first and it was the American counterculture which Turner defines as “a culture antithetical to the technologies and social structures powering the cold war state and its defence industries” (ibid). The various, social, political and military pressures of the 40s and the 50s served as the boiling pot that led to waves of political protests and personal exploration of the 1960s for which, Turner writes “much of it aimed at bringing down the cold war military-industrial bureaucracy” (2006, p. 3).

As many academic institutes began to develop courses to teach computer programming in the 1960s, a culture of openness and sharing emerged with a focus on developing an understanding of the potential of computers, as well as a promotion of ideas such as the decentralization of power, and a mistrust of authority, which were already present in the counterculture zeitgeist of that era (Levy, 1984). At the same time, according to Bynum (2000), the first ethical and social implications of computing technology were starting to appear as well as concerns about topics such as how these new technologies could facilitate authoritarian actions like privacy invasions by government agencies. Many of these concerns were captured in various artistic forms (films, books, comics etc.) as well as in scientific studies, government discussions and proposed legislation (Bynum, 2000).

Computer ethics at the time was mostly about the potential consequences of future computing technologies. As technology progressed and computers found new uses in already existing aspects of everyday life and in other fields of study, such as medicine (Bynum, 2001) new ethical considerations started to emerge. In the 1980s, two seminal works came out establishing the ethics of computing as a distinct field of study by James Moor (1985) and Deborah Johnson (1985). At the same time other thinkers, like Donald Gotterbarn (1991), were proposing that computing ethics should be mainly focused on being a professional code of conduct. Indeed there have been very important steps taken by professional bodies to establish code of ethics and good practices such as those by the ACM (1992) and IEEE (Shahriari & Shahriari, 2017). Since the 1980s a myriad of actions have been taken to promote computing ethics including new theories and considerations of ethics (e.g. Floridi, 1999; GorniakKocikowska, 1996). However, little attention has been paid by Computer Science (and related fields) into the historical – political – and economic realities that have shaped and still influence the development and evolution of current computer technology.

In the 1990s, the emancipatory belief in technology carried over from the counterculture of the 60s fused with the entrepreneurial and libertarian thinking of Silicon Valley in what Barbrook and Cameron term the ‘Californian Ideology’ (1996, p. 1). They argue that this idea of the future was adopted by an assortment of diverse groups from computer enthusiasts, to students, to investors, to activists all the way to politicians in the US (ibid). They further state that this antithetical mix of

“technological determinism” and “libertarian individualism” has become the “hybrid orthodoxy of the information age”.

Studies show that a number of socio-economic factors in the last fifty years played significant roles in the creation of digital economy as it is today. Srnicek (2017) places first the economic downturn of the 1970s with the drop in manufacturing profitability in advanced economies. Then comes the dot-com boom and bust of the 1990s where technological infrastructure saw great private investments and the internet was commercialised for the first time. Venture capital (VC) was paramount for the development of this newly-formed sector. The business models at the time were in uncharted waters and revenue generation was still a puzzle, therefore many companies would aim for growth rather than profit (Zuboff, 2019) or as Srnicek puts it, the goal of internet based companies since then has been monopolistic dominance (Srnicek, 2017). The dot-com boom years saw huge VC investments in Silicon Valley of what researchers termed as “impatient money” (Zuboff, 2019) looking for a big and quick return on investments from promising start-ups increasing the hype and financial volatility in the sector.

When the dot-com bubble burst in 2000 all these risky investments became even riskier causing panic in the market, making investors in Silicon Valley reluctant to invest further or to even abandon companies altogether leading to high start-up mortality rates (ibid). Tech start-ups from then on had to find ways to generate revenue. The new “mantra” for Silicon Valley investors became “an ability to show sustained and exponential profits” (Zuboff, 2019, p. 74). Zuboff argues that this stress for survival was what led to the defining moment in the evolution of the digital economy. She takes Google as the prime example for this where the company, struggling for its survival, and with an abundance of stored behavioural data that seemed useless at first, would realise that they can be used to run targeted, personalised ads for its users and thus generate revenue (Zuboff, 2019). That led to an increasing “behavioural value reinvestment cycle” (p.97) and would constitute what Zuboff defines as “surveillance capitalism”.

The final factor Srnicek includes in his analysis is the 2008 financial crisis which brought about low interest rates and companies with a surplus of cash looking for better investment rates in higher risk sectors such as the tech industry (Srnicek, 2017). The aforementioned economic conditions along with the social narratives of emancipation and social change through technology, the transition towards immaterial commodities also known as the ‘knowledge economy’ and the need of tech companies to better handle the commodification of the vast amount of data collected through online services heralded the creation of digital platforms as the new business model. Srnicek (2017, p. 52) writes that “often arising out of internal needs to handle data, platforms became an efficient way to monopolise, extract, analyse, and use the increasingly large amounts of data that were being recorded.”

The hope that technology will save the world has become a popular narrative for many technology corporations and it can be seen in company mottos like Google’s “don’t be evil” but the adoption of surveillance capitalist, neo-liberal business models in the form of ‘platformisation’ has come into stark contrast with some of those

narratives as can be seen from the IPO manifestos of several big tech corporations (Dror, 2013).

Digital platforms today are not only part of digital technology companies like Google, Facebook and Airbnb but also part of traditional industries such as General Electric, Siemens and others. Some researchers even argue that all these companies have effectively turned into data companies. Digital platforms, and more specifically their services, their level of control and their use of computing technologies have very significant social and ethical implications in democracy, finance, education, cultural production and labour and are defining how powerful technologies such as AI and Machine Learning are being used globally.

This paper argues that specific socio- economic circumstances have, to a large extent, formed the ground where modern computing technologies emerge and this is something that needs to be prominent in any considerations about the ethics of computing. For example when a social media platform's business model depends on capturing the user's attention for as long as possible so they can see more ads the platform's design will try to make sure that this is the case despite of what is ethical and the people who would have to work on that will have to do so within the rules of the market or the interests of the company. Addressing the issue from a computing ethics point of view is important but it might not be enough.

References

- ACM, C. M. (1992). ACM code of ethics and professional conduct. Code of Ethics.
- Barbrook, R., & Cameron, A. (1996). The Californian ideology. *Science as Culture*, 6(1), 44–72. <https://doi.org/10.1080/09505439609526455>
- Bynum, T. W. (2000). The foundation of computer ethics. *ACM SIGCAS Computers and Society*, 30(2), 6–13. <https://doi.org/10.1145/572230.572231>
- Bynum, T. W. (2001). Computer ethics: Its birth and its future. *Ethics and Information Technology*, 3(2), 109–112. <https://doi.org/10.1023/A:1011893925319>
- Dror, Y. (2013). 'We are not here for the money': Founders' manifestos. *New Media & Society*, 17(4), 540–555. <https://doi.org/10.1177/1461444813506974>
- Edwards, P. N. (1995). From 'impact' to social process: computers in society and culture. *Handbook of Science and Technology Studies*, 257–285.
- Floridi, L. (1999). Information ethics: On the philosophical foundation of computer ethics. *Ethics and Information Technology*, 1(1), 33–52.
- Gorniak-Kocikowska, K. (1996). The computer revolution and the problem of global ethics. *Science and Engineering Ethics*, 2(2), 177–190.
- Gotterbarn, D. (1991). Computer ethics: Responsibility regained. *National Forum*, 71(3), 26.
- Johnson, D. (1985). *Computer ethics*. Prentice Hall.
- Levy, S. (1984). *Hackers: Heroes of the computer revolution* (Vol. 14). Anchor Press/Doubleday Garden City, NY.

- Mahoney, M. S. (1988). The History of Computing in the History of Technology. *Annals of the History of Computing*, 10(2), 113–125. <https://doi.org/10.1109/MAHC.1988.10011>
- Moor, J. H. (1985). What is computer ethics? *Metaphilosophy*, 16(4), 266–275.
- Shahriari, K., & Shahriari, M. (2017). IEEE standard review—Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. 2017 IEEE Canada International Humanitarian Technology Conference (IHTC), 197–201.
- Srnicek, N. (2017). *Platform capitalism*. John Wiley & Sons.
- Turner, F. (2006). *From Counterculture to Cyberculture: Stewart Brand, the Whole Earth Network, and the Rise of Digital Utopianism*. University of Chicago Press. <https://doi.org/doi:10.7208/9780226817439>
- Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile books.

Regulating AI in Europe: from Ethics to Legal Rules building on the GDPR example

Costas Popotas¹ and Maria Bottis²

¹LL.M QUB, Head of Unit, Establishment and Settlement of Entitlements CJEU, Luxembourg,
PhD cand.Ionian University

²Professor, Head of the Department of Archives, Library Science and Museology, School of
InformationScience and Informatics, Ionian University, Corfu, Greece

²Director, International Society for Ethics and Technology

¹Costas.Popotas@curia.europa.eu

²bottis@otenet.gr

Extended Abstract

Keywords: Artificial Intelligence, AI Ethics, Law, European Union

The proposed contribution intends to examine the challenges and opportunities of Artificial Intelligence (henceforth AI) and how appropriate regulation can conciliate the expectations from the insertion of AI in contemporary life and conformity to the rule of law. It will, in particular, attempt to assess the European Union's efforts to improve the ethical rules applicable to artificial intelligence (AI) and shift towards a model based on legal rules.

The paper will sketch how the current surge of interest in Artificial Intelligence (AI) developed a potential of opportunities that covers an impressive spectrum of applications in all possible aspects of human activity. It will also trace the societal footprint of AI, extending over state borders and escaping the reach of national regulators. Inevitably it will describe the competitive race of big actors to gain AI dominance.

Based on recent reports, the proposed presentation will outline how fundamental rights are taken into account when using or developing AI applications¹ in crucial areas or the complex context compiled by opportunities, risks, and governance challenges associated with AI and data use².

Then the paper will turn to the challenges for the regulator. In the first place, it will analyse the policies developed gradually around deontological rules. It will compare the different definitions of AI promoted via each initiative³; the main approaches adopted to commit actors around AI Ethics, like in the European Union⁴, the United States⁵, UK, China⁶ or multilateral efforts⁷. It will also examine the shortcomings encountered.

Finally, the presentation will focus on the recent legislative developments in the European Union that indicate a convergence towards legislative solutions imposing more stringent legal rules. The evolution of the European Union policies demonstrates a clear will to lead worldwide and exemplifies how this paradigm shift towards legislative tools can obtain followers in the future.

Up to recently, even at the EU level, the attitude was to elaborate on ethical principles existing in other domains and establish moral guidelines and boundaries for the use of AI. At the level of the European Union, the first efforts concentrated on ethical rules, promoting fairness, transparency and accountability in AI. On Monday, April 8 2019, the European Commission's High Level Expert Group on Artificial Intelligence (HLEG) published its "Ethics Guidelines for Trustworthy AI"⁸, a voluntary framework to achieve legal, ethical, and robust AI. Four ethical principles were advanced as foundations of ethical AI: respect for human autonomy, harm prevention, fairness and explicability⁹.

In April 2020, the European Union took the lead worldwide and superseded the previous framework and proposed legal rules for artificial intelligence¹⁰. The proposed legislation will be an EU regulation¹¹, immediately enforceable and aspiring to universal application. It follows the EU's General Data Protection Regulation (GDPR) model closely, having the idea of compliance as the central point. The text also proposes a governance structure at the European and national levels.

The launch was followed on July 20, 2021, by a Conference¹² on the issue organised by the Slovenian Presidency of the EU.

The regulatory proposal bases itself on trust, extending the idea of trustworthy AI that was the core issue of the Guidelines. AI applications should remain trustworthy even after being placed on the market. The regulation aims to provide AI developers, deployers and users with precise requirements and obligations regarding specific uses of AI. At the same time, the proposal seeks to reduce administrative and financial burdens for businesses. The proposal is part of a broader AI package, including the updated Coordinated Plan on AI. The vision is to guarantee people and businesses' safety and fundamental rights while strengthening AI uptake, investment, and innovation across the EU.

The AI regulation will set up a risk-based approach, classifying AI applications according to four levels of risk: Unacceptable risk; High-risk; limited risk, minimal or no risk. For example, in the case of unacceptable risk, all AI systems considered a clear threat to people's safety, livelihoods, and rights will be banned. High-risk AI systems will be subject to strict obligations before being put on the market. Limited risk will concern AI systems with specific transparency obligations. Minimal risk AI systems, which correspond to most AI systems currently used in the EU, can be freely used. It will propose a list of high-risk domains and set precise requirements for AI systems in the areas and specific obligations for AI users and providers. A conformity assessment will occur before the AI system is put into service or placed on the market.

Following the Commission's proposal in April 2021, the regulation has advanced quickly through the legislative procedure in force. It has been examined at the European Economic and Social Committee¹³ and the European Committee of the

Regions. It was introduced for discussions at the Council of the European Union in October. It is thus expected to enter into force in the second half of 2022, with a European Artificial Intelligence Board overseeing the enforcement of the regulation. After a transitional period of two years, it should allow for developing procedures for risk assessment and conformity examination.

In conclusion, the paper shall retain the ambivalent position of both academics and regulators to maintain the need to implement a general framework or a unique approach amongst the key actors,

not to mention throughout the globe. It will examine the criticism against either approach

It will, though, accent the distinctive character of the proposal for AI regulation. This novel European approach will have to conciliate the rule of law and the protection of fundamental rights to counter the American and Chinese vanguard and protect "domestic" firms. Nevertheless, most importantly, it should evaluate whether the regulation will allow European AI ethics to become the yardstick for analogue developments in other legal orders, as it happened with the GDPR

References

- ¹ European Agency for Fundamental Rights Getting the future right – Artificial intelligence and fundamental rights (FRA), 14 December 2020, <https://fra.europa.eu/en/publication/2020/artificialintelligence-and-fundamental-rights> accessed on 22 February 2022
- ² AI Barometer 2020 - GOV.UK (www.gov.uk),
- ³ OECD Legal Instruments <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>;
- ⁴ EU guidelines on ethics in artificial intelligence: Context and implementation (europa.eu)
- ⁵ U.S. Chamber of Commerce - Artificial Intelligence (AI) Commission on Competition, Inclusion, and Innovation: <https://www.uschamber.com/technology/u-s-chamber-launches-bipartisancommission-on-artificial-intelligence-to-advance-u-s-leadership>
- ⁶ Beijing AI Principles Datenschutz und Datensicherheit (springer.com) : <https://link.springer.com/content/pdf/10.1007/s11623-019-1183-6.pdf>
- ⁷ Forty-two countries adopt new OECD Principles on Artificial Intelligence - OECD
- ⁸ Building trust in human-centric AI | FUTURIUM | European Commission (europa.eu)
- ⁹ The European Parliament commissioned a report on the context and implementation of the European Guidelines and on June 18 2020, it set up a special committee on artificial intelligence in a digital age, AIDA and defined its mandate - Special committee on artificial intelligence in a digital age: https://www.europarl.europa.eu/doceo/document/TA-9-2020-0162_EN.html
- ¹⁰ Regulatory framework on AI | Shaping Europe's digital future (europa.eu)

- ¹¹ <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN>
- ¹² Conference on Regulation of Artificial Intelligence – Ethical and Fundamental Rights Aspects, Collection-of-lectures-AI-Conference-si2021.eu-20.07.2021.pdf (gov.si)
- ¹³ C_2021517EN.01006101.xml (europa.eu)

Addressing ethical issues in the design of smart home technology for older adults and people with disabilities

Jonathan Turner, *Dympna O'Sullivan, Damian Gordon, Yannis Stavrakakis, Brian Keegan, Emma Murphy

ASCNet Research Group, Technological University Dublin, Dublin, Ireland

*Dympna.OSullivan@tudublin.ie

Abstract. Unique ethical, privacy and safety implications arise for people who are reliant on home-based smart technology due to health conditions or disabilities. In this paper we highlight a need for a reflective, inclusive ethical framework that encompasses the life cycle of smart home technology. We present key ethical considerations for smart home technology for older adults and people with disabilities and argue for ethical frameworks which combine these key considerations with existing models of design and development.

Keywords: Smart Home Technology, Pervasive Computing, Older Adults, People with Disabilities

1 Introduction

Smart home technology encourages independent living at home with the support of assistive technologies. Specialized assistive devices, smartphone or tablet based applications, on-body or passive sensing technology can be used to increase, maintain, or improve the functional capabilities of older adults or individuals with disabilities. Feedback from monitoring technology can be relayed to occupants or shared with informal caregivers to aid with decision making about health and wellbeing. Challenges in the ethical design and development of such technology include how to develop understandable and usable technologies so that they meet individual variations in needs and abilities so that they help to maintain autonomy, provide meaningful activities, address the emotional state of individuals and promote social inclusion (Nunes 2015). Moreover, there is great variety within user groups, such as differences in demographics (e.g., socioeconomics) and personality, but also due to the diversity of specific conditions, each with different behavioral, cognitive, and emotional consequences. We consider the following as pertinent ethical considerations when developing assistive smart home technologies.

Informed consent: The pervasive nature of some smart devices raises issues of technological understanding and consent. In addition, older adults or persons with specific disabilities might have a reduced or compromised ability to decide for themselves about the use of technology.

Privacy: Smart devices gather a broad spectrum of data about their users, ranging from in-application activity to communications to movement and location data. Combined with their pervasive nature, data can be collected and used in ways that are not always clear to end users.

Security: This involves physical security as well as security of the data and network from intrusion and cyber attacks (Karale 2021). Choosing the right technology to fit the requirements is crucial in avoiding over or unnecessary surveillance. For example a motion sensing device may be sufficient in place of a camera to determine activity.

Autonomy: Technology should be designed to accommodate existing living patterns and should offer users control and influence over their lives and well-being (FakhrHosseini 2019).

Safety: Ensuring the safety of older adults and persons with disabilities is crucial to their independence and quality of life. Safety and technological reliability are highly coupled and it is important that evaluations of smart technologies are not limited to laboratory settings rather than more complex real world environments (Pigini 2017).

Data Accuracy: The accuracy of data collected in smart spaces depends on a number of factors including device reliability, configuration or placement, misuse or misunderstanding. Smart sensors can also generate false positives and inferences and recommendations based on inaccurate data will contain errors (Aramendi 2019).

Data Sharing: Smart home data is often shared with manufacturers and third parties. This can be for varied purposes, to help improve the product or to aggregate data for analytics and insights. Older adults or persons with disabilities may wish to share data with formal or informal caregivers but they should have control over how and with whom their own data is shared. Data management policies should be available and accessible (Mocrii 2018).

Transparency: Transparency enables end users to understand the smart system. It incorporates previous factors such as privacy and data management and ensuring that these are well understood by those using the system. Transparency is important at both device and system levels (Yao 2019).

Trust: To trust decisions computed by smart systems, users need to know how that system arrives at its conclusions and recommendations. Trust is related to data accuracy and transparency above and explanation below (Cannizzaro 2020).

Explanation: Existing approaches to explanations for smart systems are tailored more towards interpretations that are more suitable for modelers and less for technically inexperienced users. The majority of smart systems do not incorporate explanation capabilities (Nikou 2019).

Acceptability: Pervasive technology requires data to understand the environment and individual. This means allowing technology access to our personal spaces. This can be intrusive if not done correctly and tailored for the cohort. Passive, low impact, low visibility, low maintenance and high reliability should be considered as high priority requirements when dealing with older adults and people with disabilities.

It is accepted that end users make trade-offs when using smart technology, for example, privacy for functionality or increased autonomy, security over privacy for

better surveillance, increased functionality or better displays for less explanations or usability for complexity. We argue that these trade-offs should not be inevitable, particularly for persons who are reliant on technology. We posit that an ethical, user driven framework incorporating a design-driven approach can reduce or eliminate these trade-offs by better understanding the requirements of end users.

2 Ethics and Smart Home Technologies

It is fair to say that software engineering has traditionally been driven by a utilitarian approach by focusing on outcomes in terms of the development of commercial products or services. However, virtue ethics, with its focus on choices that aim at the 'good life' is ideally suited for managing complex, novel, and unpredictable moral landscapes, just the kind of landscape that today's emerging technologies present (Vallor 2016). Value Sensitive Design (Friedman 2013), defined as "a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process" could be considered an example of Vallor's (2016) application of virtue ethics to technology.

The 'Human Factors & Ethics Canvas (HFEC)' introduced by (Cahill 2019) provides a bridge to integrating human factors and ethics issues with a particular focus on the collection of evidence using stakeholder evaluation methods. The HFEC allows non-ethicists such as Designers, Human Factors Researchers, Engineers, and Computer Scientists to engage in ethical issues pertaining to the emerging technology product. Frameworks such as that by (O'Keefe and O'Brien 2018) offer organizations a practical guide to implementing data ethics. Recent welcome developments have shifted the emphasis from outcomes to intentions to reduce blind spots in technology development. For example Consequence Scanning is an iterative approach that encourages organizations to consider potential consequences of products and services on people and communities (Brown 2019).

Projects involving human participants undergo ethical assessments and more recently data protection impact assessments but typically these occur at the end of the design phase. Ethical evaluation is usually late in technology development or research project and focus on the impact of the system as designed on the participants. At this point, it is arguably too late for researchers to consider questions such as "should this technology have been developed in the first place?". We argue that a framework is required that allows to reflect on ethical issues - those related to both intentions and outcomes - at challenge points throughout the technology life cycle.

3 A Framework for Inclusive Smart Home Technology

As part of the Ethics4EU project (Ethics4EU 2021), we are developing a new ethical framework - the 5D Framework - for the development of inclusive smart home technology by combining research presented above with aspects of the five-phase design thinking model proposed by the Hasso-Plattner Institute of Design (d.school), at Stanford, USA (Apiyanti 2019), as well as elements of the UK Design Council's

Double Diamond Model (Howard 2008). Crucially, the framework emphasizes that the user is at the heart of the framework - they must be the co-designers of the system; in combination with HFEC, the O'Keefe and O'Brien ethics model and the practical application of ethics in value-sensitive design (Friedman 2013). The full design team includes participants that are end-users, as well as experts in technology and relevant health domains i.e. general practitioners, occupational therapists, physiotherapists and other clinical specialists. An initial draft of the 5D framework is shown in Figure 1.

1. Discover
The full Design Team is trying to understand the needs of the end-user of the system. They, therefore, will speak to the end-user in a thoughtful and solicitous manner, as well as other parties that may provide useful insights. The end-user is also asking questions.
2. Define
The full Design Team are trying to encapsulate their findings from the Discover Stage into a series of models, noting key challenges (pinch points and pain points) as well as existing affordances. Again the end-user is a core member of the Design Team, and they are both the subject of the design, and the architect of the solutions.
3. Develop
The full Design Team are working on identifying a range of potential approaches to addressing the issues identified in the two previous stages. Again the end-user will be a vital force in the stage.
4. Deliver
The full Design Team is selecting a single potential solution from those developed in the previous stage, and it is vital that the end-user is asked and listened to.
5. Determine
The full Design Team is testing the effectiveness of their solution. The system is deployed and the team is determining what aspects of the system work well, and which are not fully serving their purpose. This section includes considerations relating to maintenance and sustainability.

Figure 1: 5D framework

4 Conclusions

Older people and persons with disabilities are vulnerable groups and their dignity, rights and privacy must be safeguarded. The development of inclusive home-based smart technology presents many unique ethical challenges, and when this is allied with these systems being developed for older adults and people with disabilities, the ethical concerns and considerations grow significantly. Assessing the ethical implications of new technologies which may have impacts we cannot predict, is very difficult. Critically, design frameworks should consider protections concerning potential negative consequences, unintended consequences and unknown future implications. In this paper we have outlined a framework for navigating some of these ethical issues using a range of techniques from Software Engineering, Human

Computer Interaction, Education, and Research Methods to produce a coherent new ethics driven approach that we have entitled “The 5D Framework” that puts the user at the heart of the process.

Acknowledgments

The authors gratefully acknowledge the support of the Erasmus+ programme of the European Union. The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

This material is based upon works supported by the Science Foundation Ireland under Grant No. 19/FFP/6917.

References

- Apiyanti, S. and Dewi, L., 2019, Design Thinking of the Inclusive Education Learning Strategy in the 5.0 Society Era”. 4th International Conference on Education and Regional Development (ICERD), pp.631-642.
- Aramendi, AA, Weakley, A., Goenaga, AA., Schmitter-Edgecombe, M., Cook, DJ, (2018). Automatic assessment of functional health decline in older adults based on smart home data, *Journal of Biomedical Informatics*, 81: 119-130.
- Brown S. (2019) Consequence Scanning Manual Version 1. London: Doteveryone.
- Cahill, J. 2019. The Human Factors & Ethics Canvas. Available at: <https://www.tcd.ie/cihs/projects/hfaecanvas.php>
- Cannizzaro S, Procter R, Ma, S, Maple C (2020) Trust in the smart home: Findings from a nationally representative survey in the UK. *PLoS ONE* 15(5): e0231615.
- Ethics4EU. 2021. www.ethics4eu.eu
- FakhrHosseini M., Lee C., Coughlin J.F. (2019) Smarter Homes for Older Adults: Building a Framework Around Types and Levels of Autonomy. In: Zhou J., Salvendy G. (eds) *Human Aspects of IT for the Aged Population. Social Media, Games and Assistive Environments. HCII 2019. Lecture Notes in Computer Science*, vol 11593. Springer, Cham
- Friedman, Batya, Peter H. Kahn, Alan Borning, and Alina Hultgren. 2013. “Value Sensitive Design and Information Systems.” In *Early Engagement and New Technologies: Opening up the Laboratory*, 59–95.
- Hasso Plattner Institute of Design (2010) An Introduction to Design Thinking: PROCESS GUIDE. Available: <https://web.stanford.edu/~mshanks/MichaelShanks/files/509554.pdf>

- Howard, T.J., Culley, S.J. and Dekoninck, E., 2008. Describing the Creative Design Process by the Integration of Engineering Design and Cognitive Psychology Literature. *Design studies*, 29(2),160-180.
- Karale A. The Challenges of IoT Addressing Security, Ethics, Privacy, and Laws, *Internet of Things*, 2021, 15:100420
- Mocrii, D., Chen, Y., Musilek, P. (2018) IoT-based Smart Homes: A Review of System Architecture, Software, Communications, Privacy and Security ,*Internet of Things*, 1–2:81-98.
- Nikou S. (2019) Factors driving the adoption of smart home technology: An empirical assessment, *Telematics and Informatics*, 45:e10128.
- Nunes F, Verdezoto N, Fitzpatrick G, Kyng M, Grönvall E, Storni C (2015). Self-care technologies in HCI: Trends, tensions, and opportunities. *ACM Trans Comput Interact* 22, 33:1–33:45
- O'Keefe K. and O. Brien, D. (2018). *Ethical Data and Information Management: Concepts, Tools and Methods* (1st. ed.). Kogan Page Ltd., GBR.
- Pigini, L., Bovi, G., Panzarino, C., Gower, V., Ferratini, M., Andreoni, G., Sassi, R., Rivolta, M.W., Ferrarin, M. (2017). Pilot test of a new personal health system integrating environmental and wearable sensors for telemonitoring and care of elderly people at home (SMARTA project). *Gerontology*, 63: 281-286
- Vallor S (2016) *Technology and the Virtues*. Oxford University Press, New York.
- Yao, Y., Basdeo, J. R., Kaushik, S., & Wang, Y. (2019). Defending my castle: A co-design study of privacy mechanisms for smart homes. In *CHI 2019 - Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (Conference on Human Factors in Computing Systems - Proceedings)*. ACM.

The rights of caregivers to their personal data in the context of disability services

Anne-Marie Tuikka^{1[0000-0002-2962-105X]} and Ville Kainu²

¹ University of Turku, Turku, Finland

² Karolinska institutet, Solna, Sweden
anne-marie.tuikka@utu.fi

Abstract. Since the 90', disability services in Finland, one of the Nordic welfare states, have been developed following the support paradigm. Services for people with disabilities vary according to their age, type of diagnosis and other factors. This study focuses on services targeted for children with disabilities and their caregivers as these services, alongside other social services, are going through process of digitalization. Hence, it would be necessary to have clear definition of personal data in the context of disability services.

Although caregivers have essential role in the lives of their children, their role is often neglected when digitalization of disability services is discussed. In this study they are given the spotlight as the rights related to personal data are approached from the viewpoint of caregivers. Possible conceptualization for this purpose is explored through analyses of both prior research and empirical data.

Keywords: Disability services, Parents, Caregivers, Children, Datenherrschaft, Personal data

1 Extended abstract

Since the 90', disability services in Finland, one of the Nordic welfare states, have been developed following the support paradigm (Saloviita 2005). This paradigm highlights the equality and full participation of people with disabilities in their own communities. To achieve this goal, people with disabilities are entitled to different type of social services. Some of these services are available to all people living in Finland, while others are dedicated for people with disabilities called disability services. Services for people with disabilities also vary according to their age, type of diagnosis and other factors. For the purposes of this study, focus is on the services targeted for children with disabilities and their caregivers.

Although caregivers always have essential role in the lives of their children, this is especially true in the case of children with disabilities. For example, they are representatives of their child for applying both health and social services. In addition, the caregiver might need to participate in the rehabilitation initiatives and provide assistive technologies. Alongside these duties, caregivers themselves might need support to maintain their wellbeing and financial resources for being able to care for

their child. These responsibilities as well as needs of caregivers themselves have been acknowledge in the Finnish disability services which aim to support and empower caregivers. However, the child with disabilities is always considered as the applicant of these services. Hence, caregivers can be considered to be secondary customers in relation to the disability services. The concept of secondary customers was first used in business literature but have later been introduced to the context of care services by Leino (2017). While her study focused on family members of people who have needed health care services, the concept of secondary customers appears to fit also to the context of disability services.

Disability services alongside other social services are going through process of digitalization in Finland. This process is closely linked to digitalization of health care in Finland, which have resulted in national data bank – called Kanta – combining patient data from all health care providers and offering patients a possibility to decide about the use of their patient data across service providers. Since 2021, social care providers are also required to join this data bank and gradually integrate more services to this (Finnish institute for health and welfare). These advances in digitalization have also had impact on legislation regarding personal data that is collected on clients of social services.

In research literature in the field of information systems as well as in the field of IT ethics, personal data in the context of social care have not been defined as thoroughly as personal data in the context of health care. However, definitions can be found from governmental sources. For example, Information Commissioner's Office in UK defines social work data as personal data that is processed in connection with social services function or to provide social care (Information Commissioner's Office). In Finland, similar term is not used although special regulation is applied to personal data produced or collected in relation to providing social services (Office of the data protection ombudsman).

Because it is the caregiver, not a child, who applies disability services and often also interacts with professionals at disability services personal data about caregivers is collected to information systems used in organizations offering disability services. Our study focuses on information systems used in one public service provider which offers services for citizens of certain municipality in Finland. Research data was collected through interviewing employees who work with caregivers and their managers in 2017–2018. Research methods included both semi-structured interviews and focus groups. The analysis of these interviews revealed that information systems used in this organization only recognized the child as a client. However, employees of disability services need to know a lot about the caregivers to make decisions regarding the service needs of the child. Hence, personal data of the caregiver is added to client data of the child. This raises a question, how the rights of caregivers toward their personal data are respected in the context of disability services.

To answer this question, we will explore the suitability of one prominent conceptualization, called *Datenherrschaft*, in the context of disability services. *Datenherrschaft* was initially designed by Koskinen (2016) to find ethically justified definition to represent the rights of people for their personal data in the context of health care. This concept refers to possession of and mastery over patient data granting each

individual the legal right to decide the use of their patient data. However, patient data can in some cases be vital for another person's well-being. Hence, individual's rights to their patient data cannot be absolute. Public healthcare providers should in certain, ethically justified situations be able to utilize this information without the consent of the individual. Still healthcare providers nor their employees should not be granted a mastery over patient data. Their rights are limited to be used to protect health and life of others, while individual who has mastery over their own patient data should have possibility to know how and why their patient data is accessed and used.

Koskinen, Kainu and Kimppa (2016, 83) have defined few instances in which someone's patient data can be used by health care professionals to "safeguard health of others". For example, preventing an epidemical disease from spreading could justify why medical professionals would have right to see and use patient data of an individual without their consent. However, having such right in a particular situation does not give mastery over patient data to health care professionals – it is still hold by the individual of whose patient data is in question. Similar situations could be defined in the context of social care, in which wellbeing of individuals is often interconnected.

Understanding if this definition of Datenherrschaft is fitting to represent the rights of caregivers to their personal data collected by disability services is at the core of this study. Although Datenherrschaft has initially been designed in the context of health care to define the rights of people to their patient data it appears to have explanatory value also in the context of social care. However, further research is needed to assure this.

References

- Finnish institute for health and welfare. Uusi asiakastietolaki tulee voimaan 1.11.2021. <https://thl.fi/fi/web/tiedonhallinta-sosiaali-ja-terveysalalla/-/uusi-asiakastietolaki-tulee-voimaan-1.11.2021>, last updated 27.8.2021
- Information Commissioner's Office. Social work data. <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/right-of-access/social-work-data/>
- Koskinen, J. (2016). Datenherrschaft – An ethically justified solution to the problem of ownership of patient information. University of Turku, Turku School of Economics, Series A.
- Koskinen, J., Kainu, V., Kimppa, K. (2016). The concept of Datenherrschaft of patient information from a Lockean perspective. *Journal of Information, Communication and Ethics in Society*, 14/1, 70-86.
- Leino, Henna M. (2017). Secondary but significant: secondary customers' existence, vulnerability and needs in care services. *Journal of Services Marketing*, 37/7, 760-770.
- Office of the data protection ombudsman. Usein kysyttyä sosiaalihuollosta. <https://tietosuoja.fi/usein-kysyttya-sosiaalihuolto>
- Saloviita, T. (2005). Paradigms of disability services in Finland. Resistance, reflection and change, 47-57.

Snooping, Stalking, Breakup and Letting Go - Examining Emerging Norms, Values, and Technological Drivers around Relationships & Break-Ups online.

Aara J. Cho¹ and Wilhelm E. J. Klein²

¹ Chinese University of Hong Kong, Hong Kong

² Hong Kong Baptist University, Hong Kong

Abstract. The proposed paper aims to examine current behaviors, trends, experiences, and observations connected to relationships online, focusing on emotionally charged situations such as snooping, stalking, breakup, and letting go during and post relationships. We propose to review a collection of papers examining these occurrences from different perspectives – from technology design, psychology, and sociology for research input. Based on these, we provide a descriptive and partially normative review of online breakup and grief behavior and its natural and "unnatural" evolution.

Keywords: ethics, emerging values, social media, breakup, moral psychology

1 Introduction

With social media, humanity has been signed up for the largest psychological experiment ever conducted, and we have no idea what the consequences and outcomes will be. This notion, or similar ones to the same effect, have long been part of academic debate. In the public sphere, this idea is gaining more traction too – e.g., recently as a central point of the viral Netflix documentary *The Social Dilemma* (Orlowski, 2020) and the *Facebook files* published by the Wall Street Journal, detailing internal documents from Facebook, leaked by whistleblower Frances Hauge (Horwitz, 2021). In both cases, ex-corporates working for large social media companies took to the public to expose what they perceived as dangers to users and society, resulting from the intentional and unintentional mechanisms of their products and design choices made by these companies. The shared goal of virtually all social media is, of course, to "drive and maintain engagement", which essentially translates to successfully nudging users to log into social media, consume and generate content. The purpose of this is to harvest behavior data which can then be sold to advertisers and other parties interested in purchasing the ability to change users' behavior in their favor.

With this underlying structure in mind, our proposed paper aims to examine social media users' attitudes, perspectives, and behavior during and after relationships, as influenced or driven by social media use. Questions like 'is it ok for me to check my partner's messages', 'is it healthy to follow your partner's social media account', 'should I feel guilty about stalking my ex on social media', 'why can't I let go of my ex connections', 'should I delete pictures of my ex from social media?' etc. will be examined in light of what they disclose about absent and emerging norms and values, as well as the mechanisms which may drive or hinder certain behaviors online.

Specifically, we will examine the following phenomena:

- a) **Snooping:** the action of logging into a partner's social media accounts and/or messenger applications, emails etc., without their consent to find out information about their private affairs.
- b) **Stalking:** the action of following a love interest, partner, or ex-partners social media accounts and online presence in an obsessive, exaggerated manner.
- c) **Breakup:** the action of severing romantic ties with a partner as manifested in an online context (e.g. unfriending, blocking, etc.)
- d) **Letting Go:** the act of psychologically moving on from a romantic relationship, not in the sense of deleting experiences (or evidence thereof), but rather in reclaiming one's power and recovering one self's emotional equilibrium, vitality, and self-worth.

To examine these, we propose to review academic research in technology design, psychology, and the social sciences, and materials from popular media, including articles, podcasts, YouTube videos addressing these issues.

2 Relevant literature & material

The following is a selection of the relevant literature we are reviewing for the proposed paper.

In *Cheating, Breakup, and Divorce: Is Facebook Use to Blame?* Clayton et al. examined the correlation between the use of social media and its impact on personal relationships by surveying the levels of Facebook use and adverse relationship outcomes for Facebook users. Their results suggest that high levels of Facebook use tend to lead to more negative relationship outcomes such as a breakup, divorce, emotional and physical cheating. Highly relevant for the proposed paper, Clayton et al. identify several mechanisms and behaviors, such as maintaining contact with ex-partners or checking photos of ex-romantic partners enabled by Facebook, as significant causes for conflict in the current relationships (Clayton et al., 2013).

Haimson et al. in *Relationship breakup disclosures and media ideologies on Facebook* shed light on how Facebook users disclose their breakup and post-

relationship on Facebook. One of their key findings is that there appears to be little to no common ground for users' behavior or process for online breaking up and after a breakup. They connect this to the concept of *media ideologies*, as introduced by Ilana Gershon (Gershon, 2010, 2011). According to Gershon, people who use social media assume that they would react the same way at the end of a relationship. However, in reality, most people tend to hold very diverse opinions and perspectives about appropriate post-breakup communication and behavior (Haimson et al., 2018). This paper clearly pertains to the proposed paper's 'ongoing evolution of norms' hypothesis.

McDaniel et al.'s *Are You Going to Delete Me? Latent Profiles of Post-Relationship Breakup Social Media Use and Emotional Distress* examined how people are engaged in social media after relationships (McDaniel et al., 2021). This includes users' monitoring, interacting and deleting posts and photos, deleting the partner's family/friends, stopping social media use, and keeping *digital possessions*. They classify social media behaviors post-breakup into four different stages; clean breakers, wistful reminiscers, ritual cleansers and impulsives – once again presenting a picture of highly diverse notions of dos and don'ts after the end of a relationship.

Several other relevant papers have already been identified to be examined and incorporated in the proposed paper. A selection is as follows: (Kanakaris et al., 2018), (Arikewuyo et al., 2021), (Lukacs & Quan-Haase, 2015), (McPeak, 2014), (Sas & Whittaker, 2013), (Seraj et al., 2021), (Reed et al., 2016).

3 Analysis and reflection

Rather than a specific philosophical lens, the proposed paper will make use of perspectives from multiple ethical frameworks, including Verbeek's persuasive technology approach (Verbeek, 2006), Friedman's value-sensitive design (Friedman & Hendry, 2019), and a general utilitarian perspective maximizing well-being for all involved parties. Further, specific attention will be paid to moral psychology and how inherent and culturally evolved moral intuitions are accentuated, dampened, manipulated, or inhibited by technological mediation. We are particularly keen to discuss the extent to which users appear to be in a situation where common sense, moral norms and rules and ways of managing circumstances do not seem to apply or seem to be challenging to translate to the online platform. So, for example, where the notion that it is unhealthy to keep seeking out one's ex-partner in real-life, to observe their social interactions and new love interests seems rather uncontroversial, the same activity is not nearly as clearly identified as unhealthy or problematic on social media. While offline, maintaining strong ties with an ex-partner's extended family is at least rather uncommon, social media users publicly wonder whether they should unfriend their ex-partner's family – or block their access to their own social media. In many ways, it seems, we are finding ourselves in a period of cultural evolution, where we are collectively negotiating, probing and testing new norms and rules for social life and relationships – an evolution which is profoundly influenced and at least partly shaped by those controlling the online platforms we inhabit.

This being said, the paper aims not to find definitive answers to how to design or not design social media. Nor is it an attempt to provide definitive normative advice on the best handling of relationships and breakups online. Instead, it seeks to provide a critical examination and exhibition of a range of questions worth asking and pursuing further research about.

Citations and references

- Arikewuyo, A. O., Eluwole, K. K., & Özad, B. (2021). Influence of Lack of Trust on Romantic Relationship Problems: The Mediating Role of Partner Cell Phone Snooping. *Psychological Reports*, 124(1), 348–365. <https://doi.org/10.1177/0033294119899902>
- Clayton, R. B., Nagurney, A., & Smith, J. R. (2013). Cheating, Breakup, and Divorce: Is Facebook Use to Blame? *Cyberpsychology, Behavior, and Social Networking*, 16(10), 717–720. <https://doi.org/10.1089/cyber.2012.0424>
- Friedman, B., & Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press.
- Gershon, I. (2010). *The Breakup 2.0: Disconnecting over New Media*. Cornell University Press. <http://ebookcentral.proquest.com/lib/cuhk-ebooks/detail.action?docID=3137991>
- Gershon, I. (2011). Un-Friend My Heart: Facebook, Promiscuity, and Heartbreak in a Neoliberal Age. *Anthropological Quarterly*, 84(4), 865–894. <http://dx.doi.org/10.1353/anq.2011.0048>
- Haimson, O. L., Andalibi, N., De Choudhury, M., & Hayes, G. R. (2018). Relationship breakup disclosures and media ideologies on Facebook. *New Media & Society*, 20(5), 1931–1952. <https://doi.org/10.1177/1461444817711402>
- Horwitz, J. (2021, October 1). The Facebook Files. *Wall Street Journal*. <https://www.wsj.com/articles/the-facebook-files-11631713039>
- Kanakaris, V., Tzovelekis, K., & Bandekas, D. (2018). *Impact of AnonStalk (Anonymous Stalking) on users of Social Media: A Case Study*. [https://www.semanticscholar.org/paper/Impact-of-AnonStalk-\(Anonymous-Stalking\)-on-users-a-Kanakaris-Tzovelekis/409ad58d8b5c1c5d248876f7f1141c4bcca3148b](https://www.semanticscholar.org/paper/Impact-of-AnonStalk-(Anonymous-Stalking)-on-users-a-Kanakaris-Tzovelekis/409ad58d8b5c1c5d248876f7f1141c4bcca3148b)
- Lukacs, V., & Quan-Haase, A. (2015). Romantic breakups on Facebook: New scales for studying post-breakup behaviors, digital distress, and surveillance. *Information, Communication & Society*, 18(5), 492–508. <https://doi.org/10.1080/1369118X.2015.1008540>
- McPeak, A. (2014). Social Media Snooping and Its Ethical Bounds. *Arizona State Law Journal*, 46(3), 845–898.
- Orlowski, J. (2020, January 26). *The social dilemma*. Netflix.
- Reed, L. A., Tolman, R. M., & Ward, L. M. (2016). Snooping and Sexting: Digital Media as a Context for Dating Aggression and Abuse Among College Students. *Violence Against Women*, 22(13), 1556–1576. <https://doi.org/10.1177/1077801216630143>

- Sas, C., & Whittaker, S. (2013). Design for forgetting: Disposing of digital possessions after a breakup. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1823–1832. <https://doi.org/10.1145/2470654.2466241>
- Seraj, S., Blackburn, K. G., & Pennebaker, J. W. (2021). Language left behind on social media exposes the emotional and cognitive costs of a romantic breakup. *Proceedings of the National Academy of Sciences*, 118(7). <https://doi.org/10.1073/pnas.2017154118>
- Verbeek, P.-P. (2006). Persuasive Technology and Moral Responsibility Toward an Ethical Framework for Persuasive Technologies. *Persuasive*, 6, 1–15.

The ethical dilemma while using technology to mitigate covid-19 pandemic restrictions

Ala Ali Almahameed¹[0000-0002-4381-0316], Jorge Pelegrín-Borondo²[0000-0003-2720-1788],
Jorge de Andrés Sánchez¹[0000-0002-7715-779X], Orlando Lima Rua³[0000-0002-1593-7440], Maria
Alesanco Llorente²[0000-0001-9142-1203], Alba García Milon²[0000-0003-0367-8012] and Mario
Arias-Oliva^{1,4}[0000-0002-6874-4036]

¹Universitat Rovira i Virgili, Tarragona, Spain

²La Rioja University, Logroño, Spain

³Polytechnic of Porto, Porto, Portugal

⁴Complutense University of Madrid, Madrid, Spain

a.mahameed82@gmail.com

EXTENDED ABSTRACT

The world is still experiencing the negative impact of the Covid-19 pandemic, although nearly two years have passed since the emergence of the virus in the Chinese city of Wuhan. Also, different variants are emerging from time to time with various spread speeds, symptoms, and impacts. Based on the global statistics on Covid-19 and up to the beginning of 2022, the number of infections reached 294,558,279, of which 5,471,212 deaths and 255,849,216 cases of recovery from the disease were recorded (Worldometer, 2022). Many countries have resorted to full or partial lockdown to confront the virus since the beginning of the pandemic, which in turn impacts social and economic activities in different ways and levels (Ozili & Arun, 2020). It caused millions of workers to lose their jobs across the world, besides the largest decline for the quarterly Gross Domestic Product (GDP) since the Great Depression, which took place between 1929 and 1932 (Koller, 2020). For instance, the International Air Transport Association (IATA) reported a drop in international passengers' demand by 75.6% below 2019 demand (IATA, 2021). In the USA, the GDP was reported the first annual decline for 2020 (3.5%) since the financial crisis in 2007, and it is considered the lowest drop since 1946 (Mutikano, 2021). And in Europe, the unemployment rate

is increased by 7.3% for January 2021 if compared to the unemployment rate in January 2020, which was 6.6% (European Commission, 2021).

Restricting the mobility of individuals has been considered an effective way to minimize and control the spread of Covid-19. These restrictions could involve social distancing, partial or complete lockdown, closing public transportations and borders, and working from home (Zhou, 2020). For instance, Fang et al. (2020) expected in their study that the lockdown, which was imposed in the Chinese city of Wuhan, reduced the outbreak of Covid-19 in the entire country by almost 58.7 %. In the same context, Carteni, Francesco, and Martino (2020) found that the number of new Covid-19 cases in Italy is related directly to the mobility of individuals across the country. They claimed that as much as travel between cities across the country is reduced, the total number of new cases will be reduced. Moreover, Nouvellet et al. (2021) studied the impact of mobility on Covid-19 transmission for 52 countries and by using mobility data from Apple and Google. Their results showed that mobility is correlated with the intensity of Covid-19 transmission and it is changing over time. Also, it indicated that for 73% of analyzed countries, the reduction in mobility is reducing transmission of Covid-19. Furthermore, Linka, Goriely, and Kuhl (2021) have studied the impact of global and local mobility on the Covid-19 outbreak. They mentioned that global mobility is correlated with the Covid-19 pandemic outbreak, especially for zero or low cases. And the impact of reducing local mobility could reduce new cases, especially when the outbreak becomes wider. And as a result, it could cause flattening the epidemic curve, and toward declining in a later stage.

Local and international health authorities and institutions have used the technology to facilitate the planning and response strategies in confronting the pandemic in a way that mitigates infections and minimizes negative consequences. For instance, the maps applications have been used to track people's movement, especially for infected and quarantined people. Also, China has used these maps to track people who had visited the city of Wuhan, where these data had been used to forecast the transmission nature of Covid-19 and guide surveillance and borders check (Hu et al., 2020). In the USA, healthcare institutions have developed a digital platform to report real-time data of the number of infected cases, the status of healthcare facilities, percentage use of ventilators, available resources, and so on (Becker's Health Information Technology, 2020). Besides, global websites, which are offering real-time data of the number of daily cases, death, and recovery on a country basis. In the same context, some researchers identified five digital technologies that have been used in health care activities during the covid-19 pandemic, which are machine learning, computing devices, natural languages processing, online databases, and mobile phones. These activities include surveillance, tracking, identification, and evaluation of the intervention. And they considered ethical dilemmas as the major barriers to implementing such technologies during the Covid-19 pandemic (Vargo et al., 2021).

With the emergence of different vaccines, different countries have taken incentive procedures to encourage citizens and residents of their territories to take the vaccine, as it represents the fastest and safest solution to this pandemic so far. For instance, some European countries are allowing vaccinated travelers who are permitted to visit Europe to land in their countries without a need for quarantine (McMahon, 2021). In

the same context, different countries and international technologies have been integrated technology to facilitate people mobility, such as the Covid-19 passport (or Vaccine Passport), which has been used at the local and international levels, by exempting people who have been vaccinated or have negative test results of the virus from the restrictive measures that are imposed by governments (Fiocchi & Jensen-Jarolim, 2021). It has been used as a free mobile application where local and international travelers can attach their Covid-19 test results or other health waivers, as well as proof of vaccination (McMahon, 2021). For instance, the UK government has approved the use of the vaccine passport as a mobile app, in order to mitigate the lockdown in the country. Besides, it could be used to encourage people to get the Covid-19 vaccine. As the government stated, it will be easy for holders of this app to prove their negative test results of Covid-19, or to prove that they have been vaccinated (Lawrie, 2021). For Chinese citizens who are traveling overseas, the vaccination status and Covid-19 test results can be shown on the Chinese social media apps WeChat, which is launched on March 8, 2021. In the same context, the IATA has introduced a digital platform for passengers to be used as a travel pass, which includes all medical information related to the Covid-19 tests and vaccinations. IATA has justified the need for the travel pass in which it will offer accurate information on passengers' health status regarding Covid- 19 (IATA, 2021).

Not all are supporting the use of the vaccine passport. They are opposing the vaccine passport from an ethical perspective, considering potential problems such as discrimination, data privacy, and freedom of movement. For instance, potential discrimination between the people who already have the vaccine, and the people who haven't could be involved in terms of classifying people based on their Covid-19 status. Also, the freedom of movement is another concern that could become a problematic issue, especially for people who would not be able to take the vaccine because of health constraints. This could imply preventing individuals who haven't got the vaccine from international travel or restricting their domestic mobility and access to different local places (Kofler, 2020). Furthermore, there could be some financial consequences that are associated with the need to do the Covid-19 test frequently for those who do not want to take the vaccine or have a medical contraindication. As well, the timeline that the world will need in order to ensure that vaccination reaches all countries is another critical concern. Despite that most western and wealthy countries have vaccinated most of their citizens, most of the African countries will be able to offer the vaccine to their people not before 2023. And the rest could be able to vaccinate their citizens in the mid of 2022 (Buchholz, 2021). This could reinforce the ethical concern regarding the ability of everyone to get the vaccine passport.

The study aims to investigate the impact of technology in mitigating covid-19 negative impacts. In this regard, and as mentioned above, the vaccine passport has been introduced to encourage people to get the vaccine and to facilitate their local and international mobility. However, the ethical impact on the attitude toward using vaccine passports shall be considered. Accordingly, the paper will investigate the impact of ethical dimensions on attitude toward using the vaccine passport by adopting the Multidimensional Ethics Scale (MES), which has been used by Pelegrín-

Borondo, Arias-Oliva, Murata, and Romero (2018) to determine the influence of ethical judgment on intention to become a cyborg. In this way, we believe that the finding of this research will contribute to the academic, business, and social fields by understanding the ethical impact on the attitude toward using the vaccine passport, and how the technology supports world efforts in the fight against the Covid-19 pandemic.

Keywords: Ethics, Covid-19, Vaccine passport, MES scale, Covid-19 Vaccine

Acknowledgements

This research was supported by Telefonica and its Telefonica Chair on Smart Cities of the Universitat Rovira i Virgili and the Universitat de Barcelona (project number 42.DB.00.18.00)

REFERENCES

- Becker's Health Information Technology. Swedish health services tap Microsoft to build app that tracks COVID-19 patients and hospital capacity.
<https://www.beckershospitalreview.com/healthcare-information-technology/swedish-health-services-taps-microsoft-to-build-app-that-tracks-covid-19-patients-hospital-capacity.html> (accessed January 05, 2022).
- Buchholz, K. (2021). Chart: Global Vaccine Timeline Stretches to 2023. Retrieved June 9, 2021, from <https://www.statista.com/chart/24064/covid-19-vaccination-timeline-global/>
- Carteni, A., Francesco, L. Di, & Martino, M. (2020). How mobility habits influenced the spread of the COVID-19 pandemic: Results from the Italian case study. *Science Of the Total Environment*, 741, 1–9. <https://doi.org/10.1016/j.scitotenv.2020.140489>
- European Commission. (2021). EU Digital COVID Certificat. Retrieved June 5, 2021, from https://ec.europa.eu/info/live-work-travel-eu/coronavirus-response/safe-covid-19-vaccines-europeans/eu-digital-covid-certificate_en
- Fang, H., Wang, L., & Yang, Y. (2020). Human Mobility Restrictions and the Spread of the Novel Coronavirus (2019-nCoV) in China. *SSRN Electronic Journal*, 191. <https://doi.org/https://doi.org/10.1016/j.jpubeco.2020.104272>
- Fiocchi, A., & Jensen-Jarolim, E. (2021). SARS-COV-2, can you be over it? *The World Allergy Organization Journal*, 14(2), 100514. <https://doi.org/10.1016/j.waojou.2021.100514>
- Hu, Zixin, Ge, Qiyang, Li, Shudi, Jin, L., & Xiong, Momiao. (2020). Artificial intelligence forecasting of covid-19 in china. *arXiv preprint arXiv:2002.07112*.
- IATA. (2021). IATA Travel Pass Initiative. Retrieved June 9, 2021, from <https://www.iata.org/en/programs/passenger/travel-pass/>

- Kofler, N., & Baylis, F. (2020, May 28). Ten reasons why immunity passports are a bad idea. *Nature. Nature Research*. <https://doi.org/10.1038/d41586-020-01451-0>
- Koller, T. (2020). Why have stock markets shrugged off the COVID-19 crisis? Retrieved June 9, 2021, from <https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/the-strategy-and-corporate-finance-blog/why-have-stock-markets- shrugged-off-the-covid-19-crisis#>
- Lawrie, E. (2021). What is a Covid passport and what are the UK's plans? Retrieved June 9, 2021, from <https://www.bbc.com/news/explainers-55718553>
- Linka, K., Goriely, A., & Kuhl, E. (2021). Global and local mobility as a barometer for COVID- 19 dynamics. *Biomechanics and Modeling in Mechanobiology*, 20(2), 651–669. <https://doi.org/10.1007/s10237-020-01408-2>
- Lucia, Mutikano (2021). COVID-19 savages U.S. economy, 2020 performance worst in 74 years. Retrieved June 9, 2021, from <https://www.reuters.com/article/us-usa-economy-idUSKBN29X0I8>
- McMahon, S., & Sampson, H. (2021). Vaccine passports: Everything travelers need to know - *The Washington Post*. Retrieved June 9, 2021, from <https://www.washingtonpost.com/travel/2020/12/08/vaccine-passport-immunity-app-covid/>
- Nouvellet, P., Bhatia, S., Cori, A., Ainslie, K. E. C., Baguelin, M., Bhatt, S., ... Donnelly, C. A. (2021). Reduction in mobility and COVID-19 transmission. *Nature Communications*, 12(1), 1–9. <https://doi.org/10.1038/s41467-021-21358-2>
- Ozili, P., & Arun, T. (2020). Spillover of COVID-19. *SSRN Electronic Journal*, (March 2020), 27.
- Pelegrín-Borondo, J., Arias-Oliva, M., Murata, K., & Souto-Romero, M. (2018). Does ethical judgment determine the decision to become a cyborg? Influence of Ethical Judgment on the Cyborg Market. *Journal of Business Ethics*, 161(1), 5-17. <https://doi.org/10.1007/s10551-018-3970-7>
- Vargo, D., Zhu, L., Benwell, B., & Yan, Z. (2021). Digital technology use during COVID-19 pandemic: A rapid review. *Human Behavior and Emerging Technologies*, 3(1), 13-24. DOI: 10.1002/hbe2.242
- Worldometer. (2022). COVID live update. Retrieved Jan 9, 2022, from <https://www.worldometers.info/coronavirus/>
- Zhou, Y., Xu, R., Hu, D., Yue, Y., Li, Q., & Xia, J. (2020). Effects of human mobility restrictions on the spread of COVID-19 in Shenzhen, China: a modelling study using mobile phone data. *The Lancet Digital Health*, 2(8), e417–e424. [https://doi.org/10.1016/S2589-7500\(20\)30165-5](https://doi.org/10.1016/S2589-7500(20)30165-5)

Care and Data: How can we use healthcare data ethically?

Ryoko Asai

Ruhr University Bochum, Bochum, Germany

ryoko.asai@rub.de

Abstract. Nowadays, we have been facing demographic changes caused by the aging of the population. Sweden and Japan are no exceptions in this respect. Living longer does not mean just only adding years to life expectancy rates. Aging should be a process where quality of life and well-being are instigated and maintained. For older people, the aging process gives them more experiences and knowledge about a life by themselves and also in their community and society. Especially in the aged society, being healthy and living independently will be very important to have a sufficient level of quality of life and establish a sustainable and participatory relationship with communities and society. In the past two years, the corona pandemic has affected older people in many aspects of their lives, and new technologies have contributed to elderly support greatly. Most of these emerging technologies utilize data collected by public authorities and also by different kinds of organizations including industries, when designing and developing new digital tools for older people. On the other hand, elderly people are particularly reluctant to adopt digital health technologies such as mobile health tools. Under the contradicting situation, what can the collected data stand for in developing a digital tool? This study aims to delineate what “data” means in developing digital technology to support older people, from the perspective of computer ethics.

Keywords: aging society, care, data in healthcare, digital technologies, ethical use

Nowadays, we have been facing demographic changes caused by the aging of the population. Sweden and Japan are no exceptions in this respect. Life expectancy in both countries is very high: about 83 years in Sweden and 84 years in Japan (The World Bank, 2021)¹. Demographic aging and other structural changes derived from aging will push up the social expenditures such as medical costs and public pension spending (OECD, 2021). For example, in Sweden, the social services act handles elderly care and the municipalities mainly take the responsibility for elderly care. That means the major cost of elderly care is financed by the municipalities and the government. Older people can get elderly care service from public and private sectors as they choose. But still those services are distributed under the responsibility

of the municipalities. Therefore, aging of society would strain the resources of the municipalities and ask them to take responsibility for more people's health and life.

In 2040, around 25 percent of the Swedish population will be 65 years or older, and most of the people in this age group will be active and healthy (Swedish Institute, 2021). Being healthy and living independently is very important for older people to have a sufficient quality of life and enhance their wellbeing. Older people pay a great attention to self-care to maintain health, prevent disease and cope with illness, as well as participate in society. However, they are commonly reckoned as reluctant to adopt new technologies and are seldomly included in the development of new technologies (Klaver et al., 2021). Therefore, it is indispensable or significant/critical to ask: How can emerging technologies support elderly's self-care and social participation? Corresponding to these ideas and questions, in the winter 2021, we kicked off the project which focuses on elucidating older people's needs and wants regarding selfcare and social participation, and identifying available digital tools to support their needs and wants. In the conference, although the project was just kicked off and is still on-going, we try to explore some ethical concerns about developing and using emerging technologies and data for older people's healthcare from the perspective of information ethics.

In the past two years, the Corona pandemic has affected older people in many aspects of their lives, from being prevented of social interaction to added responsibility on self-care and management of diseases. However, amid growing healthcare interests, many elderly people do not have enough knowledge to determine what is good to do as self-care for health by themselves. Although media provides abundant information about what is good for health, ironically, too much information can disturb their ability to take assertive decisions. In a similar way, the access to technology through mobile phones, applications, and wearable devices can be overwhelming. Given these circumstances, it is significant to delve into the needs and wants of self-care from the elderly people, and delineate effective elderly supports by the communities and new technologies.

Ongoing recent trends in personalized medicine have highlighted the importance of research in ethical, legal, and social issues (ELSI)². The project aims to apply the Ethical, Legal and Social Issues (ELSI) framework to create awareness of the ethical and social effects of emerging technologies, in order to make these technologies privacy-oriented, socially valued and culture-sensitive. Also, the previous research shows that one of the main reasons to reduced digital adoption by older people is 'perceived risks of digital health tool use', such as privacy risk, performance risk, legal concern, and trust (Klaver et al., 2021). In order to mitigate these perceived risks, the ELSI approach would be useful and also give older people some degree of security when they use digital healthcare tools and inscribe their personal and health information digitally.

In this study, as a groundwork to apply the ELSI framework, we try to elucidate visible and invisible ethical problems over data used for older people's healthcare including self-care. In order to comprehend the aging society and older people's situation, public authorities and many industries collect data from older people, their families, care givers and other stakeholders related to them. The collected data by the

authorities would initiate policy discussions and influence decision making processes of medical-healthcare systems, social care for older people and so on. Also, many companies in different industrial areas make their business strategies to gain the profit by targeting older people and/or their families and offering different services. Furthermore, AI-based medical care and healthcare systems and tools are functioned through machine learning based on tremendous amount of data. Data has become absolutely essential in daily medical/health care activities nowadays.

Although data contributes to our comprehensive understanding of older people and their lives, it is not always applicable to individual cases. Because, data shows overall figures of the target and many policies and machines function in the one-size-fits-all way. Simply put, when politics and industry speak of elderly people based on the collected data, they reflect datafied elderlies on their talks. Then, how much do datafied elderlies represent the real human elderlies and their actual life? (Nafus ed., 2016; Barassi, 2020) Moreover, regarding care, it is considered as an interaction between human and human, also between human and technology. Many social workers and care givers use the data base for recording health conditions of persons in need of care, and retrieve, sometimes share with other care givers, the data for their care activities. Again, the same concern comes up: Some care-receivers cannot give care takers or data collectors what/how they feel and think. There are many cases that care givers inscribe information instead of the care receivers. Then, how much can we trust data for the older people's care, especially those who cannot express their actual situations? How can we make data-driven digital tools trustworthy? Is it not perceived as the dataveillance of the older people? (Lupton and Williamson, 2017) If no, then, why and how is it different from the dataveillance? You may say, because it is used for care. Then, is there a risk that technology may amplify the destructive characteristics of care? (Puig de la Bellacasa, 2017)

Additionally, in terms of developing a new digital application and increasing the profit, companies seldom concern older people as a target user, rather develop a digital product for the younger generation who are used to digital technologies. The technology industry that are in the front line of innovation but where research knowledge lacks to better address the needs of older people regarding self-care and use of digital tools. This study aims to elaborate ethical concerns over data and AIbased/data-driven digital tools regarding the elderly care from the information ethics aspect. Also, we will look into a possibility to develop safe and/or trustworthy machines through reconsidering the meanings of security and trust.

Notes

1. This research project is conducted in Sweden and Japan. Due to this research background, we mention about ageing and society by referring Sweden and Japan.
2. Please see the details of Ethical, Legal and Social Issue approach in EUPATI's website: <https://toolbox.eupati.eu/resources/ethical-social-and-legal-issues-elsi-in-hita/>

Acknowledgements

This research was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant-in-Aid for Scientific Research (C) No. JP20K12551 “Social Robots and Children’s wellbeing.” and also by Vinnova Academia/Industry collaborations Reference Number 2021-04992 “Designing self-care for increased health of older people in the digital age”.

References

- Barassi, V. (2020). *Child|Data|Citizen: How Tech Companies Are Profiling Us from before Birth*, The MIT Press.
- Klaver, N., van de Klundert, J., van den Broek, R., & Askari, M. (2021). Relationship Between Perceived Risks of Using mHealth Applications and the Intention to Use Them Among Older Adults in the Netherlands: Cross-sectional Study. *JMIR Mhealth Uhealth* 2021;9(8):e26845, DOI: 10.2196/26845
- Lupton, D. and Williamson, B. (2017). The Datafied Child: The Dataveillance of Children and Implications for Their Rights, *New Media & Society*, Vol.19 (5), pp.780-794.
- Nafus, D. ed. (2016). *Quantified: Biosensing Technologies in Everyday Life*, The MIT Press.
- OECD (2021), Social spending (indicator). doi: 10.1787/7497563b-en
- Puig de la Bellacasa, M. (2017). *Matters of Care: Speculative Ethics in More Than Human World*, University of Minnesota Press.
- Swedish Institute (2021), Elderly care in Sweden: Sweden’s elderly care system aims to help people live independent lives. <https://sweden.se/life/society/elderly-care-in-sweden> (Accessed on 31 August 2021)
- The World Bank (2021), Life expectancy (data). <https://data.worldbank.org/indicator/SP.DYN.LE00.IN?view=map&year=2019>

“Be a Pattern for the World”: The Development of a Dark Patterns Detection Tool to Prevent User Loss

Jordan Donnelly¹, Alan Dowley¹, Yunpeng Liu¹, Yufei Su¹, Quanwei Sun¹, Lan Zeng¹, Andrea Curley¹, Damian Gordon¹, Paul Kelly¹, Dymphna O'Sullivan¹ Anna Becevel¹

¹Technological University of Dublin, Ireland

Damian.X.Gordon@TUDublin.ie

Abstract. Dark Patterns are designed to trick users into sharing more information or spending more money than they had intended to do, by configuring online interactions to confuse or add pressure to the users. They are highly varied in their form, and are therefore difficult to classify and detect. Therefore, this research is designed to develop a framework for the automated detection of potential instances of web-based dark patterns, and from there to develop a software tool that will provide a highly useful defensive tool that helps detect and highlight these patterns

1 Introduction

Research on dark patterns covers a range of different fields, including Cognitive Psychology, Usability, Marketing, Behavioural Economics, Design and Digital Media. There is no general agreement that explains their effectiveness, however the traditional decision-making theories, including rational choice theories have been shown to be ineffective in explaining their success (Acquisti, *et al.*, 2017). However, two things that appear to be able to explain some of their effectiveness are *cognitive biases* and *digital nudges*. Cognitive biases are short-cuts (or heuristics) that the human brain makes in decision-making due to the fundamental limitations of the information processing of the brain (Kahneman, 2011). According to Waldman (2020) the five most pervasive are: anchoring, framing, hyperbolic discounting, overchoice, and metacognitive processes such as cognitive scarcity and cognitive absorption. Digital nudges are a manipulation strategy based on the notion that small changes can have a big effect (for example, a personalized email that reminds someone to complete an enrolment form) and nudges are based on the notions of soft paternalism, positive reinforcement and compliance (Acquisti, 2009; Almuhiemedi, *et al.*, 2015). However, unlike dark patterns, nudges can be used either for positive outcomes or negative ones (Peer, *et al.*, 2020).

Chugh and Jain (2021) explored dark patterns from the perspective of consumer protection, as well as their impact on democratic political processes. The researchers make a key distinction between dark patterns and regular advertisements that are persuasive. Their research indicates that dark patterns are deliberately manipulative, whereas persuasive advertisements merely attempt to influence people to revise their preferences. They indicate that there are two major issues with dark patterns, (1) users are typically unaware that they are interacting with dark patterns, and are, therefore, unable to safeguard themselves against their effects, and (2) market forces and market competition are not penalizing organizations for using these patterns. The researchers recommend that legislation and regulations are necessary to combat these patterns.

Bongard-Blanchy *et al.* (2021) looked at the impact of dark patterns on endusers by surveying 406 participants. They found that although all of the respondents were aware of the manipulative techniques that online services use, they are nonetheless unable to combat their impact. The researchers advocate a multi-faceted approach to addressing these issues, including an education programme to explain to people about the different patterns and how they work, as well as providing information on how to resist and avoid these patterns. They also suggest that a combination of strong legal penalties and regulations are required, as well as new software tools to help detect and highlight the existence of these patterns.

Mathur, *et al.*, (2019) undertook a meta-analysis of over 11,000 shopping websites, and created a taxonomy to try to explain how dark patterns affects user decision-making, and their taxonomy has the following characteristics: Asymmetric, Covert, Deceptive, Hides Information, and Restrictive. They found that 11.1% (1254 websites) of the sites had dark patterns, and they recommend the development of plug-ins for browsers to help detect these patterns.

UX researcher Harry Brignull (2011) was a pioneering researcher in this field, and first presented definitions of dark patterns, some of which are:

1. **Sneak into Basket:** Some websites add an additional item into the customer's digital shopping basket, and it is usually the new product that is added in because of a hidden opt-out button or checkbox on a previous page.
2. **Hidden Costs:** Some websites add unexpected charges into the customer's digital shopping basket, e.g. delivery charges, etc.
3. **Trick Questions:** When registering for a new service, some websites present a series of checkboxes, and the meaning of checkboxes is alternated so that ticking the first one means "opt out" and the second means "opt in".
4. **Misdirection:** Some website purposefully focuses users' attention on one thing in order to distract their attention from another, for example, a website may have already undertaken a function and added a cost to it, and the opt out button is small.
5. **Confirmshaming:** Some websites try to guilt the user into opting into doing something, for example, "No thanks, I don't want to have unlimited free deliveries".
6. **Disguised Ads:** Some websites include advertisements that are disguised as other kinds of content or navigation, in order to get you to click on them, for

example, advertisements that look like a “download” button or a “Next >” button.

UX researcher Reed Steiner (2021) identified six types of patterns:

1. **Fake Activity:** Some websites claim that other shoppers are looking at the same products, for example, websites that claim “five other people are viewing this item right now” might not be fully truthful.
2. **Fake Reviews:** Some websites include reviews of products and services that may be fake, and exact matches with different customer names can be found on several sites.
3. **Fake Countdown:** Some websites countdown or timers, and in most cases these timers only add urgency to a sale.
4. **Ambiguous Deadlines:** Some websites indicate that a product is only on sale for a limited amount of time, but don’t mention a specific deadline.
5. **Low Stock Messages:** Some websites claim that they are low on a particular item.
6. **Deceptive High Demand:** This is similar to the low stock messages, but focuses on the high demand for a particular product.

2 Development

The software tool was developed as an overlay onto a browser page, using a combination of Python, NodeJS and Javascript. It is possible to detect six dark pattern types with this tool: Fake Activity, Fake Countdown, Fake Limited-time, Fake Lowstock, Fake High-demand, and Confirmshaming. To detect dark patterns on the webpage, the HTML of the webpage needs to be analyzed first for text extraction. A specific tag type will be assigned to each text string extracted and the pre-processed data will be ready for detection. To detect text in an image, Optical Character Recognition (OCR) is used. OCR is the technology that allows users to detect, analyze and extract texture data from the image file. If image detection is enabled (when OCR is on), the image detection will be conducted first to extract text from the source images and add the extracted text back to the pre-processed dataset for detection. If image detection is not enabled (when OCR is off), the detection will be directly conducted after data pre-processing.

User evaluation and usability testing can be done in many different forms, from a simple questionnaire about the product to letting the users use it and gathering their reactions and thoughts throughout the process using both the think-aloud protocol and cognitive walkthrough, with 96 users. Those users felt the systems was very helpful and clear in its purpose and detection process.

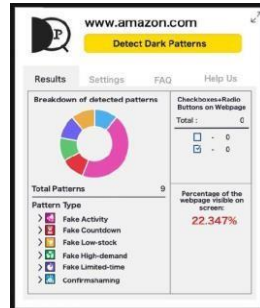


Figure 1. Dark Patterns Detection Tool

Some notable limitations of the study include the following:

1. It might be the case that some of the patterns simply cannot be detected, as they vary so much in implementation. If so, it significantly limits the efficacy of the final system - a thorough exploration of the Mathur *et al.* dataset is needed, as well as a number of further brainstorming sessions, to explore potential solutions.
2. Some sites have a special file called `Robots.txt` that prohibits the use of web scraping, and it is also the case that some sites use technologies that make them more difficult to parse, for example, frames or webpages implemented in Javascript or CSS.
3. Many shoppers use mobile applications instead of websites to purchase products and services, and the techniques outlined so far would be ineffective on these applications.

3 Conclusions

It is worth noting that sometimes the terminology itself can be a barrier, although everyone who has shopped online has experienced dark patterns, nonetheless, the terminology itself may be unfamiliar and therefore confusing, for example the terms “Roach Motel” and “Friend Spam” are opaque as to their meaning and impact (changing “Roach Motel” to “Hard to Unsubscribe”, and changing “Friend Spam” to “May use your addressbook” might a solution. It is also worth noting that the Optical Character Recognition (OCR) aspect of the system was able to read text from images to determine if there were dark patterns on the images, however, on websites that are image-heavy this proved to be prohibitive in terms of overall scan time of webpages, therefore a toggle to turn on and off the OCR features was added. Finally, perhaps one of the most significant outcomes of this research was that it created an opportunity to interrogate our fundamental understanding of the notion of a Dark Pattern. There is a fine line between dark patterns and persuasive advertisements (which do not rely on pressuring or confusing the customers). For example, an advertisement that includes the phrase: “Customers who bought this product also bought ...” were initially

classified as Dark Patterns by the system, as they are similar to a “Fake Activity” which may say something like “Other Customers are looking at this product”, but they are pressuring the customers in the same way.

References

- Acquisti, (2009) “Nudging privacy: The behavioral economics of personal information”. *IEEE Security & Privacy*, 7(6), pp. 82-85.
- Acquisti, I. Adjerid, R. Balebako, L. Brandimarte, L. F. Cranor, S. Komanduri, S. Wilson. (2017) “Nudges for privacy and security: Understanding and assisting users’ choices online”, *ACM Computing Surveys (CSUR)*, 50(3), pp. 1-41.
- H. Almuhiemedi, F. Schaub, N. Sadeh, N.I. Adjerid, A. Acquisti, J. Gluck, Y. Agarwal. (2015) “Your location has been shared 5,398 times! A field study on mobile app privacy nudging”. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pp. 787-796.
- K. Bongard-Blanchy, A. Rossi, S. Rivas, S. Doublet, V. Koenig, G. Lenzini. (2021) "I am Definitely Manipulated, Even When I am Aware of it. It s Ridiculous!--Dark Patterns from the End-User Perspective". *arXiv preprint arXiv:2104.12653*.
- H. Brignull. (2011) “Dark patterns: Deception vs. honesty in UI design”. *Interaction Design, Usability*, 338,
- Chugh, P. Jain (2021)"Unpacking Dark Patterns: Understanding Dark Patterns and Their Implications for Consumer Protection in the Digital Economy". *RGNUL Student Research Review Journal*, 7, 23.
- L. Di Geronimo, L. Braz, E. Fregnan, F. Palomba, A. Bacchelli, "UI Dark Patterns and Where to Find Them: A Study on Mobile Applications and User Perception", *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, April 2020.
- Kahneman. (2011) “Thinking, Fast and Slow”, Penguin Books.
- Peer, S. Egelman, M. Harbach, N. Malkin, A. Mathur, A. Frik. (2020) “Nudge me right: Personalizing online security nudges to people's decision-making styles”. *Computers in Human Behavior*, 109, 106347.
- Mathur, G. Acar, M. J. Friedman, E. Lucherini, J. Mayer, J. M. Chetty, A. Narayanan, A.(2019) “Dark patterns at scale: Findings from a crawl of 11K shopping websites”. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), pp. 1-32.
- R. Steiner. (2021) “Dark Patterns” . [Online]. Available from: <https://www.fyresite.com/dark-patternsa-new-scientific-look-at-ux-deception/>, 2021.06.24
- E. Waldman (2020) “Cognitive biases, dark patterns, and the ‘privacy paradox’”. *Current opinion in psychology*, 31, pp. 105-109

Chinese Soft Power: Data Politics

Nehme Khawly¹[0000-0002-2221-8532] and Mario Arias-Oliva² [0000-0002-6874-4036]

¹ Lebanese American University, Beirut, Lebanon

² Complutense University of Madrid, Madrid, Spain
nehme.alkhawly@lau.edu.lb

Abstract. The rise of globalization has brought forth innovation, rendered former barriers obsolete, arguably facilitated global development, and arguably contributed to rendering warfare as a choice of last resort for nations. Nonetheless, a great number of scholars have come out to criticize globalization as a tool that is self-serving to the powerful entities in the international system, such as the United States of America, the European Union, China, and others. One of the tools that powerful nations have taken advantage of is the development of the digital world and its subsidiary services such as social media platforms which grant them with an abundance of data that can be used to pursue their interests. This paper will build upon the revelations of the Snowden leaks back in 2013 with regard to the use of digital data by the United States of America. The former led to a domino effect across the world over the saliency of digital privacy and how personal data is used. More recently in 2019, the Facebook-Cambridge Analytica scandal renewed interest in the topic and swayed the discussion towards China and its role in the (ab-)use of digital data to pursue its personal interests. This research project will evaluate Chinese global ambitions and establish a correlation between their achievement and the use of digital data. The former allows us to establish intent, which facilitates the task of validating the role that Chinese tech-enterprises such as Huawei and social media platforms such as Tiktok play in the provision of digital data in addition to the risks those constitute with regard to privacy.

Keywords: China, Digital Data, Privacy, Security, Global Competition

1 Introduction

“Data is the new oil!” proclaimed Clive Humby, British mathematician, back in 2006. The former statement can be interpreted in different ways, but it all boils down to the following: data is highly sought-after, data is highly valuable, and data can be used for national interests. On the other hand, two aspects contradict Humby’s proclamation. The first of those is that oil is finite whereas data is infinite and has no growth cap or defined quantity that may be depleted. The second contradictory aspect rests in the fact that data is multifaceted and can be categorized in ever-changing and increasing types and categories, all of which supplements its value. Those facts have ultimately led nations to use and abuse data for the pursuit of their interests, whether at the domestic or international scale, and scandals with that regard are abundant, such as Julian

Assange's WikiLeaks debacle (Brunton, 2011), the Snowden Leaks (Weinstein, 2014), the Facebook-Cambridge Analytica scandal (Isaak & Hanna, 2018), and many more. Recently, the conversation moved towards the Eastern Hemisphere, towards China as its ambitions grew exponentially and its role as global challenger to the United States of America increasingly materialized itself. Politicians and scholars raised concerns regarding Chinese potential use of digital data, most notably as it collects it through its tech enterprises such as Huawei, or through its social media platforms such as Tiktok and WeChat. The concerns over digital data security rose as Tiktok achieved the rank of 7th most popular social network in the United States of America, 4th amongst social media platforms (Statista, 2021). COVID-19 only contributed to the proliferation of the use of Tiktok across the globe, which translates into additional data being uploaded on a daily basis. Under the Trump Administration, both Chinese-based Tiktok and WeChat were blocked from U.S. app stores (Swanson et al., 2020), in addition to tech-enterprises such as Huawei and ZTE being banned from receiving network equipment licensing in the country which was finalized by U.S. President Joe Biden (Shepardson, 2021). The former endeavors were justified under articles related to national security as well as user security, claiming that Beijing may take advantage of the services that the enterprises and platforms in question offer to collect sensitive data that may constitute a threat to national security and compromise user privacy and security.

2 Data and methods

This Paper will address the digital data and digital security by placing an emphasis on China. On the one hand, the paper in question will highlight the security risks that ease of access to digital data constitutes whether at level of the user or at the scale of national security. On the other hand, the paper will focus on China as a recipient of loads of digital data through its foreign operating enterprises and social platforms, a status it has achieved by remaining at the forefront of cutting edge technology. This approach will serve to highlight the advantages that China is capable of securing due to the use of digital data to advance its global quest, taking advantage of the lack of awareness of the global population with regard to digital privacy and security.

In terms of data collection (methods and nature), this study will rely upon the testimonies of experts for primary data which would be acquired through a series of tailored interviews per the interviewee's expertise. Moreover, in terms of primary information, this paper will rely on the results of an extensive survey that will be administered to users of social platforms such as Tiktok. With regard to secondary data, information will be traced back to accredited sources, most notably academic and professional publications. Additionally, complementary information from news sources will be relied upon when necessary.

By adopting those principles when it comes to study trajectory and data collection, this research effort will be conducted according to a narrative and analytical methodology, promoting accuracy and originality whilst averting the possibility of bias.

3 Discussion

In Scholars have studied Chinese global aspirations for over a decade. Globalization has shed the attention on the importance of data and information, all of which was bolstered by the development of the digital world to circumvent physical obstacles. The former attributed significant value to digital data, which can be harnessed by individuals, businesses, as well as nations in their pursuit of self-serving ends. With regard to China, the date in question can be provided under the Digital Silk Road, the technological extension of the Belt & Road Initiative, which has thus far tapped into wireless networking, surveillance cameras, subsea internet cables, in addition to satellite monitoring, in partnership with a multiplicity of nations (Hillman, 2021).

Moreover, China has taken advantage of increased usage of mobile devices and the extent to which individuals spent time on their screens using social platforms and rolled out Tiktok, whose user base has grown exponentially as a result of the ongoing global pandemic. Owned by a Chinese multinational company, Tiktok collects everything ranging from demographic information, to imagery and videography uploaded (inclusive of sounds), all of which constitutes an increasingly growing database of raw data which can be used for marketing purposes, or with malicious intent. The latter is becoming more alarming whereas Tiktok has, since June 2021, granted itself the permission to collect biometric information of its users including fingerprints, faceprints, and voiceprints (Perez, 2021). Such factors are plausible ground for doubts into Chinese aspirations and its quest to abuse digital data it reaps from its services, whether from Tiktok, Huawei or any other enterprise that operates across the globe, as those constitute serious violations of privacy. Hence, it is understandable that the likes of the U.S. probes into the activity of Beijing as the compiled data collected can constitute a threat to national security, physical or cyber. Nonetheless, China's data is acquired with the knowledge of its users who ratify its privacy rules when subscribing to the terms of agreement, in most cases disregarding the clauses provided, and thus the inconspicuous nature of users with regard to digital security facilitates Beijing's task as will our survey portray.

4 Conclusions

In conclusion, digital data is a double-edged sword which can be harnessed for good use such as shared development and can be used with malicious intent. The latter is apparent and is obvious when tapping into the history of digital data use by governments and their agencies, and facilitated by the lack of awareness of the global population with regard to digital privacy. Digital data is weaponized at present times by competing nations to obtain an advantage over one another. China is one of those nations to (ab-)use digital data and has devised innovative plans to acquire it, most notably through attractive social platforms such as Tiktok. Whilst the means and methods through which Beijing intends on using the data it collects constantly are yet to be observed clearly, the likes of the United States of America and allies have already

taken precautionary measures in order to hamper Chinese plans under the justification of national security, as well as user privacy and security.

Acknowledgements

This research was supported by Telefonica and its Telefonica Chair on Smart Cities of the Universitat Rovira i Virgili and the Universitat de Barcelona (project number 42.DB.00.18.00)

References

- Brunton, F. (2011). WikiLeaks and the Assange papers. *Radical Philosophy*, 166, 8-20.
- Hillman, J. E. (2021, October 21). *Mapping China's Digital Silk Road*. Reconnecting Asia. Retrieved January 17, 2022, from <https://reconasia.csis.org/mapping-chinas-digital-silk-road/#:~:text=The%20Digital%20Silk%20Road%20is,%2C%20subsea%20cables%2C%20and%20satellites.>
- Isaak, J., & Hanna, M. J. (2018). User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer*, 51(8), 56-59.
- Perez, S. (2021, June 3). *Tiktok just gave itself permission to collect biometric data on US users, including 'Faceprints and Voiceprints'*. TechCrunch. Retrieved January 17, 2022, from <https://techcrunch.com/2021/06/03/tiktok-just-gave-itself-permission-to-collect-biometric-data-on-u-s-users-including-faceprints-and-voiceprints/>
- Shepardson, D. (2021, November 12). *Biden signs legislation to tighten U.S. restrictions on Huawei, ZTE*. Reuters. Retrieved January 17, 2022, from <https://www.reuters.com/technology/biden-signs-legislation-tighten-us-restrictions-huawei-zte-2021-11-11/>
- Statista. (2021, November). *Social networks: TikTok in the United States 2021*. Statista. Retrieved January 17, 2022, from <https://www.statista.com/study/72735/social-networks-tiktok-in-the-united-states-brand-report/>
- Swanson, A., McCabe, D., & Nicas, J. (2020, September 18). *Trump administration to ban TikTok and WeChat from U.S. App Stores*. The New York Times. Retrieved January 17, 2022, from <https://www.nytimes.com/2020/09/18/business/trump-tik-tok-wechat-ban.html>
- Weinstein, D. (2014). Snowden and US cyber power. *Geo. J. Int'l Aff.*, 15, 4.

Abolitionist Technology Ethics for Secondary Education

Leah N. Rosenbloom¹ and William M. Fleischman²

¹ Brown University, Providence, Rhode Island, U.S.A.

² Villanova University, Villanova, Pennsylvania, U.S.A.
leah_rosenbloom@brown.edu

Abstract. Technology is ubiquitous in secondary education, both implicitly in the experiences of children in society, and explicitly as an educational tool in secondary school. Existing practices in technology education prioritize technical understanding, and do not explicitly address the ways in which technology can reproduce systems of oppression. We argue the need for technology ethics education in secondary schools, and specifically for ethics that encourage students to confront and dismantle systems of oppression. Using the abolitionist teaching paradigm and case studies in technology ethics, we construct a template for abolitionist technology ethics for secondary education.

Keywords: abolitionist teaching, ethics for technology education, secondary education, abolitionist technology ethics

1 Introduction

Children enter secondary schools with substantial prior knowledge of technological tools. As they grow, their understanding of technology is shaped by the ways they interact with technology inside and outside the classroom. While students' technical competency invariably increases with practice, they get limited exposure to the history, ethics, and societal impact of technologies they use every day. Therefore, it is even more important for technology education programs in secondary schools to focus on the ways in which technology impacts students and their communities, and to provide students with the context and skills necessary to confront and dismantle systems of technological oppression.

The abolitionist teaching paradigm is an educational strategy that works to reveal, address, and replace systems of oppression. We instantiate the abolitionist teaching paradigm with examples from technology ethics to develop robust technology ethics projects for secondary education.

2 Secondary School Students and Technology Education

2.1 Technology and Cognitive Development

Children experience technology from a very young age. Their interactions with technology as an educational tool---for example early learning applications on a tablet or computer---often begin in preschool (Clements & Sarama, 2002). As they grow, children develop implicit skills and fluency with technology; they passively learn to leverage technology in order to communicate, collaborate, and teach themselves new skills. Learning tutorials on YouTube have become particularly popular educational tools among high school students (Bardakci, 2019).

2.2 Technology Education Means Technical Education

Formal technology education programs generally focus on growing students' technical understanding. This mirrors the tendency of computer science educators to focus on technical skills rather than the inclusion of "soft" non-technical disciplines that can better explain the impact of technology on society (Raji et al., 2021). For children who have grown up surrounded by technology---and especially for those with a solid technical foundation---we argue that foundational learning in technology ethics, as well as history, civics, and sociology, is of equal or greater importance than increased technical understanding. Just as shop teachers would never hand children saws without proper safety training, so should technology teachers never hand children technical tools without teaching them how to avoid addiction, misinformation, and oppression.

2.3 Tools of Oppression

There is a growing body of work that studies the ways in which technology, especially machine learning but also military-adjacent technologies like computer vision and natural language processing, reproduce systems of oppression (Benjamin, 2019; Hampton, 2021; Noble, 2018). For secondary schools students, social media has been shown to lead to low self-esteem, self-harm, and abuse, especially for young girls (Noble, 2018). Technology education programs that focus on technical understanding and do not explicitly address these issues contribute to the invisibilization of technological oppression and further harm children who are impacted by this oppression. We therefore create a pedagogy of abolitionist technology ethics: technology education that works to reveal and dismantle systems of technological oppression.

3 The Abolitionist Teaching Paradigm

3.1 Principles of Abolitionist Teaching

Abolitionist teaching is an educational method that seeks to reveal, dismantle, and replace systems of oppression with freedom for all people (Love, 2019). The core principles of abolitionist teaching, summarized below, are rooted in critical pedagogy, intersectional feminism, and the spirit and methods of abolitionists.

Community. The first principle is a focus on community, both inside and outside the classroom. The practice of building community ensures that all children feel safe and supported in the learning environment.

History, Civics, and Resistance. Once community is established, abolitionist teachers asks participants to examine the historical and contemporary context of structural racism and interlocking systems of oppression, as well as tangible ways to resist them.

Intersectionality. In the spirit of Black feminist thinkers, abolitionist teachers believe that systems of oppression are interlocking---that we cannot successfully eliminate just one form of oppression while the others remain. The principle of intersectionality calls for the inclusion of all voices.

Confronting Whiteness. By “whiteness,” abolitionist teachers mean the societal norm of white supremacy and superiority. White supremacy exists---and must be confronted---on both a structural and an individual level. In order to be addressed, whiteness must first be rendered visible.

Joy and Well-Being. Similar to the principle of community, prioritizing joy and well-being ensures that both teachers and students have the physical, mental, emotional, and spiritual health necessary to confront and dismantle systems of oppression.

Freedom Dreaming and Revolutionary Spirit. Abolitionists teach us that it is not enough to simply dismantle systems of oppression; we must be ready to replace them with something better. To that end, freedom dreaming encourages students to imagine a better world, and revolutionary spirit encourages them to realize it.

3.2 Applications to Technology Ethics

The abolitionist teaching paradigm has the power to correct many different shortcomings in the area of technology education, as well as to address the overarching problem of technological oppression. For example, the practice of confronting whiteness would reveal and address the common technological paradigm of techno-saviorism, or the idea that technology can fix anything (Raji et al., 2021). Many existing case studies in technology ethics fit into this paradigm, and could be used in the classroom to illustrate the impact of systems of oppression on ethical problem-solving.

4 Toward Technology Ethics in Secondary Education

In the full version of this paper, we will instantiate the abolitionist teaching paradigm with examples from technology ethics to develop projects in abolitionist technology ethics. While we've cited an example where the point is to expose an underlying set of assumptions that perpetuate systems of oppression, there are also projects that affirmatively use technology and publicly available datasets to help expose inequities that affect children across racial and economic divides. Fleischman describes a project in which, using Harvard School of Public Health (HSPH) data in combination with traditional instruction in technical skills, middle school children were able to explore differences in life expectancy associated with demographic and socioeconomic variables such as income and race. Since the granularity of HSPH datasets provide comparisons down to the county (and, more recently, the census tract) level, students were able to relate differences in life outcomes in their immediate region to factors they could understand through lived experiences of their families and neighborhoods. (Fleischman, 2008) This sort of analysis empowers students to think about the effects of public health disparities and problems including gun violence, drug abuse, diabetes and other chronic diseases on their communities; to imagine a better environment in which to learn and grow; and to join their own efforts to those of other community organizations in advocating for changes to call into being this more ideal community.

References

- Bardakci, S. (2019). Exploring High School Students' Educational Use of YouTube. *International Review of Research in Open and Distributed Learning*, 20(2).
- Benjamin, R. (2019). Race after technology: Abolitionist tools for the new jim code. *Social Forces*.
- Clements, D. H., & Sarama, J. (Eds.). (2002). Early Childhood Corner: The Role of Technology in Early Childhood Learning. *Teaching children mathematics*, 8(6), 340-343.
- Fleischman, W.M. (2008). Getting to the Other Side – Beyond the Digital Divide. *Proceedings of ETHICOMP 2008, University of Pavia, Mantua, Italy*. (pp. 245-255).
- Hampton, L. M. (2021). Black Feminist Musings on Algorithmic Oppression. *arXiv preprint arXiv:2101.09869*.
- Love, B. L. (2019). *We want to do more than survive: Abolitionist teaching and the pursuit of educational freedom*. Beacon Press.
- Noble, S. U. (2018). *Algorithms of oppression*. New York University Press.
- Raji, I. D., Scheuerman, M. K., & Amironesei, R. (2021, March). You Can't Sit With Us: Exclusionary Pedagogy in AI Ethics Education. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 515-525).

Free Speech and Computing Professionals: Moral Considerations and Tensions (Extended Abstract)

Michael S. Kirkpatrick ¹[0000-0002-7200-4102], Emmanuelle Burton ²[0000-0002-0570-9719], and
Marty J. Wolf ³[0000-0003-0617-942X]

¹James Madison University, Harrisonburg, Virginia, USA

²University of Illinois Chicago, Chicago, Illinois, USA

³Bemidji State University, Bemidji, Minnesota, USA

kirkpams@jmu.edu

Keywords: free speech, computing professionals, platform governance

Traditionally, justifications for free speech examine rationales that focus on principles of liberty for the speaker. Greenawalt (2005), for instance, details rationales for supporting speech beyond a “minimal principle of liberty” that only considers government prohibitions on speech. Consequentialist rationales include the role that free speech plays in truth seeking, exposing abuses of authority, fostering emotional outlet and personal development, and promoting tolerance. Nonconsequentialist rationales include a Lockean notion of consent of the governed, autonomy, and personal dignity.

These approaches, which generally reflect Western European and American philosophical traditions, implicitly assume that the mechanisms for speech are scarce and the primary concern is the ability of a powerful entity to infringe on personal liberties. Balkan (2018) argues that the information age has replaced this “dyadic” model (the individual vs. a single authoritarian entity) with a pluralistic model of speech regulation. This “new school” model recognizes that private organizations (social media platforms, search engines, domain registrars) have the power to regulate speech, just as governments can. At the same time, individuals—including citizens, hackers, and trolls—have unprecedented abilities to threaten the speech and civic participation of everyone.

Under this new model of speech governance, the open nature of Internet communication virtually eliminates the traditional forms of censorship and scarcity of speech. Rather than relying on editors and publication agencies, individuals can host their own web sites and publish their work freely. Low-cost and free platforms, such as those used for social media, blogs, or general web hosting, further reduce the time and effort required to make one’s views available. Information sources have become increasingly fragmented and decentralized, providing unprecedented opportunities for influencers and trolls to have an outsized impact on social dynamics (Mazarr et al., 2019).

The common advice to counter bad or false speech with “more speech” (generally attributed to Justice Brandeis in *Whitney v. California*, 1927) may not be as effective

given the affordance of networked communication technologies. Phillips (2018) documents how malicious actors have manipulated traditional approaches to counter narratives, such as fact checking, to increase the impact of their attacks and abuse. Croeser (2016) also argues that abusers can exploit free speech arguments to protect forms of speech that are typically not protected, such as harassment, discrimination, and threats. In short, there has been a shift from a scarcity of open platforms to a scarcity of protective resources to counter the scope and scale of targeted bad speech.

1 Legal and Moral Foundations

Within the Western European and American philosophical tradition, free speech is deemed necessary for liberal democracy and encoded into fundamental laws, though with some limitations. The First Amendment of the U.S. Constitution prohibits the government from “abridging the freedom of speech,” protecting many types of speech (such as offensive speech, hate speech, obscenity, and invasions of privacy) that are frequently banned in other countries. In contrast, Article 10 of the European Convention on Human Rights allows restrictions deemed “necessary in a democratic society,” while the Canadian Charter of Rights and Freedoms permits “reasonable limits” “as can be demonstrably justified in a free and democratic society.” These legal foundations reflect the traditional focus on protecting speech as a scarce resource that plays an instrumental role in liberal society.

The rise of networked communications has highlighted key challenges with how to support this principle. Platforms muddy the distinctions between speakers, editors, publishers, and distributors that underlie common legal limits, such as defamation and harassment laws, commonly playing multiple roles. In the U.S., Section 230 of the Communications Decency Act, generally immunizes platforms from such lawsuits, as they cannot “be treated as the publisher or speaker” of their users’ speech. Kosseff (2019) argues that this law provides a necessary—if imperfect—liability shield for a variety of Internet services.

While the legal debate surrounding free speech continues to evolve in relation to new technologies, computing professionals must adapt to the changing landscape. Specifically, the legality of particular speech in the context of a functioning liberal democracy constitutes a necessary condition for ethical conduct, though not necessarily a sufficient one. Computing professionals must learn to resolve conflicting values and rights, even when speech is otherwise legally protected. This process is both iterative and context-sensitive. Speech may be acceptable or not, depending on the particular circumstances or based on the individual speaker. Making these determinations requires nuance and should not evolve into anticipatory obedience. We will explore these considerations in more depth in the full version of the paper.

2 Free Speech and the Computing Professional

The ACM Code of Ethics and Professional Conduct (“the Code”) describes key ethical principles that should be considered as part of the work of computing

professionals. These principles, such as respecting privacy, avoiding harm, and promoting fundamental human rights, may conflict with protecting free speech or exercising one's own right to free speech. Computing professionals, especially in their professional capacity, have both rights and responsibilities that make decisions about their own speech more nuanced than when they are acting in a personal capacity. For example, they "should not misrepresent an organization's policies or procedures, and should not speak on behalf of an organization unless authorized to do so." Applying this guidance can be difficult in some situations, though the speech may still be legal.

Consider a leader of an ACM special interest group (SIG) who will be making a conference presentation in their area of computing expertise. How that person's leadership position is conveyed to the audience changes the ethical calculus. When giving a presentation (with no explicit reference to the speaker's SIG leadership position) or interacting with others on social media from a personal account, it should be clear that the person is speaking in an individual capacity; anything said by the person is to be attributed to that person, not the SIG. If the speaker is billed as a leader of the SIG or using an official SIG account, however, it is reasonable for audience members to conclude that the speech is on behalf of the SIG; the speaker must ensure that any representations made are, in fact, authorized and accurately represent the position held by the SIG.

In the full paper, we will include other scenarios and demonstrate how the Code can help guide nuanced analysis of these situations. In many cases, these situations limit free speech in ways that might be inconsistent with well-understood norms stemming from functioning liberal democracies. Further, these limits may not be binary. There may be specific things that a computing professional may say in one context (how the audience understands the role of the speaker, the medium that the speech is occurring in, etc.) that are not allowed by the Code in another. Further, a nuanced variant of that same speech may well not contravene a professional's obligations to the Code.

3 Platform Governance is Unavoidable

In recent years, much of the debate surrounding free speech online has centered around platforms for user-generated content, especially social media platforms. At various times, activists, politicians, and users have raised concerns about the role that platforms have taken on in shaping the public discourse. One key insight for computer professionals to recognize is that all platforms govern speech, regardless of their intended or stated goals (Gorwa, 2019). Some forms of platform governance are uncontroversial, such as filtering spam, preventing bots and automated posting, or removing malware, partially because they have not been immediately recognized as speech. Beyond these initial steps, platforms also have incentive structures that compel the removal of content that is not welcomed by their community of users, such as reducing harmful content (e.g., violent and disturbing images, as well as hateful, racist, or sexist posts). Platforms' popularity is highly dependent on their

ability to remove this content, but their reputations also rely on keeping these mechanisms opaque and hidden (Roberts, 2016 and 2018).

Computing professionals need to recognize that there are unavoidable tensions in creating networked communication technologies, including but not limited to social media platforms. Mechanisms for reporting harmful content can be abused to suppress unpopular opinions or to force one community's moral standards on another. Seemingly benign rules can be interpreted too strictly, causing moderators to remove highly valued or historically important content. At the same time, platforms with permissive attitudes toward harassment or offensive content can be used to coordinate attacks to harass targeted users on other platforms (Massanari, 2017). In such cases, the target may be left with little recourse but to disconnect from the public sphere, giving malicious actors more protection than they would enjoy in other domains. Computing professionals who design, build, and maintain such systems need to be cognizant of these types of attacks and mitigate against them appropriately. The Code can offer guidance for principles to consider beyond just protecting free speech.

In the full version of the paper, we will examine additional examples of ethical challenges that computing professionals may face when balancing free speech rights with other values that are fundamental to full participation in a liberal society.

References

- Balkin, J. M. (2018). Free speech in the algorithmic society: Big data, private governance, and new school speech regulation. *UC Davis Law Review*, (2018 Forthcoming), Yale Law School, Public Law Research Paper No. 615, Available at SSRN: <https://ssrn.com/abstract=3038939> or <http://dx.doi.org/10.2139/ssrn.3038939>.
- Croeser, S. (2016). Thinking beyond "free speech" in responding to online harassment. *Ada: A Journal of Gender, New Media, and Technology*, No. 10. doi:10.7264/N35Q4TC4
- Gorwa, R. (2019). What is platform governance? *Information, Communication & Society*, 22:6,854-871, DOI: 10.1080/1369118X.2019.1573914.
- Greenawalt, K. (2005). Rationales for freedom of speech. In A. D. Moore (Ed.), *Information Ethics: Privacy, Property, and Power*, University of Washington Press.
- Kosseff, J. (2019). *The Twenty-Six Words That Created the Internet*, Cornell University Press.
- Massanari, A. (2017). #Gamergate and The Fapping: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 19:3, 329-346, DOI: 10.1177/1461444815608807.
- Mazarr, J. M., Bauer, R. M., Casey, A., Heintz, S. A., & Matthews, L. J. (2019). *The Emerging Risk of Virtual Societal Warfare: Social Manipulation in a Changing Information Environment*. RAND Corporation. https://www.rand.org/pubs/research_reports/RR2714.html
- Phillips, W. (2018). *The Oxygen of Amplification*. Data & Society. <https://datasociety.net/library/oxygen-of-amplification/>

- Roberts, S. T. (2016). Commercial content moderation: Digital laborers' dirty work. *Media Studies Publications* 12. <https://ir.lib.uwo.ca/compub/12>
- Roberts, S. T. (2018). Digital detritus: 'Error' and the logic of opacity in social media contentmoderation. *First Monday*, Volume 23, Number 3 - 5 March 2018. DOI: <http://dx.doi.org/10.5210/fm.v23i3.8283>

Access Control Meets Genetics: The Challenges of Information Leakage and Non-users (Extended Abstract)

Michael S. Kirkpatrick^[0000-0002-7200-4102]

James Madison University, Harrisonburg, Virginia, USA

kirkpams@jmu.edu

Computer and information systems rely on access control policy specifications to enforce constraints of what can be done with a collection of data. Traditionally, such policies are represented as a collection of tuples of the form $\langle s, o, a \rangle$, indicating that a subject (s) is authorized to perform an action (a) on an object (o). For example, the tuple $\langle \text{admin}, \text{passwords.txt}, \text{read} \rangle$ could indicate that an administrator has access to read the contents of a password file. Such policies are characterized as identity-based, as their enforcement relies on determining the identity of the subject (typically but not always a human user) attempting the requested action. Identity-based systems work well in domains where the policy is specified by someone who is considered the owner or administrator of the data, such as a document shared via a cloud storage system.

Genetic and genomic data present a unique challenge for specifying access control policies. Bioethicists and genetics researchers have used “genetic exceptionalism” to indicate that genetic information is “uniquely powerful and uniquely personal” (Annas et al., 1995). Within these communities, now argue that this exceptionalism is not appropriate when compared to other medical data (Murray, 2019) while others have shown that genetic and genomic data create privacy threats that other medical data does not (Erlich et al., 2018).

With large-scale research projects, such as the Human Genome Project, a considerable amount of work has been done to identify and resolve questions about genetic and genomic privacy (NIH, 2020). Given the collaborative nature of such research, much of the emphasis has focused on understanding research participants’ desires for privacy (Robinson, 2015) and navigating the tensions between protecting privacy while achieving the benefits of sharing data (Kaye, 2012).

Outside of the realm of genetic and genomic research, other data collections raise new concerns for privacy. One such area is the direct-to-consumer (DTC) genetic testing (Knoppers, 2010). These services can be used for many applications, including limited health screening, ancestry research, or searches for long-lost relatives. However, the implications of these services can come as a surprise to some users. One high-profile example of such a scenario emerged when police used the GEDmatch service to identify the “Golden State Killer” (Kaiser, 2018; Selk, 2018). One

particularly salient aspect of this case was the fact that the killer had never used this service; the relevant genetic information was that of a distant relative.

While this sensational case highlights an extreme example, DTC testing raises more mundane concerns. Many people who use these services get surprising discoveries of non-paternity events (NPE), undisclosed adoptions, and consanguinity, any of which can be personally devastating. Like the Golden State Killer case, these privacy threats affect individuals who may not be users of the system. As such, identity-based access control systems are inadequate and other approaches are necessary.

More recent work in access control has sought to create more flexible and expressive policy languages that do not rely on identity-based assumptions. In attribute-based access control (ABAC) (NIST, 2014), the policy may be written based on common attributes about collections of subjects or objects. As an example, a corporate ABAC system might restrict access to a system based on whether the user has completed a mandatory training program. Thus, ABAC policies support the consideration of contextual factors when granting or denying access.

Relationship-based access control (ReBAC) models extend this concept by emphasizing the importance of how entities relate to each other (Fong, 2011; Bogaerts, 2015). For example, a medical ReBAC system might restrict a patient's medical records only to physicians who have seen that patient within the past year. Thus, the policy decision rests on the existence of a relationship between the physician and patient rather than their individual qualifications, roles, or other attributes. This approach has attracted a lot of research interest in the access control field, particularly focusing on techniques to automate the dynamic identification of relationships, even as the system evolves (Bui, 2019; Iyer, 2020; Mehregan, 2016).

Another key access control approach focuses on multi-party policies (Squicciarini, 2020). These approaches focus on the challenges created by multiple owners of a collection of data. In some cases, the multiple stakeholders may be actively working together as co-owners to create a single piece of work, such as friends sharing pictures in social networks (Mehregan, 2016). In other cases, the multiple ownership claims can arise from aggregating data sources (Rosa, 2020). Regardless of the provenance, these approaches aim to reconcile competing policy intentions.

As noted by Clayton (2019), genetic information is complicated and defies easy classification. Everyone's full genome is unique; Clayton notes that even monozygotic ("identical") twins also have small differences in their genomes. At the same time, though, each individual's genome contains segments of genetic information that is shared by relatives, ethnic groups, or the full human population. Consequently, this dual nature challenges the foundational concepts of data ownership or privacy that subjects in genomic research or customers of DTC genetic testing companies mistakenly take for granted.

Given this ambiguous nature of genetic information, existing approaches to access control can express some, but not all, desired policy goals. ABAC could be applied to common segments associated with medical conditions; for instance, someone researching a particular condition that is linked to a known gene variant could be granted access based on the presence of that variant. Multi-party policies could be

applied to segments that are shared between close relatives. And ReBAC policies could establish controls over the sharing of family networks. However, all implicitly assume the individual under consideration is a knowing and consenting user of the system.

The nature of genetic information creates novel challenges for the access control community that will be difficult to overcome. First, knowledge of the existence (or lack thereof) of a genetic relationship itself constitutes an information leak. It is a common principle of genetic privacy that an individual has the right to choose ignorance of a testing result; many people do not want to know whether a test result was positive or negative, particularly if there is no known intervention for the condition.

A second challenge for access control is that the stakeholders in these cases extend beyond the users of the system. As noted above, the Golden State Killer was not a user of GEDmatch and therefore could not have consented to their information sharing policies. Removing the particulars of the law enforcement scenario, one can imagine a similar scenario in which GEDmatch could be used to trace an undisclosed adoption against the will of the parent, based on the genetic information of the parent's sibling. Relatedly, the stakeholders of these cases are not limited to the individuals whose genetic information is under consideration. The discovery of an unexpected relationship can also impact spouses, partners, and others.

A third challenge revolves around the issue of consent. Given the biological nature of genetic data, the children of DTC consumers and genomic research participants are implicated by the presence of their parents' genetic information in these databases. As such, they are denied the meaningful ability to restrict access to their own genetic information, which may have long-term ripple effects. Cohn (2020) argues that the lack of consent makes the use of DTC testing for ancestry research morally objectionable. Additionally, there are legal implications in this lack of consent, as raised in *Maryland v. King*, a legal case brought to the U.S. Supreme Court. The Court found that it was legal for a police department to collect genetic samples from an arrested person without their consent. It is important to note that this sample can be kept, even if the person is never charged with a crime nor convicted. As noted above, this lack of consent can have secondary effects on others beyond that particular individual.

In summary, recent advances in access control have made significant strides toward more flexible policy models that can incorporate complex relationships and attributes of subjects and objects within a computing system. These models are based on a potentially flawed assumption that users are the most relevant stakeholders to make policy decisions. Within the context of genetic information, these models are insufficient to address the concerns and rights of all relevant stakeholders. As such, there is a need for the access control community to develop newer approaches that account for the unique challenges of genetic information.

References

- Annas, G. J., Glantz, L. H., & Roche, P. A. (1995). Drafting the genetic privacy act: Science, policy, and practical considerations. *The Journal of Law, Medicine & Ethics*, 23(4), 360-366.
- Bogaerts, J., Decat, M., Lagaisse, B., & Joosen, W. (2015). Entity-based access control: Supporting more expressive access control policies. From ACSAC 2015: Proceedings of the 31st Annual Computer Security Applications Conference, 291-300.
- Bui, T., Stoller, S. D., & Le, H. (2019). Efficient and extensible policy mining for relationshipbased access control. From SACMAT '19: Proceedings of the 24th ACM symposium on Access control models and technologies, 161-172.
- Clayton, E. W., Evans, B. J., Hazel, J. W., & Rothstein, M. A. (2019). The law of genetic privacy: applications, implications, and limitations. *Journal of law and the biosciences*, 6(1), 1 – 36.
- Cohn, B. C. (2020). *Direct-to-consumer genetic ancestry testing: A morally objectionable practice*. [Thesis, Johns Hopkins University]. <https://dspace-prod.mse.jhu.edu/handle/1774.2/63300>
- Erlich, Y., Shor, T., Pe'er, I., & Carmi, S. (2018). Identity inference of genomic data using long-range familial searches. *Science*, 362, 690-694.
- Fong, P. W. L. (2011). Relationship-based access control: protection model and policy language. From CODASPY '11: Proceedings of the first ACM conference on Data and application security and privacy, 191-202.
- Iyer, P. & Masoumzadeh, A. (2020). Active learning of relationship-beased access control policies. From SACMAT '20: Proceedings of the 25th ACM symposium on Access control models and technologies, 155-166.
- Kaiser, J. (2018). We will find you: DNA search used to nab Golden State Killer can home in on about 60% of white Americans. *Science magazine*. <https://www.sciencemag.org/news/2018/10/we-will-find-you-dna-search-used-nab-golden-state-killer-can-home-about-60-white>.
- Kaye, J. (2012). The tension between data sharing and the protection of privacy in genomics research. *Annual review of genomics and human genetics*, 13, 415-431.
- Knoppers, B. M. (2010). Consent to 'personal' genomics and privacy. Direct-to-consumer genetic tests and population genome research challenge traditional notions of privacy and consent. *EMBO reports*, 11(6), 416-419.
- Mehregan, P. & Fong, P. W. L. (2016). Policy negotiation for co-owned resources in relationshipbased access control. From SACMAT '16: Proceedings of the 21st ACM symposium on Access control models and technologies, 125-136.
- Murray, T. H. (2019). Is genetic exceptionalism past its sell-by-date? On genomic diaries, context, and content. *The American Journal of Bioethics*, 19:1, 13-15.
- NIH Genome Research Institute (2020). Privacy in genomics. <https://www.genome.gov/about-genomics/policy-issues/Privacy>.

- NIST (2014). Guide to attribute-based access control (ABAC) definition and considerations. *NIST Special Publication 800-162*. <https://doi.org/10.6028/NIST.SP.800-162>.
- Rizvi, S. Z. R., Fong, P. W. L., Crampton, J., & Sellwood, J. (2015). Relationship-based access control for an open-source medical records system. From SACMAT '15: Proceedings of the 20th ACM symposium on Access control models and technologies, 113-124.
- Robinson, J. O., Slashinski, M. J., Chiao, E., & McGuire, A. A. (2015). It depends whose data are being shared: considerations for genomic data sharing policies. *Journal of law and the biosciences*, 2(3), 697-704.
- Rosa, M., Di Cerbo, F., & Lozoya, R. C. (2020). Declarative access control for aggregations of multiple ownership data. From SACMAT '20: Proceedings of the 25th ACM symposium on Access control models and technologies, 59-70.
- Selk, A. (2018). The ingenious and 'dystopian' DNA technique police used to hunt the 'Golden State Killer' suspect. *Washington Post*. <https://www.washingtonpost.com/news/truecrime/wp/2018/04/27/golden-state-killer-dna-website-gedmatch-was-used-to-identify-joseph-deangelo-as-suspect-police-say/>.
- Squicciarini, A. C. (2020). Multi-party access control – 10 years of successes and lessons learned. From SACMAT '20: Proceedings of the 25th ACM symposium on Access control models and technologies, 189-190.