



Embedding artificial intelligence in society: looking beyond the EU AI master plan using the culture cycle

Simone Borsci^{1,2} · Ville V. Lehtola³ · Francesco Nex³ · Michael Ying Yang³ · Ellen-Wien Augustijn⁴ · Leila Bagheriye^{5,6} · Christoph Brune⁷ · Ourania Kounadi^{4,8} · Jamy Li^{9,10} · Joao Moreira¹¹ · Joanne Van Der Nagel¹⁰ · Bernard Veldkamp¹ · Duc V. Le¹² · Mingshu Wang^{4,13} · Fons Wijnhoven¹⁴ · Jelmer M. Wolterink⁷ · Raul Zurita-Milla⁴

Received: 14 July 2021 / Accepted: 21 December 2021 / Published online: 22 January 2022
© The Author(s) 2022

Abstract

The European Union (EU) Commission’s whitepaper on Artificial Intelligence (AI) proposes shaping the emerging AI market so that it better reflects common European values. It is a master plan that builds upon the EU AI High-Level Expert Group guidelines. This article reviews the masterplan, from a culture cycle perspective, to reflect on its potential clashes with current societal, technical, and methodological constraints. We identify two main obstacles in the implementation of this plan: (i) the lack of a coherent EU vision to drive future decision-making processes at state and local levels and (ii) the lack of methods to support a sustainable diffusion of AI in our society. The lack of a coherent vision stems from not considering societal differences across the EU member states. We suggest that these differences may lead to a fractured market and an AI crisis in which different members of the EU will adopt nation-centric strategies to exploit AI, thus preventing the development of a frictionless market as envisaged by the EU. Moreover, the Commission aims at changing the AI development culture proposing a human-centred and safety-first perspective that is not supported by methodological advancements, thus taking the risks of unforeseen social and societal impacts of AI. We discuss potential societal, technical, and methodological gaps that should be filled to avoid the risks of developing AI systems at the expense of society. Our analysis results in the recommendation that the EU regulators and policymakers consider how to complement the EC programme with rules and compensatory mechanisms to avoid market fragmentation due to local and global ambitions. Moreover, regulators should go beyond the human-centred approach establishing a research agenda seeking answers to the technical and methodological open questions regarding the development and assessment of human-AI co-action aiming for a sustainable AI diffusion in the society.

Keywords Artificial intelligence · Human artificial intelligence interaction · AI Policies · Symbiosis of AI and Humans · Societal impact · Trust

1 Introduction

The European Union Commission’s whitepaper (ECWP) on Artificial Intelligence (AI) anticipates a common market for AI and aims to secure civil rights in a future EU society embedded with AI systems (EC EUWP 2020; MSI-NET 2018). The ECWP builds upon four pillars identified by the AI High-Level Expert Group (AIHLEG) for driving the incorporation of AI in the EU society, i.e., respect for

human autonomy, harm prevention, fairness, and explicability (AIHLEG 2019). This European Commission’s (EC) master plan aims to build and regulate the EU AI market while acknowledging the importance of balancing common principles with the specific interests of the stakeholders. In line with John McCarthy who defined, in 1956, AI as the “science and engineering of making intelligent machines, especially intelligent computer programs” (Rajaraman 2014), the EC recognises that this science is currently fragmented and needed to be strengthened to facilitate the embedding of AI in our societies (e.g., lighthouse centres): “Europe cannot afford to maintain the current fragmented landscape [...]. It is imperative to create more synergies and

✉ Simone Borsci
s.borsci@utwente.nl

Extended author information available on the last page of the article

networks between the multiple European research centres on AI” (ECWP 2020, p. 6).

A unified definition of AI is currently missing in literature (Wang 2019). That said, AI systems in the context of the EC plan are intended as all those “systems that display intelligent behaviour by analysing their environment and taking actions—with some degree of autonomy—to achieve specific goals” (EU COM 237, 2018, p. 1). By adopting such a wide AI definition, the EC aims for providing a general framework to regulate different types of AI without, however, focusing on specific and contextual details. Embracing such a general perspective on AI sets a clear limitation to the present work, namely, that we can only marginally discuss specific implementations of such systems and hence ethical analysis of these solutions are beyond the scope of this manuscript. This wide perspective on AI, however, is also an advantage as it offers the possibility to look from a general standpoint at how the EC plan could match or clash with current societal, technical, and methodological boundaries. Additionally, it enables us to formulate our message so that can be directly useful to the EC when monitoring the regulations’ impact. Therefore, here we are not proposing how to solve issues regarding specific AI solutions but reviewing the conditions (societal, technical, and methodological) in which the EC plan will operate.

The EC plan is pushing for a change of culture around the design of AI to transition the EU digital society toward an AI-embedded society where technology is not only interactive (digital) but also proactive in proposing solutions and performing actions. The EC plan grounds the EU diffusion of AI to the “values and fundamental rights such as human dignity and privacy protection” (ECWP 2020, p. 6) aiming for a human-centric approach to AI (COM 237 2018; ECWP 2020). This centrality of people rights for AI development is connecting the EC plan with the UN declaration of Human Rights (UN General Assembly 1948; UNESCO 2021). Moreover, this human-centric approach is a key difference with, for instance, the US National Science Technology Council (NSTC) which mainly intends AI as a set of transformative technologies putting at the centre the value of these in terms of social and economic empowerment (NSTC 2019). The USA’s approach to AI is centred around utility and cost-effectiveness, a vision which is in line with the one of the Chinese’s government (State Council of China 2017), where AI tools are seen as enablers of a data-driven economic transformation and the new focus of international competition in terms of industrial upgrading.

The EC recognizes the economic driving force of AI as well, however, the central assumption of the EC plan is that this transformation can only happen while putting citizens at the centre by enabling trust toward technology and people who are making and governing AI. The EC key objective is to emerge as “a quality brand for AI” (European

Political Strategy Centre 2018) more than seeking short-term economic advantages. This marks another difference with competitors (e.g., China and USA, etc.) where trust is seen as a technical aspect, i.e., trust toward the product. Conversely, the EC considers trustworthiness a strategic aspect that should be enabled systemically among operators, people and technologies: “Building an ecosystem of trust is a policy objective in itself and should give citizens the confidence to take up AI applications and give companies and public organisations the legal certainty to innovate using AI” (ECWP 2020, p. 3).

In this article, we start by reviewing the AIHLEG-guidelines intended as the skeleton of the ECWP, then we consider some criticisms, and propose a Culture Cycle Framework (CCF, Hamedani and Markus 2019) to expose gaps in the EC plan. Then, we discuss the societal, technological, and methodological factors that may enable the diffusion of trustworthy AI in the EU. Our goal is to identify challenges and opportunities and suggest future directions for the embedding of AI in our society, and we discuss these towards the end of the article.

1.1 Status quo: AIHLEG guidelines and the EU values driving manufacturers

The EU’s AIHLEG-guidelines—in line with laws representing EU values—advise AI developers to design robust AI to avoid harm for end-users due to foreseeable risks and failures (AIHLEG 2019, p. 5). The guidelines offer four “ethical imperatives” (AIHLEG 2019, p. 11) according to which practitioners should conduct themselves (pages 12 and 13): (i) Respect for human autonomy, (ii) Prevention of harm, (iii) Fairness, and (iv) Explicability. EU regulators then propose seven main requirements that should be evaluated throughout the AI system’s lifecycle to ensure adherence to these principles. We list these requirements, and associate them with relevant imperatives, in Table 1.

Although the AIHLEG declares a non-hierarchical organization of principles, more AI requirements seem to be associated with imperatives ii and iii, harm prevention and fairness, than with imperatives concerning autonomy (i) and explicability (iv). This is concerning for three reasons. First, though the EC’s proposal prioritizes harm prevention over the other pillars, the relative lack of attention to issues of autonomy and explicability is not addressed explicitly, but only emerges through an analysis of the AIHLEG-guidelines and the ECWP. The seven requirements are design principles meant to guide the AI implementation, and the four imperatives should be used to support engineering processes in cases of conflicts between systems’ functions, ethics, and regulation (AIHLEG 2019, p. 13 and 24). It is well known that it could be necessary to violate a principle, partially or completely, to fully operationalise

Table 1 Relationships among requirements and imperatives for trustworthy AI systems according to EU AI High-Level Expert Group guidelines (AIHLEG 2019)

Requirements for trustworthy AI	Imperatives of trustworthy AI			
	i	ii	iii	iv
	Respect for human autonomy	Prevention of harm	Fairness	Explicability
Human agency and oversight , i.e., AI systems should respect fundamental rights, human agency, and human oversight (associated with Imperative i)	X	X		
Technical robustness and safety , i.e., AI systems should be resilient to attack and security, fall back plan and general safety, accuracy, reliability, and reproducibility (associated with Imperative ii)		X		
Privacy and data governance , i.e., Including respect for privacy, quality and integrity of data, and access to data (associated with Imperative ii)		X		
Transparency , i.e., Including traceability, explainability, and communication (associated with Imperative iv)				X
Diversity , i.e., the avoidance of unfair bias, accessibility and universal design, and stakeholder participation (associated with Imperative iii)			X	
Societal and environmental wellbeing , i.e., sustainability and environmental friendliness, social impact, society, and democracy (associated with Imperative ii and iii)		X	X	
Accountability , i.e., Auditability, minimisation, and reporting of negative impact, trade-offs, and redress. (associated with Imperative iii)			X	

the systems' functionalities (Hollnagel 2009). This practical fact is also recognised in the AIHLEG indicating that there are “no fixed solutions” (AIHLEG 2019, p. 13) for design and ethics trade-offs. In case of conflicts between requirements and ethical imperatives, harm prevention (Table 1, imperative ii) implicitly prevails as the pillar that should guide the decision-making. The importance of this imperative to ‘not harm’ is reinforced in the ECWP: “While AI can do much good, including by making products and processes safer, it can also do harm. This harm might be both material (safety and health of individuals, including loss of life, damage to property) and immaterial (loss of privacy, limitations to the right of freedom of expression, human dignity, discrimination for instance in access to employment), and can relate to a wide variety of risks. A regulatory framework should concentrate on how to minimise the various risks of potential harm [...]” (ECWP 2020, p. 10). With this focus on harm prevention (and conversely, safe use) the EC plan draws upon classic precautionary engineering principles and human-centred design approaches (ISO 9241–11; ISO 9241–210). This centrality of harm prevention is also recognised by the ECWP: “risks can be caused by flaws in the design of the AI technology, be related to problems with the availability and quality of data or to other problems stemming from machine learning. [...] A lack of clear safety provisions tackling these risks may, in addition to risks for the individuals concerned, create legal uncertainty for businesses” (ECWP 2020, p. 12). This hidden hierarchical organization of the imperatives could be considered a minimal theoretical issue.

A second concern is that when it comes to describing methods for assessing the safe use of AI, the guidelines are not enlightening, suggesting that methods that can be “[...] either complementary or alternative to each other, since different requirements [...] may raise the need for different methods of implementation” (AIHLEG 2019, p. 21). Later, the guidelines suggest that: “due to the non-deterministic and context-specific nature of AI systems, traditional testing is not enough” (AIHLEG 2019, p. 22). Recently, Federici et al. (2020) suggested in a systematic review on AI for health that a unified framework to benchmark, assess and selectively support the AI development is lacking. Highlighting a concerning gap of knowledge and wide use of different (more and less reliable) methods to assess the safety and user experience (UX) with AI systems. If the thesis of Federici et al. (2020) is correct, then the key methodological challenge for AI developers is how to generate reliable and replicable evidence to support the design of safe AI.

The third concern is that safe AI should come by a comprehensive extension or new perspectives on metrics and practices to assess the interaction with AI. Without such an extension, available methods for assessing AI may be used by developers to identify shortcuts and rapidly fulfil the list of trustworthiness criteria proposed by the guidelines (AIHLEG 2019, p. 26), instead of aiming for rigorous reliability and quality stress tests of their systems. Following Rességuier and Rodrigues (2020), without a clear set of indications on how to operationalize the EU values into practice, requirements and imperatives may be exposed to the risk of being conveniently used by different stakeholders

to achieve their practical goals. This risk is also connected to the fact that the EU framework of requirements is unspecific and broad, aiming to support the development of any type of AI. Based on risks in usage, regulators will establish additional standards for different AI asking designers to comply with (all or some of) the EU requirements and to provide evidence regarding safety to access the common market. The quality of AI in Europe will depend on the ability of future regulators to establish such specific standards and enforce EU values in the design process of AI systems that aim to access the EU market (De Gregorio 2021). As recognized by the EU Commission only: “a solid European regulatory framework for trustworthy AI will protect all European citizens and help create a frictionless internal market for the further development and uptake of AI” (ECWP 2020, p. 10).

1.2 Is EU policy enough? Safe to use AI for a frictionless market

The main EC agenda is to build a market based on the concept of “enabling trustworthy AI” with a strong focus on data (ECWP 2020 p. 3) pointing toward an unspecified approach of safety-by-design (Hale et al. 2007). Rieder et al. (2020) recently highlighted that the EU plan emphasises the concept of trustworthy AI as a ‘ready-made brand’, even though it is yet unclear what can be concretely operationalised in terms of AI safety.

Data access, re-use, exchange, and interoperability are considered central to the idea of the EU AI market and to facilitate international exchanges based on trustworthiness (ECWP 2020, p. 8). Nevertheless, as recently clarified by a wide survey across Europe that involved a total of 1215 AI stakeholders (EU robotics-AI team 2020; EU robotics and artificial intelligence team AI 2020) there is a concern regarding the diffusion of AI in the EU, with 95% of the respondents expressing uncertainties regarding the real possibility to ensure trustworthiness and the fear of exposing citizens to unsafe or harmful AI systems.

It should be highlighted that, compared for instance to the USA and China, the EU only recently attempted to identify a comprehensive way for embedding AI in society by proposing an original approach compared to its competitors, i.e., putting at the centre the users and their privacy, protected by the General Data Protection Regulation (GDPR EU Regulation 2016/679 2016). The EU approach to AI might result in an over-regulatory effort based on the fear of developing a market at the expense of society (Pereira et al. 2020). Nevertheless, it seems that China is going to enact a new regulation, the Personal Information Protection Law, which in many instances resembles the GDPR (Determann et al. 2021) by enforcing at least companies to minimise the collection and use of personal data. It should be remarked that there is quite a distance between the

Chinese’s social-construction idea of governance where AI is applied to monitor, control and support citizens (Roberts et al. 2021), and the EC proposal of designing with citizens at the centre respecting their rights. Nonetheless, it is worth noting that the EC idea of protecting privacy to strategically build trust toward the market is spreading outside Europe.

Researchers are, however, concerned about this focus on the trustworthiness of AI. Whenever AI systems are based on a limited dataset with unavoidable margins of uncertainties then risks in the usage of AI should be assessed against advantages associated with the usage of such systems (Cabitza et al. 2020). Certainly, trustworthiness cannot only be built over data (Mons 2020) but by developing interaction protocols, strategies, and procedures of interaction to minimise and even anticipate errors in the humans-AI interaction (Cabitza et al. 2020; Rajih et al. 2017). How this interaction will be designed (e.g., by acknowledging data limitation, maximising understandability, etc.) and tested will determine the balance between risks and safety of AI and enable a trustworthy market. This leads us to the current situation: when such interactions with AI are not yet routine, and we must at the same time plan for a future world in which such interactions are ubiquitous. To model such a situation, the next section will shine some insights into the EU proposal for embedding AI in our societies through the lens of the CCF (Hamedani and Markus 2019).

1.3 The gaps in the European vision on AI through the culture cycle framework

We analyse the EC plan through the CCF (Hamedani and Markus 2019; Markus and Kitayama 2010). This framework characterizes sociocultural factors enabling socio-cultural change as four interacting elements (i) ideas of changes, (ii) how these ideas are institutionalized (institutions) in terms of principles, regulations, and guidelines, (iii) interactions, intended here as practices to operationalize these ideas in real-world design and assessment of AI systems. This element contains the methods (i.e., the how) by which it is possible to achieve the institutionalized ideas; and (iv) individuals, intended here as the potential realizers of the impact of the ideas institutionalized in the society (Hamedani and Markus 2019).

This framework is usually applied to map the elements that can facilitate or prevent cultural changes. This approach is used here to highlight what was provided for, and what was left undiscussed (missing) in the EC proposition of changing the culture to enable an AI market across the EU (see Fig. 1).

A change of culture is intended here as a modification (bottom-up or top-down) of explicit and implicit patterns that have been historically embodied in institutions, practices, and artefacts (Adams and Markus 2003; Hamedani and

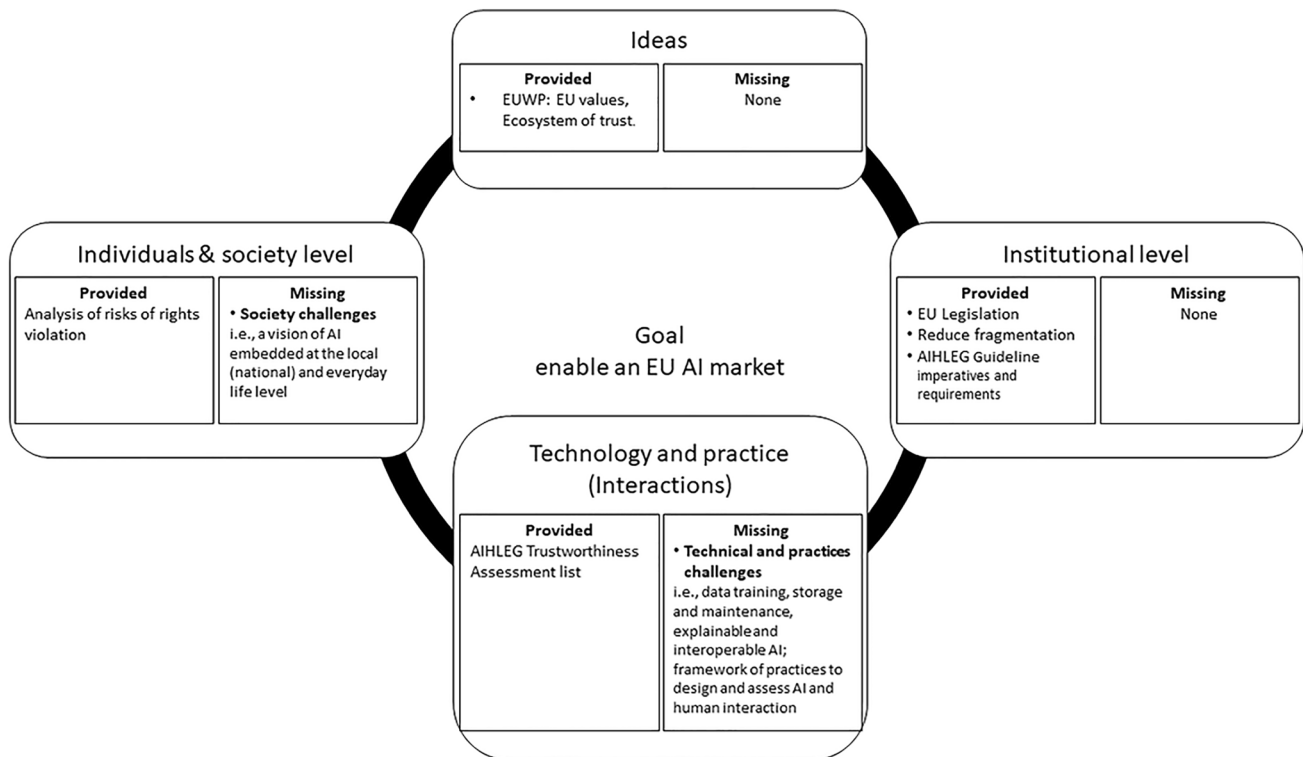


Fig. 1 The EU initiative of enabling a Market of AI. This is observed from the lens of the CCF (adapted with permission from, Hamedani and Markus 2019). Elements provided by the EU documentation to enable the culture change, as well as aspects that are not provided

(missing) are detailed for each component of the model: ideas, the institutionalisation of ideas, technology and practice (interactions) and individuals

Markus 2019). Bottom–up changes happen from individuals to the institutional level by incorporating into formal regulation new habits, ideas and practices that became commonplace patterns in the society. Conversely, top–down changes occur from the institutional level enforcing new regulations that will affect behaviours, practices, and artefacts. People could accept and participate proactively in the change coming from the top or adversely react (culture clash) by feeding up a new cycle from the bottom asking the institutions to adjust the regulation accommodating the (positive/negative) instances coming from the society (Hamedani and Markus 2019). The EC plan is a top–down cycle that aims to bring the EU to compete in the AI global race while concurrently establishing the rules for balancing its internal market.

Starting with *ideas*, the ECWP puts EU values at the centre. The goal of the EC to enable an EU AI market is represented at the *institutional level* by: (i) EU legislation, (ii) the intention to reduce fragmentation and (iii) the definition of AIHLEG-guidelines to mediate a sustainable trade-off between EU values and needs for an AI market. The AIHLEG-guidelines consist of regulations and principles with which AI applications and their development should comply. We did not identify missing points in terms of the institutionalisation of the idea. Nevertheless, when

it comes to the operationalisation of the ideas in terms of how AI should be designed and assessed (i.e., *interactions*, see Fig. 1), only a high-level list of assessment criteria is provided, without a reflection on the challenges and open questions at the technical and methodological level. The gap between defining ethical principles and the operationalised ideas in the real world is not discussed in the EU proposal demanding a self-emergent mechanism from the market to identify solutions to these operational issues. Finally, at *individuals'* level (namely how AI will be used in society by end-users) the ECWP and its associated documentation (MSI-NET 2018) described the potential negative impact of AI on human rights without providing, however, a clear vision on which type of benefits and threats people and states will have to deal with to decide how to implement AI in everyday life. Without a clear vision of the society to be realised, EU states and citizens are left with an open space that could be filled in different ways by states creating differences within the EU.

The race to develop AI follows from the global economic race, as AI is seen to play a key role in increasing, e.g., production efficiency and leading to a positive balance of trade and, ultimately, to public good. For example, similarly to the US (NSTC 2019), the EU attempts to lever AI to become a

global champion in domains such as industry, health, transport, finance, agrifood, energy/environment, forestry, earth observation and space (ECWP 2020). As recently shown with simulation models by Cimpeanu et al. (2020), the use of AI for the public good can come from diversity and cooperation among operators in a market. In this sense, to win the AI development race between nations, something more is required than aiming for AI systems that are safe-for-all (see, Table 1, imperative ii).

The following sections will discuss the gaps we identified in terms of challenges and opportunities regarding AI from the societal (Sect. 2), technical (Sect. 3), and methodological perspective (Sect. 4) to identify strategies to close the culture cycle and to safely embed AI systems in EU societies.

2 Societal integration of AI

The embedding of AI into Western societies can be thought of as a continuum of the regeneration of these societies through creative destruction (Reinert and Reinert 2006). Considering the role of AI in the regeneration of contemporary society brings up a key question: what part of this regeneration is an actual improvement and how can AI be harnessed to improve our societies?

To explore possible answers to this question, we adopted the societal innovation framework, which considers not only the added monetary value but also the added symbolic values of innovation (Lehtola and Stähle 2014). Lehtola and Stähle (2014) separate the economic value that is the traditional measure of innovation, e.g., decrease in production costs, and the symbolic value that innovation represents, e.g., the introduction of gender balance in organizations. Under this framework, the inclusion of AI in our societies differently affects the state and civil society. However, this complexity is often reduced into a balance between two

forces. The driving force behind AI is monetary benefits, while the force counter-acting this rises from people reacting to associated symbolic threats. For example, imposing AI-based surveillance on individuals may be cost-effective in terms of police work (monetary benefit) but may also be seen as a privacy violation (symbolic threat). To account for the changes in the modern global society, we have extended (see Table 2) the model from Lehtola and Stähle (2014) by adding another stakeholder, namely the *international corporations* which are playing a relevant role in the domain of AI development and diffusion.

Following the model of Lehtola and Stähle (2014), we argue that the market-driven AI is a successful societal innovation, i.e., it is successfully embedded into the society, when the balancing of economic factors against the symbolic factors in Table 2 succeeds. For policymakers, it is then important to understand the interplay between AI and the different societal players, namely, the state (Sect. 2.1), civil society (Sect. 2.2), and international corporations (Sect. 2.3).

2.1 State and AI

The course of a nation in Western democracies is decided by elected representatives. The supervision of AI is then, ultimately, their responsibility. Arguably, a challenge follows from the fact that technological systems are typically supervised by experts, but specific expertise is not a typical selection criterion of representatives in Western democracies (Held 2006). In other words, we may end up in situations where leaders are not personally able to understand the technological functions over which they have oversight. And if the elected rulers cannot rule as intended, it is a democratic crisis. Advances in human supervisory control of AI (Musić and Hirche 2017; Peters et al. 2015) allow managing how information is presented to human operators and how the input from the operators is used by systems, and vice versa,

Table 2 Artificial intelligence monetary benefit and the symbolic threat from the perspective of the state, civil society, and international corporations (adapted with permission from: Lehtola and Stähle 2014)

Stakeholders	Monetary benefit by AI	The symbolic threat by AI
State	Increased productivity of state functions (e.g., health care, education, security, city planning, infrastructure construction, and maintenance)	Loss of human supervision, de-stabilization by fake media campaigns (e.g., case Cambridge Analytica), espionage, cyberattacks, hybrid wars
Civil society	Increased productivity of company assets, new business models Potential to deviate less attractive jobs to a machine Higher standard of living: Better services for consumers Personalized healthcare Personalized education	Loss of privacy (e.g., personal data managed by corporations outside Europe), theft of biometric identity (e.g., fake videos), police state (e.g., remote biometric identification), being bossed by the AI Involuntary psychological profiling and destabilization by polarization in social media (e.g., personalized electoral campaigns)
International Corporations	Increased productivity of company assets, new business models (e.g., platform economy)	Loss of competitive advantage and market share due to competitor AI investments

for decision-making. Leveraging these techniques, viable solutions to implement human supervision over AI activities could include:

1. The supervisor of AI delegates the validation of AI output to human experts, when AI is dealing with a domain of human expertise. This is a widespread practice used in, e.g., evaluating construction work. However, human expert judgement becomes subjective if data are hard to interpret, such as with breast cancer images (Kerlikowske et al. 2018). Therefore, multiple experts may be needed.
2. The supervising human looks at outputs generated by different independent AIs or data sources and validates the result only if outputs are coherent. For example, Lehtola and colleagues found that expert ship captains, responsible for human lives and shipping yard property, do not rely on sole sources of information, but information from multiple independent sources before they make decisions about ship routes (Lehtola et al. 2020). This suggests that human-AI decision-making builds on exchanges between reliable information sources. Nevertheless, when signals are not coherent, methods of arbitration should be established to support that appropriate actions are performed by AI or operators (Musić and Hirche 2017)
3. The supervising human does not supervise AI per se but its responsible developers, for example, using a transparent credibility scoring system. To understand such a scoring system, one must first understand that developers' moral obligation towards good is ultimately established from that the code and data are open source (Von Krogh et al. 2012). Hence, a prerequisite for such a system is that the code and data are, in fact, open-source. The power of openness is that it allows for whistleblowing from inside and outside the development hierarchy. The scoring system is for rewarding timely action and development efforts: the credibility of the responsible person(s) is raised whenever reported errors are fixed promptly. It may also be used to establish a hierarchy among developers if a high credibility rating allows for taking over more responsibilities. Openness is also connected to security, see Sect. 3.2.

Properly wielding the AI certainly expands the options of a state to fill its role and better serve civil society. For instance, its role in ensuring the protection of vulnerable populations with public healthcare (Ruof 2004) may be enhanced by more usable services that streamline the relationship between citizens and public administrations (Wirtz et al. 2019). In emergencies, such as the COVID19 outbreak, AI tools supported experts in handling the crisis with applications, e.g., treatment, medication, screening, forecasting

and contact tracing (Lalmuanawma et al. 2020). The beneficial role of AI, for other health-related hazards, has been examined too (Choi et al. 2017; Colubri et al. 2019; Reddy et al. 2020). Decision-makers should also be aware of the technical caveats, such as inaccurate predictions following the data generalization that may lead to wrong decisions.

2.2 Civil society and AI

Technological applications of AI offer multifaceted benefits for the private sector, NGOs, and individuals, i.e., members of civil society. AI tools increase the productivity of companies and introduce new business models based on automation (Aghion et al. 2017). AI systems can support members of civil society with more efficient ways of communication, information retrieval, customer services, travelling/commuting, job finding, and healthcare (Faggella 2020). This better efficiency may be used for work, leisure, or educational purposes, as similar AI functions could be transferable in different domains of our life, e.g., email filters, smart replies, etc. AI algorithms for the recommendation of products and goods, as well as AI-driven approaches to organise and personalise services, are enabling us to perform tasks more efficiently—e.g., grocery shopping. This gain in terms of efficiency is driven by monetary factors and it is supposed to bring more satisfaction and create a return on investment, nevertheless, this is also counterweighted by the (symbolic) threats related to data privacy (Sect. 2.2.1) and to organizational development where AI is assessing human behaviour (Sect. 2.2.2).

2.2.1 Privacy issues: every data we give can be used against us

AI excels in combining different data, without taking a stance whether it is personal or not. Facial photographs, friends' networks, personal preferences, and status updates leave a digital trail that must be protected by the state from the profit- and intelligence-seeking agents. The main concerns of civil society are AI “breaching fundamental rights” and “the use of AI that may lead to discriminatory outcomes” (EU robotics-AI team 2020). These include unwanted remote biometric physiological identification and behavioural tracking.

Citizens' acceptability of systems that can monitor and track people is questionable even in cases of real needs, such as the recent use of track-tracing applications for COVID19. Altmann et al. (2020) recently reported that the approval rate of citizens towards track-tracing apps in different countries was no more than 68% during the pandemic crisis, showing that many citizens were still reluctant in accepting such tools invading their privacy. One example of such an algorithm is Facebook's AI system, which flags users that may

be considering self-harm and looks for ways to respond to these cases, such that the person in need gets automated helpline resources (Constine 2017). The question of which ends justify which means is in the end balanced by a parliamentary process.

Within the EU, citizens are protected by GDPR (EU Regulation 2016/679 2016), but this regulation leaves room for interpretations that can yield significantly different implementations. There are gaps between technical and legal (Haley et al. 2016; Kounadi and Leitner 2014) and ethical and legal requirements (Staunton et al. 2019).

The bits of data we spread around while interacting with systems could potentially be used to harm us. Despite GDPR and other measures, citizens must be vigilant and learn to draw the line regarding what is acceptable and what is violating (even potentially) civil rights. These lines must be somehow communicated to EU policymakers, who should make adjustments to accommodate local interests.

2.2.2 Tailored utopia and hidden Taylorism

Education is commonly provided for classes (or masses) because tailoring to fit individual needs would be prohibitively costly. However, AI could investigate user preferences and patterns, allowing for personalized learning. Detailed insights into learning processes can be obtained, and education adjusted to the needs of students, both at individual and institutional levels (Williamson 2016). Real-time feedback using AI can be applied to predict which students are at risk and teaching resources can be allocated to effective interventions. Certainly, properly done AI assessment promotes equality, since personal biases of teachers are omitted. However, there is the risk of other biases, as AI only performs to the limit of the data it was trained for. Hence, AI might, in a hidden way, maintain historical patterns and inequities. The importance of proper human oversight is then highlighted to resolve any unwanted dystopic elements.

AI also plays a part in the rise of the platform economy that enables a wide range of human activities. The nature of these activities, including benefits and harms, will be determined by “the social, political, and business choices we make” (Kenney and Zysman 2016, p. 61). On one hand, the option to participate in different activities is beneficial, especially to micro-providers in developing economies (Lehdonvirta et al. 2019). On the other hand, if the low wage labour markets in the West boil down into fragmented jobs bossed by AI, this may be seen by some not as a tailored utopia but as a Taylorian dystopia. The state should react appropriately. The turbulence in the job market caused by the diffusion of AI-driven systems has set out a call for welfare that should mitigate the collateral loss of obsolete jobs and help to close the gap between these old jobs and the new ones (Birhane and van Dijk 2020).

2.3 International corporations and AI

Corporations are not exposed to the same benefits and threats as a national company that is part of civil society, because they are not bound by the borders of any single country. This creates opportunities for corporations to evade unfavourable rules. These include well-known cases of tax planning strategies within the EU involving corporations, e.g., Starbucks, Google, Apple, Amazon (Cerioni 2016). Consequently, The Organisation for Economic Co-operation and Development started the “Pillar Two” (OECD 2020) initiative issuing a global minimum tax in all states to avoid tax evasion by corporations.

EU Data Protection Agencies have been vigorously enforcing violations of regional/national data protection law in recent years against tech companies, but few changes have been made to their business models of exchanging free services for personal data (Houser and Voss 2018). This is problematic, given the potential for societal disruption based on such data. For instance, Cambridge Analytica set a well-known example by swaying the electorate through social network users’ participation in their psychological manipulation (Berghel 2018). In other words, Cambridge Analytica used the social media input of individuals to create psychological profiles of them and then used these profiles to manipulate them to vote in the wanted fashion.

2.4 Future society: how the EU may avoid an AI crisis

The EU consists of multiple nations with diverse cultures, which are in different phases of embedding AI. It seems likely that the execution of the EC plan may work out differently in these nations. An analogous example would be the implementation of the European currency, the Euro, which led to different outcomes among nations and ultimately into turbulence in the common market (Hall 2012). Currency as a technology relies heavily on trust. For example, in trust that it is not counterfeited, in that it is accepted by everyone, and in that the value does not suddenly disappear. To maintain such trust and eliminate any abuse, regulations need to be enforced. With AI things are likely to be similar and the successful integration of AI in our society is inherently connected with *how much people can trust that other stakeholders are using AI on similar or agreed terms*. The stability of the Euro was used by all nations to obtain loans with low rates, but then it was revealed that only some of the nations faced problems in paying back those debts (Hall 2012). In retrospect, it was a regulation crisis. With AI, one could speculate that drastic differences in the enforcement of regulations between EU nations could similarly lead to a crisis.

Ultimately, each nation in the EU is going to develop its national adaptation of the EU standards and supervise

their enforcement. We are not claiming that this contextual adaptation is an issue per se, we just want to highlight that there is a risk of clashing regulations within countries which could potentially create disparities in the market and for the end-users. The risk of inappropriate use of AI by EU single states to gain economic or strategic advantages is also recognised by the EU: “If the EU fails to provide an EU-wide approach, there is a real risk of fragmentation in the internal market, which would undermine the objectives of trust, legal certainty and market uptake” (ECWP 2020, p. 10). Solutions to issues associated with overseeing AI, protecting personal data, and efficiently and effectively using AI are likely to emerge not only in national but also in EU-wide context. Integration of the different perspectives and solutions to embed AI in the EU states becomes a key question.

Looking at recent attempts to operationalise Human-AI interaction (HAI), researchers are pointing toward the idea that AI systems should provide a platform over which humans and AI may safely act together (co-act) and compensate for each other's abilities and limitations in a symbiotic relationship (Abbass 2019; Abbass et al. 2016; Bousdekis et al. 2020; Fletcher et al. 2020; Hamann et al. 2016; Jarrahi 2018; Peeters et al. 2021). To co-act AI systems should be designed as tools with programmed intentionality that may evolve autonomously by affecting, as well as being affected by, the exchanges with humans in positive and negative ways. Under this perspective, a safe human-AI relationship cannot be designed by prioritizing humans or AI needs; instead, a systems approach (of design and evaluation) should be adopted, putting at the centre the emergent relationship of humans and AI in the specific context of use aiming for a sustainable impact of this relationship on the society. How this type of future symbiotic relationship will be implemented is an open question. However, we can clarify what technical challenges wait in implementing AI systems that may co-act with humans, in the next section.

3 Technical challenges and opportunities

Several technical challenges should be understood while attempting to embed AI into our societies in a constructive way. AI decision-making is ideally seen by the EC as transparent, reliable, and traceable (ECWP 2020). Technically, this means investigating the following three topics (ECWP 2020): (i) training AI models, (ii) AI data storage and maintenance, and (iii) keeping the AI explainable.

3.1 Training challenges in the era of deep learning

Machine learning consists of computer-based methods of finding regularities in data. These methods, as in common data science, may start with some insights into the

relevant variables that may determine the effectiveness of an approach, like finding the time needed for the delivery of some packages while knowing possible influencers like traffic and location. Deep learning is used as a further step of analysing data available by creating neural networks with multiple layers to, more precisely, predict possible arrival times of packages and thus is more data-driven than theory-driven.

Deep learning (DL) is part of a broader family of AI methods based on artificial neural networks with representation learning (LeCun et al. 2015). The power of DL is that it builds this connection through a *training process*. DL has transformed the field of AI, and now some computational model architectures rival human-level performance in tasks such as image recognition (Krizhevsky et al. 2017) and object detection (Girshick et al. 2014). Despite the power of DL methods, these are black-box systems and thus difficult to interpret. Such techniques lack ways of representing causal relationships and reasoning structures (Marcus et al. 2014).

The success of DL has been fuelled by large sets of labelled data. The mainstream training methods are based on fully supervised learning, which fine-tunes the models using supervised backpropagation. This approach usually requires a huge amount of labelled data, which is the main challenge of supervised learning with DL architectures.

Ideally, training data represent a realistic distribution of real scenarios. Once trained, these models serve as generic feature extractors and can be applied to a wide range of problems using fine-tuning techniques (Oquab et al. 2014). However, this is not typically the case. One of the big challenges is limited training data, and thus the ability to generalize from this data to the real world is often insufficient. Furthermore, training data remains a problem in many domains (e.g., biology, geo-spatial, imaging) due to the lack of publicly available data, the extremely high variability in the objects to detect and classify, the lack of labelled data, and barriers to data sharing such as privacy concerns. Many specialized tasks cannot be easily addressed by only refining pre-trained networks, and not sufficiently large datasets to train a Convolutional Neural Network (CNN) from scratch exists. The implementation of new datasets usually faces two challenges: (i) data collection generates concerns about data usage (e.g., privacy concerns) and (ii) data labelling is often time-consuming and requires expert knowledge to be accurate. For instance, in biomedical imaging publicly available data are scarce. Hospitals are reluctant to share their data for fear of breaching privacy regulations. Many commercial companies are eagerly offering to label large datasets, often omitting that the quality of their work may vary, as it strongly depends on task complexity, e.g., labelling of medical images is more complicated than persons' detection. Statistical methods to judge the quality of labelling

and “weight” the results achieved by different operators have been already proposed (Schlesinger et al. 2017) without solving the uncertainty issues of available benchmarks. Concurrently, it is known that modern AI methods are subject to error to adversarial attacks (Zeng et al. 2019), e.g., minor perturbation of the input data. Therefore, how to design robust deep neural networks remains an open question.

One approach for coping with a limited amount of labelled data is to utilize multi-modal data, such as transfer learning and semi-supervised learning. There are two main benefits of using multi-modal data: (1) Multiple sensors observing the same data can make more robust predictions by detecting changes only when both modalities are present (Zhang et al. 2019); (2) The fusion of multiple sensors can capture complementary information that may not be captured by individual modalities (Cao et al. 2019). Nevertheless, when it comes to multi-modal sensory, AI systems are still behind human performance, for instance, in understanding what is going on in a specific scenario.

The growing need for retrieving heterogeneous information and operating upon it to ensure AI multi-modal learning inspires researchers toward the idea that to progress AI capabilities computing should be on the edge, intended as distributed computing in which data are captured near the source (Chen et al. 2019; Zhou et al. 2019).

3.2 AI data storage and maintenance

Security of data storage and maintenance is a serious concern as adversarial attacks could lead to incorrect model predictions, with severe effects in critical scenarios such as clinical decisions or autonomous driving. Concurrently, the wide usage of AI-powered by large-scale data and DL algorithms in areas of high societal impact raises a lot of concerns regarding fairness, accountability, and transparency of decision-making (Raji et al. 2020). Such systems can potentially lead to discriminatory decisions towards groups of people or individuals based on inherent bias in the data (Buolamwini and Gebru 2018). As an example, Datta et al. (2015) showed that Google’s ad-targeting system was displaying more highly paid jobs to men than to women. Such incidents call for methods that explicitly target bias and discrimination in AI systems while maintaining their predictive power. Similarly, for any specific AI data storage solution, the ethics of that solution can be further investigated once the implementation is known.

For security purposes, AI algorithms, software and methods should be well documented, and ideally, open-source, to minimize the risks of hacked or malfunctioning black-box solutions. Codes for national security purposes can be open source, which enables straightforward auditing options, without being publicly disseminated. A trade-off between openness, control over algorithms, and commercial

or national interest should be established, nevertheless, at least well-documented capabilities and limitations behind the code are extremely important to provide.

Several initiatives in EU countries are proposing centralized databases solutions of data storage, e.g., Health Data Hub, (see: <https://www.health-data-hub.fr/>) and German Medical Informatics Initiative (Gehring and Eulenfeld 2018). On the one hand, centralized solutions offer clarity in terms of administration. On the other hand, decentralized approaches like federated learning (Rieke et al. 2020) may be more realistic for training local models with large datasets overall, while record-keeping in medical datasets could be possible only at a coarse-scale and not at a patient level.

3.3 Explainable AI: what does it do and why?

AI solutions have reached surprising results in very few years. However, these exciting performances have been counterposed by a limited understanding of their rationale: CNNs were outperforming traditional approaches in many domains, without scientists being able to explain it. AI algorithms optimize their ability to perform according to training data, but AI systems have no understanding of the real goal of a human designer: besides good performances, they can lead to incredibly wrong (or unexpected) results. This has induced the public and a large part of the scientific community to take AI as a “black box” for a long time. Only more recently, the explanation of the algorithms’ behaviour has become one of the goals of the AI scientific community (Arrieta et al. 2020). The aim is to increase trust in the reliability of the solutions, despite the complexity of the real world.

The explainability of AI depends on the programmed ability of the system to extrapolate information from the input. Data quality (format, accuracy and completeness) enable AI to operate upon information and perform predictions and recommendations. The ability to interpret and share data among AI in an unambiguous way, so-called semantic interoperability (Brennan et al. 2014), is essential for correct machine interpretation.

Recently van Drunen et al. (2019) suggested that algorithm transparency is a key element for explainable and trustworthy AI, however, while algorithm and information transparency are important, for communicating possible contextual risks, when algorithms are transparent stakeholders may be able to manipulate AI systems in the directions of their preference and bias the systems’ outcomes. Also, data sources may be incomplete or biased towards showing what actors want to show (UNESCO 2021). Moreover, AI systems may learn to misbehave from the information exchange with humans (Bartlett et al. 2022; Mann and O’Neil 2016; Noble 2018). Advancements in approaches to ensure security, adaptability, transparency, and explainability are at the

core of the future bidirectional exchanges between humans and AI (Ezer et al. 2019). HAI is opening an extraordinary opportunity for AI systems to access a world of dynamic data from which systems may constantly grow. To take advantage of the exchange with humans, AI should be designed to interoperate with humans intended here as the possibility for AI to be comprehensible in terms of actions more than explainable. However, enabling human-AI interoperability is only the starting point of several (design and assessment) challenges that practitioners must deal with.

4 Design and assessment: challenges and opportunities

4.1 Looking for consensus on practices for interaction between humans and AI

The imperatives proposed by the EC plan (AIHLEG 2019; ECWP 2020) define the limits around which manufacturers must design the HAI. Policymakers should understand that from the methodological point of view, we are currently in a transition period in which researchers are working on defining best practices to design and assess HAI and to adequately address the social and technical challenges discussed in Sects. 2 and 3. Adjustments to the human–computer interaction framework aiming at addressing these challenges are emerging. Recently, Amershi et al. (2019) proposed a list of design principles for usable AI, and Wallach et al. (2020) attempted to map the relationship between AI and UX. Concurrently, Google (2019) recently proposed a set of tools and recommendations to design human-AI collaborative systems. Along this line, Fletcher et al. (2020) proposed an exploratory work to understand how to benchmark and demonstrate the value of Human-AI teams.

Independently from which perspectives will aggregate more consensus, the common trend is to conceive AI systems as co-active agents (Johnson et al. 2011) that may work in a symbiotic relationship with humans by creating additional value sustainable at the individual and collective level (Abbass et al. 2016; Bousdekis et al. 2020; Fletcher et al. 2020; Hamann et al. 2016; Jarrahi 2018; Peeters et al. 2021). To put at the centre of the design the exchange between AI and humans instead of the user alone is a sort of ‘Copernican revolution’ for the product design (Shneiderman 2020). This new way of design thinking requires to focus on developing a sustainable relationship based on trust and safe co-actions (Burggräf et al. 2021), however, it also opens several ethics questions regarding the complexity of co-action in human-AI teams regarding shared supervision of actions and autonomy (Musić and Hirche 2017; Shneiderman 2020). For instance, as recently reported by Micocci et al. (2021), the diffusion

of AI systems to support diagnostics will certainly bring advantages but it could also expose clinicians and patients to the risks of passive adherence toward the AI indications. As highlighted by Russell et al. (2015) these open questions regarding HAI can only be solved by advancements of design and assessment methods and by understanding “what trade-offs can be made” (Russell et al. 2015, p. 108). As suggested by the US NSTC (2019) there is a growing need to establish new techniques to design and assess AI that can act autonomously by producing understandable and explainable outputs, actions and decisions, but also to define how to design co-active systems that can intuitively be used by multiple users for different purposes.

In line with the EC plan, experts are suggesting that to design AI systems means to develop tools that may interoperate (conjointly exchange data) with humans to perform (together and in parallel) toward a certain goal in a fair, inclusive, responsible and satisfactory way maximising the value of interaction (Fletcher et al. 2020; Google 2019; Johnson et al. 2011). An important intermediate step is to understand how each of the various proposed design methods contributes to maximising the value of the HAI. The understanding of which methods best address which aspects of AI design can clarify issues of safety, performance and ethics in AI since it is unlikely that any single method will address all these aspects (despite what authors may claim). Traditionally, safety and performance/optimization have been the main focus of research in human-autonomy teams (Shah et al. 2011). For instance, Abbass (2019) suggested a design method to enable a safe and trustworthy co-action by providing AI systems with specialised sub-modules or middleware systems that could manage the relationship by taking responsibility of presenting “to the human information at sufficient pace and form suitable for the human to understand and act on, while simultaneously able to translate back and forth the information with its internal components” (Abbass 2019, p. 163). Responsible AI also requires designers to implement systems by acknowledging and respecting individual differences, i.e., inclusiveness and fairness. More recently, scholars in human-autonomy teams have developed methods to address ethical considerations. Design methods proposed by Cimpeanu et al. (2020) to embed AI in a socially constructive way in our society not only focus on safety and promoting safe practices, but on enhancing social inclusion and autonomy of people as well (Table 1, imperatives i, ii, and iii). This means that future AI must be able to (i) communicate proactively with humans by exchanging comprehensible information (i.e., Table 1, explicability, imperative iv), (ii) enable the achievement of short- and long-term goals and needs of people and (iii) be ethical in the inclusion of those who must deal with barriers due to economic and social status, health, well-being (Coeckelbergh 2013).

This section has argued that an important unresolved issue for future implementation of HAI is to draw a line in terms of who is responsible for what, and how humans and AI will co-act together during the interaction to address safety, performance, and inclusion. The next section will propose a perspective on the HAI.

4.2 Extending the perspective on HAI: value, responsibility, and agency

From a utilitarian point of view, AI is bringing monetary and non-monetary (symbolic) value to humans. Concurrently, by performing programmed processes of adaptive fitness AI systems are learning even beyond the original intentionality gaining value from feeding their algorithms with information obtained by the interaction with humans.

The process of AI adaptation to humans’ behaviour and needs is not coming without risks. AI systems might develop unexpected and unwanted misbehaviour (Bartlett et al. 2022; Mann and O’Neil 2016; Noble 2018), with an open question: to whom does the responsibility of this belong?

Certainly, these risks may be compensated by the advantages of having systems trained, through the interaction with humans, which may anticipate when certain dangers arise like, for instance, algorithms to reduce self-harming behaviour from social media (Scherr et al. 2020). The unavoidable uncertainties about how human-AI mutual exchange of information will work out in real-world pave the road to the idea of responsible agency (Akata et al. 2020; Coeckelbergh 2020; Eggink et al. 2020; van Riemsdijk 2020), in which responsibility during the

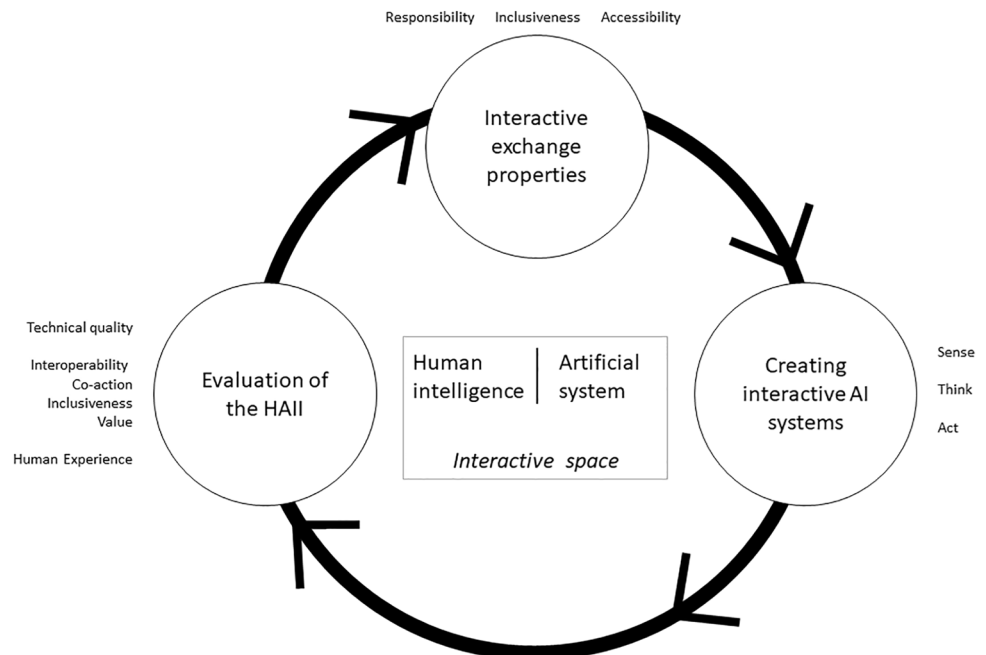
interaction is not located solely in the human or machine but emerges through the interaction and interdependence of both.

Humans and AI are mediators of each other during the interaction, and these two must be designed (AI) or learn how (Humans) to collaborate in a resilient way to achieve their goals (Verbeek 2015). Responsibility for the HAI results will be, therefore, shared by end-users and designers of the AI. As highlighted by Ryan (2020) responsibility and trustworthiness are inherently connected, as the more systems will be designed to collaborate adaptively with people minimising humans-errors the more end-users will perceive such systems as trustworthy—intended here as the ability of the systems to emit signals to gain and maintain trust (Amaral et al. 2020). While past discussions of HAI have been focused on dispositional qualities such as self-confidence and attention span, or system operation qualities such as workload and system complexity (Hoff and Bashir 2015; Lazanyi and Maraczi 2017), society-level ethical norms are also emerging as considerations for trust in automation (Awad et al. 2018; Bonnefon et al. 2016) together with human factors aspects (Rajih et al. 2017).

4.2.1 The future of interaction with AI systems

We argued for the importance of responsibility and inclusiveness in the design of the Human-AI interactive exchanges. Building upon our analysis, we can speculate about three main moments (Fig. 2) that might potentially characterise the future cycles of AI development:

Fig. 2 Cycles of design and assessment of human artificial intelligent interaction



- Interactive exchange properties: which elements of potential friction may emerge in the interaction? Who is responsible for the exchange? Can AI ensure understandability, inclusiveness, and accessibility?
- Creating interactive AI systems: How can we realise the interaction with and through AI? How the system will sense, think, and act during the exchange?
- HAII Evaluation: Which are the technical limitations? How can we assess the trustworthiness and quality from the technical and human experience perspective? How can we assess the value of the exchange in terms of interoperability, co-action and inclusiveness of the AI?

In executing the EC plan, it is necessary to highlight that the HAII quality is not only a matter of trustworthiness toward designers or the technical component, and it is mainly associated with the collaborative exchange that people will experience with the AI. This perceived quality will be determined by enabling adaptive human-AI communication and making responsibilities clear, especially those related to the AI training.

For designing AI architectures, the AIHLEG-guidelines suggest the adoption of the so-called sense-think-act cycle (AIHLEG 2019, p. 21). The use of this model enables designers to reflect on the three steps: the use of sensors to capture the data (sense) that are necessary for the decision-making (think) and to enable the AI to perform wanted tasks (act). Building upon this model (Fig. 2), we are proposing, in line with Johnson et al. (2014), that it is possible to design models of human-AI co-action and interdependency so that, for instance, robots can assist humans during scenarios of disasters. This designed co-activity is an essential component of future ways to develop interactive systems by designing modalities in which humans may assist the robot and vice versa in the execution of tasks accounting for the context. Consider a simple example where a robot can determine where to stand to pick up a hose, but it cannot account for obstacles that will prevent the correct execution of the task. The robot's position needs to be directed by the human, while the human needs information from the robot to reliably determine the optimal position for the robot to execute the task.

To ensure assessment of HAII (see Fig. 2), practitioners need to identify or adapt methods that may support with evidence the designer's decision-making during iterative cycles of testing. Short interaction studies in the lab are not sufficient for understanding the full extent of the human experience of AI technologies in daily life, and a future framework needs to account for a transition from in-vitro research to long-term in-vivo generalizations by the incorporation of intervention-based research building on automatic measurements in the physical world. While consensus around the assessment practice of AI is still far from being established,

we believe that the following three elements should be considered potential key elements for the HAII evaluation:

1. *Oversight on the technical quality* of AI builds on the accuracy and reliability of AI conclusions based on data acquired from citizens or test environments. To assess these aspects, experts must review the systems functioning to ensure usability, accessibility, safety and reliability at the technical level. Moreover, experts should identify ways to take into account in their reviews the potential societal impact of systems by looking, for instance, at the functions allocated to AI and humans (Abbass 2019).
2. *Co-action, inclusiveness, interoperability, value and responsibility of HAII*. AI and humans must interoperate conjointly to add (a society sustainable) value to each other. The assessment of this co-activity should be realised at the “act” level of the design-cycle, to ensure that humans and AI can satisfactorily exchange information and safely achieve their goals and to prevent errors due to miscommunication. This type of assessment incorporates (objective and subjective) usability measurements as well as the accessibility that could be tested with end-users in formative and summative ways (ISO 9241–11 2018). Moreover, this phase of assessment will benefit economics models to establish value-added of the HAII in specific contexts (Borsci et al. 2018b).
3. *Human experience*. Experts should identify appropriate ways to test how people perceive and learn to use (or anticipate) the use of AI and adjust their behaviour to maximise the benefit of the AI. This goes beyond the performance of people aided by AI, and their strategies to optimise the exchange, level of satisfaction, trust, and acceptance. Experience assessment should be performed at the “act” level of the design cycle to investigate needs for adjustment in the ability of AI to exchange information with humans, but also regarding the human behaviour and procedures during the interaction. Under the framework of UX (ISO 9241–210 2010), trust cannot only be considered a quality of the product (trustworthiness) but also as a subjective measurable aspect that should be continuously calibrated over time (Abbass 2019; Borsci et al. 2018a; De Visser et al. 2020; Wijnhoven and Brinkhuis 2015).

The concept of trust toward systems will certainly be a component of the future paradigm of AI development, but we are questioning whether this concept should be a central one, or if this is simply one among the other factors that determine the quality of the HAII. In line with other practitioners' perspectives (Rieder et al. 2020; Ryan 2020) we appreciated the EC intention to achieve a culture of “Trustworthy AI for Europe” (AIHLEG 2019, p. 35), however, it

is unclear whether trust toward AI is the ultimate goal, or if trustworthiness is intended as a convenient way to push the idea of an AI market by diverting responsibility and accountability from developers and users (Ryan 2020). While saying this, we still do recommend that the (national and international) political decision-makers encourage “prosocial and safe-conduct” of operators (Cimpeanu et al. 2020, p. 16). We believe that regulators should be able to define strategies to mitigate risks and to define rules to steer during this transition period in which methods for assessing the HAI are running behind the diffusion of AI systems.

5 Conclusion

The EC master plan has an ambitious vision of societal change. Regulators are working behind the scenes to articulate the rules, requirements, and needs of different stakeholders. This work is going to affect the AI market in Europe and the future ways of conceptualising and designing AI systems for such market (Berg 1997; Star and Strauss 1999; Strauss 1988; Suchman 1996). Our analysis identifies two main gaps between what may be essential for the success of the EC plan, and what is provided by this proposal. These two gaps should be considered by policymakers.

First, the EC plan does not provide a coherent vision on how to drive future decision-making processes at state and local levels. The diffusion of AI systems in our societies is already sparking reactions from individuals and states looking for local solutions to global issues (Birhane and van Dijk 2020). The risk is to have a plethora of local regulations about AI due to different contextual national ambitions, which may result in a very fragmented EU AI market. This can bring an AI crisis rooted in national differences or dystopic applications of AI. The EU should complement its programme with rules and compensatory mechanisms. For instance, the recent EU proposal for harmonized rules on AI (Artificial Intelligence Act 2021) is going toward the direction of building common regulations. However, as noted by Floridi (2021) this proposal is still lacking a pragmatic approach regarding how AI safety is going to be ensured at the design level. In this sense, the EUWP should be intended as a lean plan that is in constant evolution by continuous adjustments. We see the updating of the EU plan as the next cycle of culture change. This updated plan should support, for instance: (i) the heterogeneity of the market by hindering the big players from imposing dominant solutions (Cimpeanu et al. 2020), (ii) ensure safe practices of oversights at the state level, and (iii) include mechanisms for individuals to give feedback regarding violations of civil rights, e.g.,

privacy. Such an updated proposal would increase the understandability of the legal and regulatory landscape driving the practitioners in the right direction.

This brings us to the second gap we identified in the EC plan, concerning a lack of risk analysis regarding open questions at technical and design levels. The EU and its competitors agree that the future is data-driven, however, current technical and methodological practices are running behind the diffusion of AI. The EU safety-first agenda will not suffice for countering any problems because there are concerning technical issues that may potentially expose the EC plan to unforeseen risks such as lack of: appropriate AI training data, safety approaches against adversarial attacks, and practices to balance transparency, explainability, and security. Furthermore, assessment methods are missing to fulfil the currently vague concept of AI and market trustworthiness. While the full impact of AI cannot be predicted until general adoption takes place (Collingridge 1980), we can take measures to speed up the CCF by anticipating the next round. Therefore, we reframed the CCF of the EU plan, based on our analysis (see Fig. 3), to suggest potential future directions that should be considered to operationalise safe and trustworthy (i.e., sustainable) AI in EU societies. We anticipate multiple update cycles for the plan. In these upcoming cycles, the EU plan should also be adjusted to accommodate the bottom-up concerns (so-called ideas in Fig. 3) raised from the open consultation (EU robotics-AI team 2020). These concerns should then become addressed by future regulations (institutional level).

In terms of future directions, AI will become ubiquitous in workplaces and homes with potential unknown effects on societies and civil rights that cannot be left unregulated. The possibility to co-act symbiotically with AI is opening extraordinary opportunities, however, how this symbiotic interaction will be designed to be sustainable (instead of disruptive) bringing added value to our societies is still unknown. To stimulate a discussion, we proposed to operationalise the EU imperatives by putting the human-AI exchange instead of the human alone at the centre of the design process (see Fig. 2). We are sure that several competing methodological models will arise in future. It would be wise for the EC to push for a worldwide research agenda to achieve agreement on methodological frameworks to ensure systematic ways to assess and compare the quality of interaction with AI systems. This will accelerate the culture cycle and speed up the adoption across countries of a revised EC plan, and lead to a successful embedment of AI in our society.

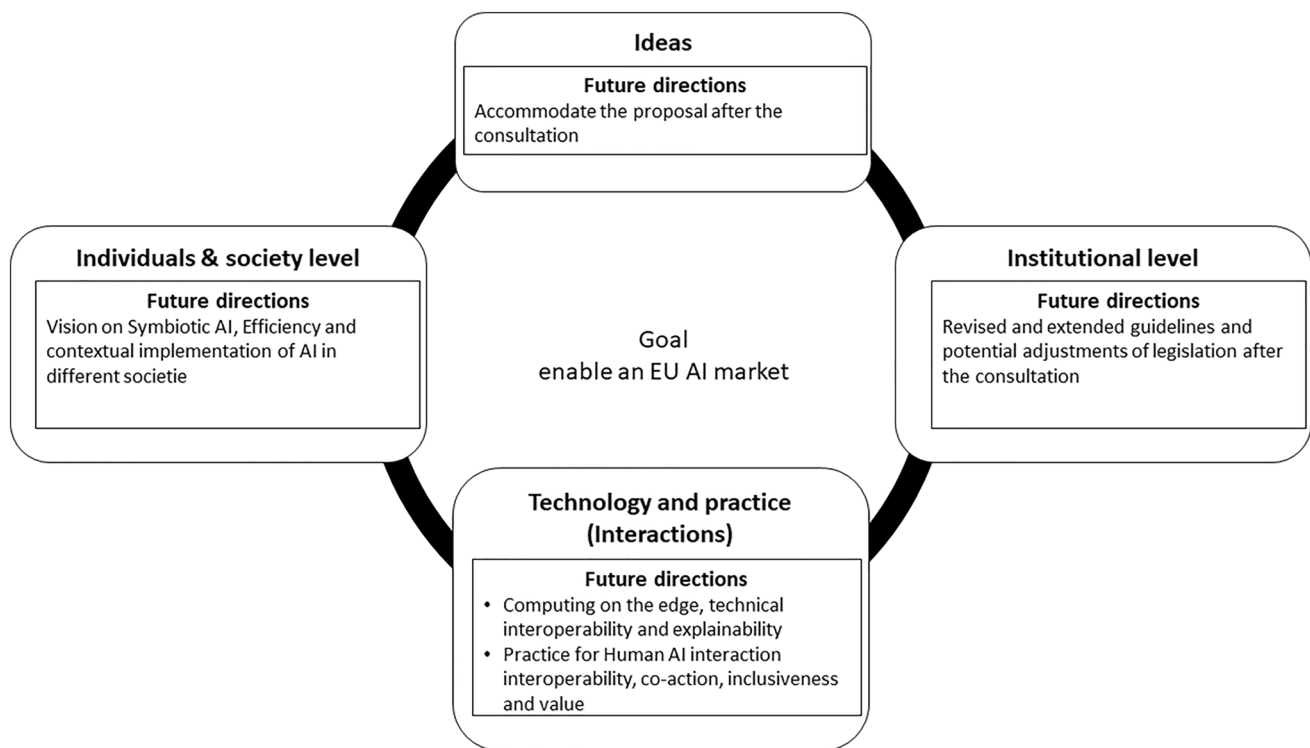


Fig. 3 Revised representation of the EU initiative of enabling a regulated Market of AI observed from the lens of the Culture Cycle Framework (Adapted with permission from, Hamedani and Markus 2019). Elements of these future adaptations of the EU documentation

should discuss and define how to enable the culture change in terms of ideas, the institutionalisation of ideas, the translation of ideas in terms of technology and practice (interactions) and at individuals and society level

Acknowledgements The support from the Digital Society Institute (DSI) of the University of Twente is acknowledged. Also, we would like to credit Dr Birna van Riemsdijk who contributed in terms of conceptualization of the present work, and the formalization of the original draft; Stephanie Hessing who contributed in terms of conceptualization and by reviewing and editing the final draft.

Author contributions All the authors significantly contributed to the article. Simone Borsci conceived and presented the initial idea, and he prepared the first concept draft together with Ville Lehtola, Francesco Nex, and Michael Ying Yang acting as coordinators of the contributions from Ellen-Wien Augustijn, Leila Bagheriye, Christoph Brune, Rania Kounadi, Jamy Li, Joao Moreira, Joanne VanDerNagel, Bernard Veldkamp, Le Viet Duc, Mingshu Wang, Fons Wijnhoven, Jelmer M. Wolterink, and Raul Zurita-Milla. The process of integration and revision of the draft was performed by the coordinators. All the co-authors reviewed the integrated version in an iterative process by significantly reshaping the text. The final version of the article was then drafted by the coordinators and reviewed by the entire interdisciplinary group.

Funding We received no funding for this work, but we were supported by the Digital Society Institute of our University (University of Twente) and this institute is going to pay for the Open Access publication. We acknowledge the support of the Digital Society Institute in the acknowledgement section.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long

as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abbass HA (2019) Social integration of artificial intelligence: functions, automation allocation logic and human-autonomy trust. *Cognit Comput* 11(2):159–171. <https://doi.org/10.1007/s12559-018-9619-0>
- Abbass HA, Petraki E, Merrick K et al (2016) Trusted autonomy and cognitive cyber symbiosis: Open challenges. *Cognit Comput* 8(3):385–408. <https://doi.org/10.1007/s12559-015-9365-5>
- Adams G, Markus HR (2003) Toward a conception of culture suitable for a social psychology of culture. In: Schaller M, Crandall CS (eds) *The psychological foundations of culture*. Lawrence Erlbaum Associates Publishers, pp 344–369
- Aghion P, Jones BF, Jones CI (2017) Artificial intelligence and economic growth. In: Agrawal A, Gans J, Goldfarb A (eds) *The economics of artificial intelligence*. University of Chicago Press. <https://doi.org/10.7208/9780226613475-011>

- AI High-Level Expert Group (AIHLEG) (2019) Ethics guidelines for trustworthy AI. European Commission
- Akata Z, Balliet D, Rijke Md et al (2020) A Research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer* 53(8):18–28. <https://doi.org/10.1109/MC.2020.2996587>
- Altmann S, Milsom L, Zillesen H et al (2020) Acceptability of app-based contact tracing for COVID-19: Cross-Country Survey Study. *JMIR Mhealth Uhealth* 8(8):e19857–e19857. <https://doi.org/10.2196/19857>
- Amaral G, Guizzardi R, Guizzardi G, Mylopoulos J (2020) Ontology-based modeling and analysis of trustworthiness requirements: preliminary results. In: Dobbie G, Frank U, Kappel G, Little SW, MHC (eds) *Conceptual modeling*. Springer, Cham, pp. 342–352. https://doi.org/10.1007/978-3-030-62522-1_25
- Amershi S, Weld D, Vorvoreanu M, et al. (2019) Guidelines for human-AI interaction. In: Proceedings of the 2019 chi Conference on human factors in computing systems. p 1–3. <https://doi.org/10.1145/3290605.3300233>
- Arrieta AB, Díaz-Rodríguez N, Del Ser J et al (2020) Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion* 58:82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Artificial Intelligence Act-Proposal for a regulation of the European Parliament and the Council laying down harmonised rules on Artificial Intelligence and amending certain Union legislative acts (2021). Retrieved December 01, 2021, from <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>
- Awad E, Dsouza S, Kim R et al (2018) The moral machine experiment. *Nature* 563(7729):59–64. <https://doi.org/10.1038/s41586-018-0637-6>
- Bartlett R, Morse A, Stanton R, Wallace N (2022) Consumer-lending discrimination in the FinTech era. *Journal of Financ Econ* 143(1):20–56. <https://doi.org/10.1016/j.jfineco.2021.05.047>
- Berg M (1997) *Rationalizing medical work: decision-support techniques and medical practices*. MIT press, Cambridge, MA
- Berghel H (2018) *Malice domestic: The Cambridge analytica dystopia*. *Computer* 51(5):84–89. <https://doi.org/10.1109/MC.2018.2381135>
- Birhane A, van Dijk J (2020) Robot rights? Let’s talk about human welfare instead. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. p 207–213. <https://doi.org/10.1145/3375627.3375855>
- Bonnefon J-F, Shariff A, Rahwan I (2016) The social dilemma of autonomous vehicles. *Science* 352(6293):1573–1576. <https://doi.org/10.1126/science.aaf2654>
- Borsci S, Buckle P, Walne S, Salanitri D (2018a) Trust and human factors in the design of healthcare technology. In Bagnara S, Tartaglia R, Albolino S, Alexander T (eds) *Advances in intelligent systems and computing*. Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) (Vol. 824, pp. 207–215). Springer, Cham. https://doi.org/10.1007/978-3-319-96071-5_21
- Borsci S, Uchegbu I, Buckle P et al (2018b) Designing medical technology for resilience: integrating health economics and human factors approaches. *Expert Rev Med Devices* 15(1):15–26. <https://doi.org/10.1080/17434440.2018.1418661>
- Bousdekis A, Apostolou D, Mentzas G (2020) A human cyber physical system framework for operator 4.0–artificial intelligence symbiosis. *Manuf Lett* 25:10–15. <https://doi.org/10.1016/j.mfglet.2020.06.001>
- Brennan R, Walshe B, O’Sullivan D (2014) Managed semantic interoperability for federations. *J Netw Syst Manag* 22(3):302–330. <https://doi.org/10.1007/s10922-013-9291-3>
- Buolamwini J, Geburu T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler SA, Wilson C (eds) *Conference on fairness, accountability and transparency* (Vol. 81, pp. 1–15). Proceedings of Machine Learning Research. <https://proceedings.mlr.press/v81/buolamwini18a.html>
- Burggräf P, Wagner J, Saßmannshausen TM (2021) Sustainable interaction of human and artificial intelligence in cyber production management systems. In: Behrens B-A, Brosius A, Hintze W, Ihlenfeldt S, Wulfsberg JP (eds) *Production at the leading edge of technology*. Lecture Notes in Production Engineering. Springer, Berlin, pp 508–517. https://doi.org/10.1007/978-3-662-62138-7_51
- Cabitza F, Campagner A, Sconfienza LM (2020) As if sand were stone. New concepts and metrics to probe the ground on which to build trustable AI. *BMC Med Inf Decis Mak* 20(1):1–21. <https://doi.org/10.1186/s12911-020-01224-9>
- Cao Y, Guan D, Wu Y et al (2019) Box-level segmentation supervised deep neural networks for accurate and real-time multispectral pedestrian detection. *ISPRS J Photogramm Remote Sens* 150:70–79. <https://doi.org/10.1016/j.isprsjprs.2019.02.005>
- Cerioni L (2016) Quest for a new corporate taxation model and for an effective fight against international tax avoidance within the EU. *The. Intertax* 44: 463. Retrieved December 20, 2020, from <https://kluwerlawonline.com/journalarticle/Intertax/44.6/TAXI2016038>
- Chen Z, He Q, Liu L, et al. (2019, 9–11 Aug. 2019) An Artificial Intelligence Perspective on Mobile Edge Computing. 2019 IEEE International Conference on Smart Internet of Things (SmartIoT), Tianjin, China.
- Choi S, Lee J, Kang M-G et al (2017) Large-scale machine learning of media outlets for understanding public reactions to nation-wide viral infection outbreaks. *Methods* 129:50–59. <https://doi.org/10.1016/j.jymeth.2017.07.027>
- Cimpeanu T, Santos FC, Pereira LM, et al. (2020) AI development race can be mediated on heterogeneous networks. *arXiv preprint* 2012.15234.
- Coeckelbergh M (2013) *Human being@ risk: Enhancement, technology, and the evaluation of vulnerability transformations* (Vol. 12). Springer Science & Business Media. <https://doi.org/10.1007/978-94-007-6025-7>
- Coeckelbergh M (2020) Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Sci Eng Ethics* 26(4):2051–2068. <https://doi.org/10.1007/s11948-019-00146-8>
- Collingridge D (1980) *The social control of technology*. St. Martin’s Press
- Colubri A, Hartley M-A, Siakor M et al (2019) Machine-learning prognostic models from the 2014–16 Ebola outbreak: data-harmonization challenges, validation strategies, and mHealth applications. *EClinicalMedicine* 11:54–64. <https://doi.org/10.1016/j.eclinm.2019.06.003>
- COM 237 Report from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Artificial Intelligence for Europe (2018). Retrieved December 20, 2020, from [https://ec.europa.eu/transparency/documents-register/detail?ref=COM\(2018\)237&lang=en](https://ec.europa.eu/transparency/documents-register/detail?ref=COM(2018)237&lang=en)
- Constine J (2017) Facebook rolls out AI to detect suicidal posts before they’re reported. Retrieved December 20, 2020, from <https://techcrunch.com/2017/11/27/facebook-ai-suicide-prevention>.
- Datta A, Tschantz MC, Datta A (2015) Automated experiments on ad privacy settings: a tale of opacity, choice, and discrimination. *Proc Priv Enhanc Technol* 1:92–112
- De Gregorio G (2021) The rise of digital constitutionalism in the European Union. *Int J Const Law* 19(1):41–70. <https://doi.org/10.1093/icon/moab001>


















- De Visser EJ, Peeters MM, Jung MF et al (2020) Towards a theory of longitudinal trust calibration in human–robot teams. *Int J Soc Robot* 12(2):459–478. <https://doi.org/10.1007/s12369-019-00596-x>
- Determann L, Ruan ZJ, Gao T, Tam J (2021) China's draft personal information protection law. *J Data Prot Priv* 4(3):235–259
- Eggink W, Ozkaramanli D, Zaga C, Liberati N (2020) Setting the stage for responsible design. In: Design Research Society, DRS 2020: Synergy, Brisbane, Australia.
- EU Regulation 2016/679 (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 (2016). Retrieved December 20, 2020, from http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf
- EU robotics and artificial intelligence team AI (2020) White Paper on Artificial Intelligence: Public consultation towards a European approach for excellence and trust. Retrieved December 20, 2020, from <https://ec.europa.eu/digital-single-market/en/news/white-paper-artificial-intelligence-public-consultation-towards-european-approach-excellence>
- EU robotics-AI team (2020) White Paper on Artificial Intelligence: public consultation towards a European approach for excellence and trust. Retrieved December 20, 2020, from <https://ec.europa.eu/digital-single-market/en/news/white-paper-artificial-intelligence-public-consultation-towards-european-approach-excellence>
- European Commission (2020) WHITE PAPER on artificial intelligence—a European approach to excellence and trust. In: COM 65 final. https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en
- European Political Strategy Centre (2018) The age of artificial intelligence: Towards a European strategy for human-centric machines. EPSC Strategic Notes 29:1–14. <https://doi.org/10.2872/481078>
- Experts on internet intermediaries (MSI-NET) (2018) ALGORITHMS AND HUMAN RIGHTS Study on the human rights dimensions of automated data processing techniques and possible regulatory implications. Retrieved December 23, 2020, from <https://edoc.coe.int/>, <https://edoc.coe.int/>
- Ezer N, Bruni S, Cai Y, et al. (2019) Trust engineering for human-AI teams. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting 63(1): 322–326. <https://doi.org/10.1177/1071181319631264>
- Faggella D (2020) Everyday examples of artificial intelligence and machine learning. *Emerj*. Retrieved December 23, 2020, from <https://emerj.com/ai-sector-overviews/everyday-examples-of-ai/>
- Federici S, de Filippis ML, Mele ML et al (2020) Inside pandora's box: a systematic review of the assessment of the perceived quality of chatbots for people with disabilities or special needs. *Disabil Rehabil Assist Technol* 15(7):832–837. <https://doi.org/10.1080/17483107.2020.1775313>
- Fletcher SR, Johnson T, Adlon T et al (2020) Adaptive automation assembly: Identifying system requirements for technical efficiency and worker satisfaction. *Comput Ind Eng* 139:105772. <https://doi.org/10.1016/j.cie.2019.03.036>
- Floridi L (2021) The European Legislation on AI: a brief analysis of its philosophical approach. *Philos Technol* 34(2):215–222. <https://doi.org/10.1007/s13347-021-00460-9>
- Gehring S, Eulenfeld R (2018) German Medical Informatics Initiative: Unlocking data for research and health care. *Methods Inf Med* 57(Suppl 1):e46–e49. <https://doi.org/10.3414/ME18-13-0001>
- Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR 2014 : 27th IEEE Conference on Computer Vision and Pattern Recognition, Columbus, Ohio. p 580–587. <https://doi.org/10.1109/CVPR.2014.81>
- Google (2019) PAIR. People + AI Guidebook. <https://pair.withgoogle.com/guidebook>
- Hale A, Kirwan B, Kjellén U (2007) Safe by design: where are we now? *Saf Sci* 45(1):305–327. <https://doi.org/10.1016/j.ssci.2006.08.007>
- Haley DF, Matthews SA, Cooper HL et al (2016) Confidentiality considerations for use of social-spatial data on the social determinants of health: Sexual and reproductive health case study. *Soc Sci Med* 166:49–56. <https://doi.org/10.1016/j.socscimed.2016.08.009>
- Hall PA (2012) The economics and politics of the euro crisis. *Ger Polit* 21(4):355–371. <https://doi.org/10.1080/09644008.2012.739614>
- Hamann H, Khaluf Y, Botev J et al (2016) Hybrid societies: challenges and perspectives in the design of collective behavior in self-organizing systems [Perspective]. *Front Robot AI*. <https://doi.org/10.3389/frobt.2016.00014>
- Hamedani MYG, Markus HR (2019) Understanding culture clashes and catalyzing change: a culture cycle approach. *Front Psychol* 10:700. <https://doi.org/10.3389/fpsyg.2019.00700>
- Held D (2006) Models of democracy. Stanford University Press, Stanford
- Hoff KA, Bashir M (2015) Trust in automation: Integrating empirical evidence on factors that influence trust. *Hum Factors* 57(3):407–434
- Hollnagel E (2009) The ETTO principle: efficiency-thoroughness trade-off: why things that go right sometimes go wrong. Ashgate Publishing Ltd.
- Houser KA, Voss WG (2018) GDPR: The end of Google and facebook or a new paradigm in data privacy. *Rich JL Tech* 25:1
- ISO (2018) ISO 9241–11 Ergonomic requirements for office work with visual display terminals—Part 11: Guidance on usability. CEN, Brussels
- ISO (2010) ISO 9241–210:2010 Ergonomics of human-system interaction—part 210: Human-centred design for interactive systems. Brussels, BE: CEN. Retrieved September 01, 2020, from <http://eu.i2.saiglobal.com/management/home/index>
- Jarrahi MH (2018) Artificial intelligence and the future of work: human-AI symbiosis in organizational decision making. *Bus Horiz* 61(4):577–586. <https://doi.org/10.1016/j.bushor.2018.03.007>
- Johnson M, Bradshaw JM, Feltovich PJ et al (2011) Beyond cooperative robotics: the central role of interdependence in coactive design. *IEEE Intell Syst* 26(3):81–88. <https://doi.org/10.1109/MIS.2011.47>
- Johnson M, Bradshaw JM, Feltovich PJ et al (2014) Coactive design: designing support for interdependence in joint activity. *J Human-Robot Interact* 3(1):43–69. <https://doi.org/10.5898/JHRI.3.1.Johnson>
- Kenney M, Zysman J (2016) The rise of the platform economy. *Issues Sci Technol* 32(3):61
- Kerlikowske K, Scott CG, Mahmoudzadeh AP et al (2018) Automated and clinical breast imaging reporting and data system density measures predict risk for screen-detected and interval cancers: a case–control study. *Ann Intern Med* 168(11):757–765. <https://doi.org/10.7326/M17-3008>
- Kounadi O, Leitner M (2014) Why does geoprivacy matter? The scientific publication of confidential data presented on maps. *J Empir Res Hum Res Ethics* 9(4):34–45. <https://doi.org/10.1177/1556264614544103>
- Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. *Commun ACM* 60(6):84–90. <https://doi.org/10.1145/3065386>
- Lalmuanawma S, Hussain J, Chhakchhuak L (2020) Applications of machine learning and artificial intelligence for Covid-19

- (SARS-CoV-2) pandemic: a review. *Chaos Solitons Fractals*. <https://doi.org/10.1016/j.chaos.2020.110059>
- Lazanyi K, Maraczi G (2017) Dispositional trust—do we trust autonomous cars? In: 2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, Serbia. p 000135–000140. <https://doi.org/10.1109/SISY.2017.8080540>
- LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553):436–444. <https://doi.org/10.1038/nature14539>
- Lehdonvirta V, Kässi O, Hjorth I et al (2019) The global platform economy: a new offshoring institution enabling emerging-economy microproviders. *Jmanag* 45(2):567–599. <https://doi.org/10.1177/0149206318786781>
- Lehtola VV, Stähle P (2014) Societal innovation at the interface of the state and civil society. *Innovation* 27(2):152–174. <https://doi.org/10.1080/13511610.2014.863995>
- Lehtola VV, Montewka J, Salokannel J (2020) Sea Captains' views on automated ship route optimization in Ice-covered Waters. *J Navig* 73(2):364–383. <https://doi.org/10.1017/S0373463319000651>
- Mann G, O'Neil C (2016) Hiring algorithms are not neutral. *Harv Bus Rev* 9:2016. Retrieved January 15, 2020, from <https://hbr.org/2016/12/hiring-algorithms-are-not-neutral>
- Marcus G, Marblestone A, Dean T (2014) The atoms of neural computation. *Science* 346(6209):551–552
- Markus HR, Kitayama S (2010) Cultures and selves: a cycle of mutual constitution. *Perspect Psychol Sci* 5(4):420–430. <https://doi.org/10.1177/1745691610375557>
- Micocci M, Borsci S, Thakerar V et al (2021) Attitudes towards trusting artificial intelligence insights and factors to prevent the passive adherence of GPs: a pilot study. *J Clin Med* 10(14):3101. <https://doi.org/10.3390/jcm10143101>
- Mons B (2020) Invest 5% of research funds in ensuring data are reusable. *Nature* 578(7796):491–491. <https://doi.org/10.1038/d41586-020-00505-7>
- Musić S, Hirche S (2017) Control sharing in human-robot team interaction. *Annu Rev Control* 44:342–354. <https://doi.org/10.1016/j.arcontrol.2017.09.017>
- National Science Technology Council (2019) The national artificial intelligence research and development strategic plan: 2019 update. National Science and Technology Council (US)-Committee on Artificial Intelligence. Retrieved October 01, 2021, from <https://www.hSDL.org/?abstract&did=831483>
- Noble SU (2018) Algorithms of oppression: How search engines reinforce racism. NYU Press
- OECD (2020) Tax challenges arising from digitalisation—report on Pillar Two Blueprint. <https://doi.org/10.1787/abb4c3d1-en>
- Oquab M, Bottou L, Laptev I, Sivic J (2014) Learning and transferring mid-level image representations using convolutional neural networks. In: CVPR 2014 : 27th IEEE Conference on Computer Vision and Pattern Recognition, Columbus, Ohio. p 1717–1724. <https://doi.org/10.1109/CVPR.2014.222>
- Peeters MM, van Diggelen J, Van Den Bosch K et al (2021) Hybrid collective intelligence in a human–AI society. *AI Soc*. <https://doi.org/10.1007/s00146-020-01005-y>
- Pereira LM, Santos FC, Lenaerts T (2020) To regulate or not: A social dynamics analysis of an idealised ai race. *J Artif Intell Res* 69:881–921
- Peters JR, Srivastava V, Taylor GS et al (2015) Human supervisory control of robotic teams: integrating cognitive modeling with engineering design. *IEEE Control Syst Mag* 35(6):57–80. <https://doi.org/10.1109/MCS.2015.2471056>
- Rajaraman V (2014) John McCarthy—father of artificial intelligence. *Resonance* 19(3):198–207. <https://doi.org/10.1007/s12045-014-0027-9>
- Raji ID, Smart A, White RN, et al. (2020) Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: FAT* '20: Conference on fairness, accountability, and transparency, Barcelona, Spain. p 33–44. <https://doi.org/10.1145/3351095.3372873>
- Rajih E, Tholomier C, Cormier B et al (2017) Error reporting from the da Vinci surgical system in robotic surgery: A Canadian multispecialty experience at a single academic centre. *Can Urol Assoc J Journal De L'association Des Urologues Du Canada* 11(5):E197–E202. <https://doi.org/10.5489/auaj.4116>
- Reddy A, Soni B, Reddy S (2020) Breast cancer detection by leveraging Machine Learning. *ICT Express* 6(4):320–324. <https://doi.org/10.1016/j.ict.2020.04.009>
- Reinert H, Reinert ES (2006) Creative destruction in economics: Nietzsche, Sombart, schumpeter. In: Drechsler JGBW (ed) Friedrich Nietzsche (1844–1900): economy and society. Springer, pp 55–85
- Rességuier A, Rodrigues R (2020) AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data Soci*. <https://doi.org/10.1177/2053951720942541>
- Rieder G, Simon J, Wong P-H (2020) Mapping the stony road toward trustworthy AI: expectations, problems, conundrums (October 23, 2020). In: Marcello P, Teresa S (Eds) *Machines we trust: perspectives on dependable AI*. MIT Press, Cambridge, MA. <https://ssrn.com/abstract=3717451>
- Rieke N, Hancox J, Li W, et al. (2020) The future of digital health with federated learning. arXiv preprint arXiv: 2003.08119.
- Roberts H, Cows J, Morley J et al (2021) The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. *AI & Soc* 36(1):59–77. <https://doi.org/10.1007/s00146-020-00992-2>
- Ruof MC (2004) Vulnerability, vulnerable populations, and policy. *Kennedy Inst Ethics J* 14(4):411–425. <https://doi.org/10.1353/ken.2004.0044>
- Russell S, Dewey D, Tegmark M (2015) Research priorities for robust and beneficial artificial intelligence. *AI Mag* 36(4):105–114
- Ryan M (2020) In AI we trust: ethics, artificial intelligence, and reliability. *Sci Eng Ethics* 26(5):2749–2767. <https://doi.org/10.1007/s11948-020-00228-y>
- Scherr S, Arendt F, Frissen T, Oramas MJ (2020) Detecting intentional self-harm on Instagram: development, testing, and validation of an automatic image-recognition algorithm to discover cutting-related posts. *Soc Sci Comput Rev* 38(6):673–685. <https://doi.org/10.1177/0894439319836389>
- Schlesinger D, Jug F, Myers G, et al. (2017) Crowd sourcing image segmentation with iastaple. In: ISBI 2017: IEEE 14th International Symposium on biomedical imaging, melburne, Australia. p 401–405. <https://doi.org/10.1109/ISBI.2017.7950547>
- Shah J, Wiken J, Williams B, Breazeal C (2011) Improved human-robot team performance using chaski, a human-inspired plan execution system. In: Proceedings of the 6th International Conference on Human-robot interaction, Lausanne, Switzerland. p 29–36. <https://doi.org/10.1145/1957656.1957668>
- Shneiderman B (2020) Human-centered artificial intelligence: Three fresh ideas. *AIS Trans Human-Comput Interact* 12(3):109–124. <https://doi.org/10.17705/1thci.00131>
- Star SL, Strauss A (1999) Layers of silence, arenas of voice: the ecology of visible and invisible work. *Comput Supported Cooper Work (CSCW)* 8(1):9–30. <https://doi.org/10.1023/A:1008651105359>
- State Council Document No. 35 New Generation of Artificial Intelligence Development Plan (2017). Retrieved October 01, 2021, from http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm
- Staunton C, Slokenberga S, Mascalzoni D (2019) The GDPR and the research exemption: considerations on the necessary safeguards for research biobanks. *Eur J Human Genet* 27(8):1159–1167. <https://doi.org/10.1038/s41431-019-0386-5>

- Strauss A (1988) The articulation of project work: an organizational process. *Sociol Q* 29(2) 163–178. Retrieved November 15, 2021 from <https://www.jstor.org/stable/4121474>
- Suchman LA (1996) Supporting articulation work. In: Kling R (ed) *Computerization and controversy: value conflicts and social choices*, 2nd edn. Academic Press, San Diego, pp 407–423
- UN General Assembly (1948) Universal declaration of human rights. Retrieved October 01, 2021, from <https://www.un.org/en/about-us/universal-declaration-of-human-rights>
- UNESCO (2021) SHS/IGM-AIETHICS/2021/JUN/3 Rev.2—Draft text of the recommendation on the ethics of artificial intelligence. Retrieved October 21, 2021, from <https://unesdoc.unesco.org/ark:/48223/pf0000373434>
- van Drunen MZ, Helberger N, Bastian M (2019) Know your algorithm: what media organizations need to explain to their users about news personalization. *E.int Data Privacy Law* 9(4):220–235. <https://doi.org/10.1093/idpl/ipz011>
- van Riemsdijk MB (2020) Artificial intelligence, data science & intimate computing. Retrieved 1st of December 2019 from <https://intimate-computing.net/intimate-computing-vulnerability/artificial-intelligence-data-science-intimate-computing/#responsible-agency>
- Verbeek P-P (2015) COVER STORY beyond interaction: a short introduction to mediation theory. *Interactions* 22(3):26–31. <https://doi.org/10.1145/2751314>
- Von Krogh G, Haefliger S, Spaeth S, Wallin MW (2012) Carrots and rainbows: Motivation and social practice in open source software development. *MIS Q* 36(2):649–676. <https://doi.org/10.2307/41703471>
- Wallach DP, Flohr LA, Kaltenhauser A (2020) Beyond the buzzwords: on the perspective of AI in UX and Vice Versa. *International Conference on human-computer interaction*, vol 12217. Copenhagen, Denmark. pp 146–166
- Wang P (2019) On defining artificial intelligence. *Journal of Artificial General Intelligence* 10(2):1–37. <https://doi.org/10.2478/jagi-2019-0002>
- Wijnhoven F, Brinkhuis M (2015) Internet information triangulation: Design theory and prototype evaluation. *J Assoc Inf Sci Technol* 66(4):684–701. <https://doi.org/10.1002/asi.23203>
- Williamson B (2016) Digital education governance: data visualization, predictive analytics, and ‘real-time’ policy instruments. *J Educ Policy* 31(2):123–141. <https://doi.org/10.1080/02680939.2015.1035758>
- Wirtz BW, Weyerer JC, Geyer C (2019) Artificial intelligence and the public sector—applications and challenges. *Int J Public Adm* 42(7):596–615. <https://doi.org/10.1080/01900692.2018.1498103>
- Zeng X, Liu C, Wang Y-S, et al (2019) Adversarial attacks beyond the image space. In: *CVPR 2019: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, Long Beach, CA, USA. pp 4297–4306. <https://doi.org/10.1109/CVPR.2019.00443>
- Zhang Z, Vosselman G, Gerke M et al (2019) Detecting building changes between airborne laser scanning and photogrammetric data. *Remote Sens* 11(20):2417. <https://doi.org/10.3390/rs11202417>
- Zhou Z, Chen X, Li E et al (2019) Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proc IEEE* 107(8):1738–1762. <https://doi.org/10.1109/JPROC.2019.2918951>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Simone Borsci^{1,2}  · Ville V. Lehtola³  · Francesco Nex³  · Michael Ying Yang³  · Ellen-Wien Augustijn⁴  ·
Leila Bagheriye^{5,6}  · Christoph Brune⁷  · Ourania Kounadi^{4,8}  · Jamy Li^{9,10}  · Joao Moreira¹¹  ·
Joanne Van Der Nagel¹⁰  · Bernard Veldkamp¹  · Duc V. Le¹²  · Mingshu Wang^{4,13}  · Fons Wijnhoven¹⁴  ·
Jelmer M. Wolterink⁷  · Raul Zurita-Milla⁴ 

Ville V. Lehtola
v.v.lehtola@utwente.nl

Francesco Nex
f.nex@utwente.nl

Michael Ying Yang
michael.yang@utwente.nl

Ellen-Wien Augustijn
p.w.m.augustijn@utwente.nl

Leila Bagheriye
leila.bagheriye@donders.ru.nl

Christoph Brune
c.brune@utwente.nl

Ourania Kounadi
ourania.kounadi@univie.ac.at

Jamy Li
jamy@ryerson.ca

Joao Moreira
j.luizrebelomoreira@utwente.nl

Joanne Van Der Nagel
j.e.l.vandernagel@utwente.nl

Bernard Veldkamp
b.p.veldkamp@utwente.nl

Duc V. Le
v.d.le@utwente.nl

Mingshu Wang
mingshu.wang@glasgow.ac.uk

Fons Wijnhoven
a.b.j.m.wijnhoven@utwente.nl

Jelmer M. Wolterink
j.m.wolterink@utwente.nl

Raul Zurita-Milla
r.zurita-milla@utwente.nl

¹ Department of Learning, Data Analysis, and Technology, Faculty of Behavioural, Management and Social Sciences, University of Twente, Enschede, The Netherlands

- ² Department of Surgery and Cancer, Faculty of Medicine, NIHR London IVD, Imperial College of London, London, UK
- ³ Department of Earth Observation Science, Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, The Netherlands
- ⁴ Department of Geo-Information Processing, Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, The Netherlands
- ⁵ Computer Architecture Design and Test for Embedded Systems (CAES), Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS), University of Twente, Enschede, The Netherlands
- ⁶ Foundations of Natural and Stochastic Computing, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands
- ⁷ Department of Applied Mathematics, Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS), University of Twente, Enschede, The Netherlands
- ⁸ Department of Geography and Regional Research, University of Vienna, Vienna, Austria
- ⁹ Department of Mechanical and Industrial Engineering, Ryerson University, Toronto, Canada
- ¹⁰ Department of Human Media Interaction (HMI), Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS), University of Twente, Enschede, The Netherlands
- ¹¹ Services and Cyber-Security (SCS) Group, Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS), University of Twente, Enschede, The Netherlands
- ¹² Pervasive System Group, Department of Computer Science, Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS), University of Twente, Enschede, The Netherlands
- ¹³ School of Geographical & Earth Sciences, University of Glasgow, Glasgow, UK
- ¹⁴ Department of Industrial Engineering and Business Information Systems, Faculty of Behavioural, Management and Social Sciences, University of Twente, Enschede, The Netherlands