AI Ethics as Applied Ethics

Hallamaa, Jaana

2022-04-07

Hallamaa , J & Kalliokoski , T 2022 , ' AI Ethics as Applied Ethics ' , Frontiers in computer science , vol. 4 , pp. 12 . https://doi.org/10.3389/fcomp.2022.776837

http://hdl.handle.net/10138/342701
https://doi.org/10.3389/fcomp.2022.776837

frontiers | Frontiers in Computer Science

ORIGINAL RESEARCH
published: 07 April 2022
doi: 10.3389/fcomp.2022.776837

AI Ethics as Applied Ethics

Jaana Hallamaa and Taina Kalliokoski*

Faculty of Theology, University of Helsinki, Helsinki, Finland

The need to design and develop artificial intelligence (AI) in a sustainable manner has motivated researchers, institutions, and organizations to formulate suggestions for AI ethics. Although these suggestions cover various topics and address diverse audiences, they share the presupposition that AI ethics provides a generalizable basis for designers that is applicable to their work. We propose that one of the reasons the influence of current ethical codes has remained modest, may be the conception of the applied ethics that they represent. We discuss bioethics as a point of reference for weighing the metaethical and methodological approaches adopted in AI ethics, and propose that AI ethics could be made more methodologically solid and substantively more influential if the resources were enriched by adopting tools from fields of study created to improve the quality of human action and safeguard its desired outcomes. The approaches we consider to be useful for this purpose are the systems theory, safety research, impact assessment approach, and theory of change.

Keywords: AI ethics, applied ethics, bioethics, safety research, systems approach

Edited by:
Rebekah Ann Rousi,
University of Vaasa, Finland

Reviewed by:
José Juan Cañas,
University of Granada, Spain
Bernd Carsten Stahl,
De Montfort University,
United Kingdom

****Correspondence:***
Jaana Hallamaa
jaana.hallamaa@helsinki.fi

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *14 September 2021*
Accepted: *26 January 2022*
Published: *07 April 2022*

Citation:
Hallamaa J and Kalliokoski T (2022) AI Ethics as Applied Ethics. Front. Comput. Sci. 4:776837. doi: 10.3389/fcomp.2022.776837

INTRODUCTION

A plethora of suggestions already exists on ethically developing artificial intelligence (AI) and designing ethically sound AI systems[1]. The existing formulations for AI ethics cover various topics and address diverse types of audiences. We use the term "AI Ethics" to refer to guidelines, declarations, ethical codes, and generalizations that display a range of styles and authors, from high-profile declarations by governmental agencies to normative rules-of-thumb for users and practical checklists for designers written by private companies, professional associations and non-profit organizations (IEEE Global Initiative on Ethics of Autonomous Intelligent Systems, 2017; The Montréal Declaration for a Responsible Development of Artificial Intelligence, 2018; AI HLEG, 2019; Jobin et al., 2019; Mittelstadt, 2019, 501; Hagendorff, 2020). Despite these differences, the various guidelines of AI ethics have a common goal: to support and improve the ethical development, design and deployment of AI. Another factor that these formulations share is their presupposition that AI ethics provides a general basis—with respect to theory, principles, values, etc.—that the designers can then apply to single cases, thereby improving the ethical sustainability

[1]We follow the description of AI system made by European Commission High-Level Expert Group on Artificial Intelligence in the document "Ethics guidelines for trustworthy AI" (2019) that goes as follows: "Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behavior by analyzing how the environment is affected by their previous actions".

of AI development (Morley et al., 2019; Hagendorff, 2020). Our use of the term "AI Ethics" does not cover the ethical inquiries of academic philosophers concerning the ethics of AI, even though we discuss their ideas.

The evidence gathered from AI ethics applications shows that attempts to improve ethical sustainability have thus far not been particularly effective (Hagendorff, 2020, 109–110). Furthermore, ethical codes thus far seem to have had a minimal effect on the moral decision-making of software engineers (McNamara et al., 2018) and the ethicality of AI applications, partly because they lack reinforcement mechanisms (Hagendorff, 2020, 113). AI ethicists have made suggestions on rectifying this situation (cf., Mittelstadt, 2019; Morley et al., 2019; Hagendorff, 2020). The suggested measures are likely to remain without the desired effect in that AI ethics means applying ethical theories to create morally acceptable practices, if the adopted approach is ineffective or incorrect (Rességuier and Rodrigues, 2020).

To discuss this matter, we first create an outline of the applied ethics model that seems to represent the current versions of AI ethics. We then consider the arguments concerning why AI ethics has not had much effect on reality, and suggest how the situation could be rectified. To offer a point of comparison and better understand the nature of AI ethics as applied ethics, we present a brief history of bioethics—the most developed and well-known field of applied ethics. By analyzing the methodological discussions on bioethics and critically considering the ways in which bioethics influence real-life problem solving, we wish to demonstrate that the conception of applied ethics, adopted in AI ethics to date, is too narrow and partly inconsistent. We conclude our article by suggesting that AI ethics could be made more methodologically solid and substantively more influential, if its resources were enriched by adopting tools from fields of study created to improve the quality of human action and safeguard its desired outcomes. The approaches we consider to be useful for this purpose are the systems theory, safety research, impact assessment approach, and theory of change.

AI Ethics—Theoretical Models for Practical Applications

From a philosophical perspective, the different formulations of AI ethics represent an understanding of AI ethics guidelines as formulations of applied ethics. The documents explore how to conceptualize, analyze, and assess ethically relevant features of AI design and application as well as how to determine methods of directing, regulating, and governing AI design and its use in an ethically sustainable manner.

Even if AI ethics documents differ in how they balance the objective of guaranteeing financial profitability and social sustainability, they share two important assumptions. First, the similarities between the suggested formulations of AI ethics indicate that there exists a common will to develop AI and AI-based technologies in a morally responsible way. Second, these documents share an understanding of how the proposed guidelines are intended to improve the world. Despite the substantial differences in AI ethics formulations, they propose to offer practical ethical guidance and support for those who design, develop, produce, and use AI applications and systems to perform the task that the AI in question is supposed to accomplish in a morally decent way.

To achieve the aim of creating more ethically sustainable methods of designing and using AI, AI ethics documents and initiatives present high moral values, ethical guidelines, principles, definitions, etc. Those who engage with AI should then apply the normative apparatus (explicated in the AI ethics code) to detect the problems and dilemmas they encounter in relation to AI, and to solve them in an ethically sound manner that is assisted by moral values and ethical principles. This view reflects an understanding of applied ethics, by applying the moral theory to empirical practice. Hence, the task of AI ethics is to analyze, clarify, and solve practical problems by relying on ethical theories and principles, and to study the reality in which AI applications and systems are designed and used (Mittelstadt, 2019, 501).[2]

To highlight the understanding of applied ethics that the current AI ethics documents represent, we examine one of the best-known versions of AI ethics in detail, i.e., "*Ethics guidelines for trustworthy artificial intelligence* (2019)," an AI ethics document created by High-level expert group on artificial intelligence (AI HLEG) determined by the European Commission (EC). We have chosen to highlight this document because it is the most extensive attempt thus far to create an ethical tool for guiding AI development and design. The AI HLEG comprises various actors and institutions (such as private companies, government agencies, universities, and professional associations) that have developed ethical guidelines for AI development and design (Jobin et al., 2019).

The term "trustworthy AI" is also used in AI strategy documents published in the United States, which can be interpreted as "ethical AI." To advance trustworthiness—that is, fair, non-discriminatory, transparent, safe, and secure AI—the United States' government funds research and development, helps its institutions and private actors to regulate AI applications, and advances standardized assessment tools to evaluate AI systems (National Artificial Intelligence Initiative, 2020).

The Institute of Electric and Electronics Engineers (IEEE) also presents standardization as a way to improve ethical AI design. The IEEE's *Ethically aligned design*—a global initiative on the ethics of autonomous and intelligent systems—sets five goals for developing intelligent autonomous systems: (1) No infringement on human rights; (2) Prioritization of human wellbeing; (3) Ensuring the accountability of designers and operators; (4) Ensuring the transparency of the system's operations; and (5) Minimizing the risks of misuse. To further its ambitious endeavor, the IEEE has translated this document into multiple languages, created various standards that rely on globally approved metrics and assessment tools, and opened an anonymous ethical transgressions' reporting channel for whistleblowers (IEEE Global Initiative on Ethics of Autonomous Intelligent Systems, 2017; Dignum, 2020, 222–223). However,

[2]Canca (2020) is a good example of such understanding of AI ethics as applied ethics.

these ethical guidelines do not have the same potential legislative power to regulate the development, deployment, and use of AI systems as the AI HLEG's *Trustworthy AI* document.

The AI HLEG's *Trustworthy AI* has been prepared in cooperation with several experts from different fields. Even though the expert group was set and the paper published by the EC, it is not officially an European Union (EU) document. However, it is the most prestigious of all the current AI ethics documents, because it serves as an ethical framework for the EU's proposal of AI legislation. *The AI Act* was published in 2021. In our analysis, we refer to other recent AI ethics codes.

AI Ethics as Applying Moral Principles to Practical Problems

The AI HLEG's *Trustworthy AI* (2019) aims to combine two things. First, the ethics document has been compiled and published to provide a competitive asset for European national economies, particularly against the United States and China. Second, it aims to introduce norms for developing *human-centric AI*, based on European values. Interestingly, these objectives concern two different moral anthropologies: the first promotes the aims of the *homo economicus* to succeed in international competition, and the second promotes the values of the European humanistic tradition.[3]

The main goals of the ethical guidelines for *human-centric AI* are fostering a commitment to using AI in the service of humanity and the common good, aiming to improve human welfare and freedom, and handling the arising risks appropriately and proportionately. Developers have noted that there is currently an important window of opportunity for technological and economic success. Making use of this opportunity is intended to evoke more trust in socio-technical AI environments among citizens and in societies in general. Embedding the principles of *Trustworthy AI* in products and services would pave the way for developing ethically high-class AI solutions which could have a competitive advantage in the global market. The idea is that being ethical will enhance economic competitiveness, which will serve both moral anthropologies voiced in the document.

To realize its goals, the document sums up two norms—one forward-looking and positive and the other cautious and negative—as guidelines for strategic planning. The development of AI should aim to maximize the benefits of AI systems, while simultaneously preventing and minimizing the risks associated with using these applications. According to Renda (2020, 653), ethics guidelines should be considered as part of an overall strategy for protecting European citizens from the abuses of digital technology, and for strengthening Europe's position in the global digital development competition.

The corpus of European AI ethics comprises a system of values and principles, where applying the norms to concrete cases and issues will translate into ethically sound action, i.e., a mechanism that mediates moral ideals into the reality of citizens and societies. The primary normative level of the document comprises ethical ideals that present a basis for a good and virtuous life. These ideals provide a framework for understanding morally relevant features, and for conceptualizing the moral constituents of a given activity within its context.

The second level presents guidelines for deliberation and moral reasoning to help developers, designers, and users think before acting and make various decisions during the development and maintenance processes. Another asset that serves ethical deliberation is the criterion for choosing goals and actions. The purpose of these tools is to ensure that the chosen goals (and the actions designed to attain them) are morally acceptable. The document also contains checklists and rules-of-thumb for supporting practitioners in making it a routine procedure to check their products, applications, and services for moral flaws. The last level of normative assessment takes place *ex post facto*, when the choices made (and the outcomes that resulted from them) are scrutinized and judged.

Influence of AI Ethics on AI Design and Development

"The Guidelines received, overall, a warm welcome by policymakers inside and outside of Europe, as well as by large and small companies and civil society," wrote Renda (2020, 662) in the Oxford Handbook of Ethics of AI, which creates a hopeful atmosphere; however, the actual changes in the actions of AI system developers are yet to be documented.

Furthermore, studies on the influence of other ethical guidelines for engineers are not encouraging. Studying effects of the code of ethics published by the Association of Computing Machinery (ACM) on software engineers' ethical decision-making (McNamara et al., 2018), the researchers concluded that the ethical codes had negligible influence on the work of the engineers who took part in the research. There were no meaningful differences between the decisions made by two groups of software engineers and software engineering students: those who had acquainted themselves with the ACM code of ethics and those who had not (ACM, 2018; McNamara et al., 2018).

There may be several reasons why ethical codes do not have a greater effect on real-life decision-making, and those who find the situation problematic have suggested ways to improve the situation. According to Mittelstadt (2019, 502–503), one of the main reasons for the negligible influence of ethical codes on AI design is that it is challenging to apply the existing ethical principles to AI design and development. In cases where technologists and engineers have received ethical training, the emphasis has usually been on morality as an individual trait (Howard, 2019, 7–8). Hagendorff's suggestion for more effective AI ethics follows the same line of thought. According to this view, virtue ethics can change the mindsets of individual agents in AI design (Hagendorff, 2020, 112–113). If people are educated to adopt virtues, in their families and at schools, as well as within various communities and even in companies, they could cultivate a moral character, thereby improving ethical decision-making practices in organizations that develop and deploy AI applications.

[3]For the philosophical differences in moral anthropologies, refer to Hallamaa (1994).

Nevertheless, following Hagendorff's suggestion to rectify the current situation would be time consuming and the effects of the enterprise are far from certain, if not completely contested (Harman, (1998–1999); Doris, 2002; Fossheim, 2014). Hagendorff emphasized the role of individual actors. Floridi (2016) also stressed on individual responsibilities. He insisted that everyone whose actions are causally relevant in leading to some consequences of collective action must be held morally accountable for the outcomes.

Hagendorff and Floridi relied on teaching individuals to act morally. Currently, the thinking that engineers and AI designers adopt during their education does not focus on ethical issues. Different fields of expertise have established their own approaches to handle the phenomena they encounter. The respective approaches have developed gradually, in relation to the specific tasks allotted to educated experts in society. Any professional training takes years to complete; after which people do not reach seniority before decades of practice. A list of ethical principles has a meager chance to affect an expert's mind in the rapidly changing field of AI development, where designers often work under considerable pressure (Mittelstadt, 2019, 108–109; Hagendorff, 2020, 502–503).

Instead of recommending ethics training, Coeckelbergh (2020, 145–165) presented various methods to improve the ethical design and development of AI. He suggests that each discussion of solving a case should make, as a starting point, an effort to answer six questions. Before fixing the course of action, there should be a clear understanding of why and when measures are needed, on what level the intervention should be made, and who should take part in the required action. The answers acquired through this method will then provide background knowledge for determining the nature, extent, and urgency of the problem that the agents are handling.[4]

Abbas et al. (2019, 76–78) suggested an ethics code corresponding the Hippocratic Oath for technologists as a solution. The oath would bind professionals to deliberate on ethical consequences, truthfulness, and responsibility. However, the suggested code does not advice practitioners on how to manage the time-consuming application of the listed ideals in the competitive and hectic corporate environment wherein most AI technologists work.

Hagendorff (2020, 108–109) identified that the companies responsible for the development of AI follow the instrumental logic of economic enterprises. It is not easy to combine the goals of profit making and efficiency with ethical thinking, irrespective of whether the recommended model prioritizes a value approach or grounds morality on normative principles. To overcome the gap between the principles and adopted practices, Morley et al. (2019) formulated suggestions and checklists to help AI developers consider the key ethical AI principles (Floridi et al., 2018) for the complete duration of design processes.

Existing organizational structures seldom encourage employees to consider moral concerns. AI as a promising branch for investors provides the aim of the ethically sustainable development of AI as a secondary role (Rosenberg, 2017). In this setting, there is a risk that moral considerations become ethical whitewashing (Vincent, 2019).

Suggestions to improve the ethical impact on AI assume that individual actors will be able to act morally by relying on a set of virtues, being held morally accountable, or applying a list of deliberative principles. Such measures may improve the ethical quality of AI development; however, if they fail, it may partly be because they all consider AI ethics as applying a theoretical or generalizable notion of ethics to practice, thereby solving the moral problem at hand. There is another field of applied ethics, known as bioethics, within which such an understanding of the nature of applied ethics has been criticized. To determine whether the criticism is also relevant in terms of AI ethics, it would be beneficial to consider the development of bioethics as a branch of applied ethics as well as the arguments of bioethicists concerning the nature and methodology of applied ethics.

BIOETHICS AS APPLIED ETHICS

The search for ethical principles to guide the development of AI technologies resembles the boom in applied ethics during the final decades of the 20th century. Rapid scientific and technological development created a need for a new type of ethical thinking. A fresh field of study, known as bioethics, emerged; in its wake, several other strands of human activity received ethical attention. During the past decades, bioethics as well as business ethics, sports ethics, and the ethics of professionalism have been established as branches of applied ethics within practical philosophy (Dittmer, 1995). As bioethics is the best established and most influential of the fields of applied ethics and offers the most interesting example of the evolution of applied ethics, its role, and its methodological discussions, it is worth studying the development of bioethics to determine whether it could offer some assistance to AI ethics.

A Short History of Bioethics

Bioethics emerged owing to the concerns of several scientists from various fields of research, who recognized the threats posed by various scientific and technological innovations to people and societies. A pioneer of the field, Van Rensselaer Potter, expressed in 1970 the need for a novel approach, called bioethics, as a new conjunction of scientific knowledge and a moral appreciation of the converging evolutionary understanding of human nature (Jonsen, 2012, 3).

Since Potter's use of the term, it has become customary to conceptualize bioethics from a narrower perspective. It is typically known as the ethical analysis of a range of moral questions posed to medical practices by advances in biomedical sciences and technologies (Jonsen, 2012, 3). Animal and environmental ethics complement the perspective of biomedical ethics, and together they cover the issues of the wider conception of bioethics (Gordon, 1995).

The urgency for an innovative ethical approach was visible in the scholars' aim to make bioethics an established form of academic discourse. The planning of the Encyclopedia of

[4]For more examples of the suggestions how to remedy the problem, refer to Mittelstadt (2019, 504).

Bioethics began in 1972, and 2 years later Dan Callahan published an article titled "Bioethics as a Discipline." He aimed to model an academic field that would combine traditional philosophical analysis with sensitivity toward human emotions and the ongoing social and political influences affecting the practice of medicine. The new discipline was proposed to serve people who encounter crucial decisions arising within medicine (Jonsen, 2012, 3).

Owing to the need for a new ethical approach, several developments took place after World War II. The accumulation of biological knowledge and rapidly emerging technological innovations made it possible to treat ailments and conditions that had earlier been incurable. For example, the invention of the artificial kidney and ventilator in the 1960s helped prolong the lives of numerous patients who would have otherwise died (Jonsen, 2012, 4).

Improving medication and refined surgical techniques, as well as various life-prolonging technological innovations, offered better prospects for numerous people; however, they also created several unprecedented problems. Artificial life support does not always help patients heal, regain consciousness, and restore their ability to move and communicate; instead, it may leave them in a permanent vegetative state. Such cases made it necessary to change the definition of death (Jonsen, 2012, 6; Overview: Brain Death, 2019; Redefining Death, 2019).

During the 20th century, research in molecular biology changed the understanding of organic life (Judson, 1996). Discoveries in genetics and cellular metabolism, mapping the human genome, and discovering stem cells with refined biological techniques have paved the way for genetic diagnostics, screening, and gene therapy as standard practices in medicine. New methods of controlling and manipulating human biology have made it possible to postpone death and manage fertility, gestation, and birth in an unprecedented manner (Jonsen, 2012, 7–9).

Along with the range of new options based on biological and technical knowledge, another remarkable change took place during the latter part of the 20th century. Economic growth, improving living standards, and widening opportunities for education, while considering the political ideals of democracy and equality, made the welfare of a society a general societal goal. There is now a growing demand for improved access to professional medical care. The traditional doctor–patient relationship has become insufficient for resolving the emerging problems of medical treatment and care. Awareness of the crimes against humanity committed by medical doctors of the Third Reich accentuated the need to protect the rights of those who take part in research with people's health as its objective. Consequently, requiring informed consent became a prerequisite for ethically sound medical research. Medically justified paternalism had to give way to patient autonomy as the basis of the doctor–patient relationship (Jonsen, 2012, 5–12).

The pioneers of bioethics were confident in their ability to safeguard moral concerns in synchrony with advancing technology and changing social conditions. They also envisaged a future for bioethics as an academic discipline. Should AI ethics advance along a similar path and establish itself as a partner in AI technology development and an autonomous field of study? To better understand how bioethics could—or should not —serve as a model for AI ethics as applied ethics, we will first examine, some of the criticism raised against the forms that bioethics has adopted and has been given in institutional settings, and second, the discussion concerning the relationship between theory and practice in bioethics.

Critical Viewpoints of Institutionalized Bioethics

The list of current master's programs in bioethics (List of Masters Programs in Bioethics, 2021), compiled by the Wikipedia collaborators, consists of approximately 100 academic programs provided by universities around the world. Bioethics has become an established academic discipline and the number of programs in higher education suggests that there is a need for professionals in this field. This situation may indicate that the hopes of the founding figures of bioethics have materialized, but there are also those whose views of the situation are more sinister. Bioethics in its different forms has been a target of diverse types of criticism over the past 30 years. Acquainting oneself with the arguments and viewpoints formulated in the discussion highlights the doubts concerning the agenda and the role bioethics has acquired as applied ethics.

The starting point of bioethics is in philosophy, which has provided methodological tools for the discipline. From the beginning, there has been a strong emphasis on bioethics as a critical discipline based on critical thinking in analytic philosophy as an evaluation concepts, positions, and arguments (Árnason, 2015). The central concepts of bioethics stem from the context in which the basic problems of the field were first formulated, that is, from the techno-medical and socio-cultural circumstances in the US. Hence, a highly individualistic conception of autonomy has taken priority over other moral concerns, leading to an emphasis on individual rights, choices, and welfare. Other traditionally important moral concepts, such as responsibility, obligation, and duty have been sidelined from discussions. Topics such as interpersonal relationships, human dependency and interdependency, community values, public health, social solidarity, trust, and the common good have received less attention (Fox and Swazey, 2005). From a more general perspective, the critical analysis and concern in current bioethics have not covered the effects of existing power relations on health care, making the bioethical enterprise inherently socially conservative. Instead of remaining detached from the development that it is supposed to critically analyze, bioethics runs the risk of joining the ranks of those who take part in the discussion as promoters of technological innovations (Ashcroft, 2004; Elliott, 2005; Árnason, 2015).

Another strand of criticism relates to the status bioethics has gained not only within healthcare institutions, but also in many other organizations. The offer of master's programs in bioethics indicates that there is an academic labor market for bioethicists. Bioethicists are not involved in treating, helping, assisting, or supporting patients, but work as part of organizational systems. Elliot calls the development bureaucratization of bioethics, during which bioethics has become a self-contained,

semiprofessional entity of bureaucratic structures. The fact that bioethicists have a position in the system lends them social authority. Owing to their status, the advice of the bioethicist then gains, possible undue, importance and relevance over other opinions (Elliott, 2005).

Bureaucratized bioethics has also found its way to institutions responsible for health policy both at the national and international levels (Littoz-Monnet, 2021). A global network of National Bioethics Committees (Global Summit of National Bioethics Committees, 2021), as well as of centers for bioethics (Global Network of WHO Collaborating Centers for Bioethics, 2021) has been established under the auspices of the World Health Organization. Bioethicists have been hired by pharmaceutical and biotechnology companies as well as by for-profit non-institutional review boards. Although the various institutional positions differ from each other, they all lend authority to the bioethicist, which is distinct from other types of authority in institutional bureaucracies. Elliot finds the development alarming, because the bureaucratic authority of bioethics is often not based on either academic or clinical merits but simply on memberships in different bioethical advisory boards or committees (Elliott, 2005).

Instead of applauding the institutional success of bioethics, critics point out that bioethics may risk becoming unable to accomplish its main function: the critical assessment of advances in biomedicine. Working as an employee within an institution involves adopting the duties, allegiance, and professional identities required by the employer thereby serving the interests of the institution. Such loyalties are likely to shape the conceptual agenda of bioethics as a field of study. Instead of formulating critical viewpoints, bioethicists may limit themselves to suggesting ways to improve the system rather than demanding a fundamental reform backed by ethical concerns. Bioethics loses its critical potential, as its practitioners become biased between the conflicting interests of different parties (Ashcroft, 2004; Elliott, 2005).

The critical remarks concerning the institutionalization and bureaucratization of bioethics provide a useful reference point for the future development of AI ethics. A brief overview of bioethics shows that institutional success does not necessarily support the basic task of the field. Next, we highlight a critical discussion within bioethics concerning its theoretical basis.

Bioethics and Discussion Concerning the Nature of Applied Ethics

The term "bioethics" has many uses, highlighting the relationship between theory and practice. First, it is the name of a disciplinary framework for various moral topics in relation to life sciences, human beings, animals, and nature. Second, it is an interdisciplinary approach that integrates various types of empirical data to solve practical problems. As an approach, bioethics claims to offer ethical guidance for practical problems and conceptual clarification of new types of complex issues. Additionally, its aim is to elaborate structured arguments by critically examining judgments and considerations in topical discussions. Bioethics employs moral philosophy when issuing problems arising from the biological nature of human beings. However, it can also contribute to the opposite, as the study of biological facts may give rise to specifications of ethical concepts, such as defining and understanding the notion of personhood (Gordon, 1995; Jonsen, 2012, 11–13).

A central feature of bioethics has been the aim of solving real-life problems and forming guiding practices and policies. The ideal that has motivated the development of bioethics is to create a practical, applicable moral philosophy, and to not concentrate on speculative analysis. The adopted line of study presents bioethics as a form of discourse that promotes public debate on issues related to biomedicine, thereby encouraging people to find ways to resolve upcoming issues (Jonsen, 2012, 12–13).

Such general descriptions of bioethics suggest that bioethics, even when emphasizing the importance of applicability, provides a theoretical basis for deducing practical solutions. If not, at least some principles are provided that people exercising bioethics can then apply to practical cases, thereby solving (or suggesting how to solve) the problems at hand. This view is intuitively appealing, but has been contested (Flynn, 2021).

The pioneers of bioethics were philosophers and theologians who represented different traditions of moral philosophy, such as Kantian deontology, varieties of utilitarianism, or Thomistic thinking. Thereafter, virtue, feminist, and narrative ethics were added to the pool of bioethical approaches. The differences between the basic theoretical assumptions suggest that the practical solutions deduced from them would also differ, thus reflecting the variety of background theories. However, this was not the case. In bioethics, different theoretical assumptions have not led to different suggestions concerning practical solutions to concrete problems, and sharing a background theory does not necessarily lead similar recommendations. Such observations suggest that the view of bioethics as applied ethics—deducing practical outcomes from theoretical principles—does not correspond with the actual role and practice of bioethics (Gordon, 1995; Flynn, 2021, 503; see also Mittelstadt (2019) who does not question the deductive view of bioethics).

The debate concerning the nature of bioethics as applied ethics has taken place among both practice-minded bioethicists and those who take a more theoretical approach to the discipline. Remarks concerning the discrepancy between the differences of opinion on the theoretical and practical levels of bioethics are the weightiest theoretical arguments against the deductive view of bioethics. The bioethical approach of considering bioethics as a theory and practice, which are linked by applying the basic values and principles offered by the theory to the practical problem at hand, is known as principlism (Gordon, 1995; Jonsen, 2012; Flynn, 2021).

The critical voices against the principlistic approach gained impetus in the 1980s. The critics noted that the approach often generated more theory, instead of accounting for real-life issues and considering their acuteness in people's lives. According to their view, abstract theory should give way to each actual case as the starting point of a bioethicist's consideration.

The suggestion, referred to as casuistry, was that top-down principlistic approaches should be replaced by a bottom-up type of reasoning and problem solving (Clouser and Kopelman, 1990; Gordon, 1995; Jonsen, 2012; Flynn, 2021).

Casuistry did not replace principlism as the main approach in bioethics; however, it did demonstrate the need to modify deductive approaches. The reason why casuistry did not gain more support was the criticism that a mere case description does not help in solving practical problems. To make a normative decision, at least one normative premise is required in the form of values and principles. The solution suggested was that, in lieu of abstract principles representing high morality, the theoretical starting point of a theory of ethics, bioethical considerations should focus on mid-level principles (Gordon, 1995; Flynn, 2021). The most well-known suggestions of such mid-level norms are the four principles of autonomy, non-maleficence, beneficence, and justice, which was first defined in 1984 and repeatedly redefined by Beauchamp and Childress (2019).

The gap between principlism and casuistry has been further bridged by methodological considerations borrowed from the discussion on how to best define the basic principles of societal justice. Based on John Rawls's concept of reflective equilibrium (Rawls, 1971), bioethicists have developed methods to balance each other's theoretical notions, moral principles, cultural and social conceptions, and facts concerning acute practical problems (Flynn, 2021).

By critically considering all aspects of a case against each other, it is possible to reach a conclusion that can serve as a suggestion for managing the problem. During the deliberative process, each of the discursive elements and discussion parties affects the other elements. Consequently, the empirical observations and considerations based on them may affect theoretical conceptions and modify basic moral principles, and vice versa (Flynn, 2021).

What could the discussion about bioethics contribute to the discussion of AI ethics? The subject matter of AI ethics is in many respects different from that of bioethics, as the applications and systems using AI do not relate to any overarching topic, unlike the focus on health and well-being in bioethics. However, the bureaucratization, and the conceptual and methodological developments in bioethics over the past decades warrant further examination.

The concept of applied ethics implicit in AI ethical models follows the general pattern of how bioethics has been understood as an application of ethical theory to moral practice (Mittelstadt, 2019, 501) and there are signs of bureaucratization of AI ethics (Rességuier and Rodrigues, 2020). AI ethics has followed the deductive view of applied ethics, which has not been successful in realizing the desired change thus far. Is there something that could complement the deductive and principlistic approach, and what could the resources for doing that be?

Bioethics began with the aim of establishing itself as a novel form of ethical thought that would form a discipline. Is this a path that AI ethics should try to follow, promoting professorships in AI ethics in universities? The bioethical endeavor has been successful in making bioethical considerations a part of the standard procedures of medical research. Moreover, it is not possible to conduct research without, at least nominally, pre-examining one's project from an ethical perspective. Would establishing AI ethics committees improve the ethical sustainability of AI?

In simple terms, ethical reasoning can be implemented through three channels: improving the moral quality of human agents, establishing a set of regulations and control systems to discipline their application, or establishing that moral decency is beneficial. AI ethical models rely primarily on the first two techniques. We shall now discuss, whether it is possible to formulate the third technique, connecting moral considerations to other features of AI design and development—and suggest how it could be done.

DETECTING THE MORALLY RELEVANT FEATURES OF AI DESIGN AND DEVELOPMENT

The development of bioethics shows that there is a link between the concept of bioethics as applied ethics and the adopted methodological approaches. In bioethics, by enriching the methodological apparatus, the deductive and principlistic perspectives were modified with a view that connected it more closely to the reality of actual problems. By adopting discursive methods that aim to reach a reflective equilibrium between the features of reality, which are ethically relevant at the time and in the context of decision making, bioethics has renewed both the conception of applied ethics and its methodological resources.

Following the example of research in bioethics, developing AI ethics as applied ethics presupposes two things. First, we must focus on the specific features of AI that are significant in attempts to improve the ethical sustainability of AI design and development. The second presupposition is to identify methodological tools that could help account for the morally relevant features of AI.

It is characteristic of AI that its products and outcomes are not just devices, but programs and applications within larger, often extremely complex systems, such as the health and welfare data generating system (Apotti, 2021) and taxation system (My Tax, 2021). It is impossible to extract the AI from the rest of the system. Most of the development of these applications is carried out commercially within the market economy. The use of AI is significantly more widespread than any other research topic handled by other fields of applied ethics. Unlike biomedical ethics, AI ethics is not tied to institutions, established professions, or educational traditions. Moreover, no culturally determined roles exist to support those who work in AI design that are comparable to the role positions in healthcare settings. There is no general conception of an agency that could be tied to a person who utilizes AI (Mittelstadt, 2019, 501–502). In most cases, the users are fully competent agents. They are citizens, in both professional and private settings, who use AI to accomplish their goals. AI is applicable to almost any human activity, which provides innumerable possibilities for its use. Such features of AI have moral relevance, and should receive due attention in the formulations of AI ethics (Boddington, 2017, 93).

The methodological developments toward a discursive approach in bioethics indicate that the effects between the theoretical, conceptual, factual, and practical levels are more complex than those proposed by the deductive and principlistic models of applied ethics. Recently, bioethicists have suggested that bioethics should adopt a systemic approach to the subject matter (Stoeklé et al., 2019, 24–25).

For AI ethics, the systemic approach is more urgent than it is in bioethics, as it is typical for AI applications to become part of various socio-technical systems, wherein humans and technological applications, laws, social norms, and non-intelligent infrastructures interplay in a web of actions and agents (Dignum, 2020, 216). Conversely, Powers and Ganascia (2020, 48–49) identified that the EU ethics guidelines' recommendations for a human-centered approach imply that the opposite, that is, a machine-centered approach, could be possible.

According to the socio-technical systems (STS) theory a system comprises interconnected elements that may form subsystems.[5] The system is always more than the set of its elements, as the nature of the system is determined by the set of relations between the elements and subsystems within the system. It forms an environment in which the subsystems exist, and has a structure that is linked to its function; Any function may be produced by various types of structures—a feature that the STS expresses as the principle of equifunctionality. The system cannot be captured completely through any single level of its structure, thus expressing the principle of excluded reductionism (Ropohl, 1999, 188–192).

A system becomes socio-technical when information technology is used to mediate between cognitive and social interactions. This means that socio-technical systems combine the technical and nontechnical elements of a system, such as people, regulations, processes, and cultural aspects. Non-technical elements are an integral part of a system and its workings. Owing to the constantly varying and vague nature of the non-technical components, it is difficult to design, coordinate, and run socio-technical systems in an exact manner (Appelbaum, 1997, 453; Mariani, 2016, 157).

The workings and functioning of a socio-technical system do not follow a straightforward causality. Instead, such systems have emergent properties that cannot be attributed to, or derived from, the individual parts. The models, based on the link between cause and effect, are not sufficient to account for the relations and functioning of the system and its parts. The effects of the work and its functioning evolve from the relationships and dependencies that prevail among the components of the system. Socio-technical systems function non-deterministically; that is, it is not possible to determine the outcome of the system at a given time, as it may change depending on the situation (Mariani, 2016, 158), which makes it difficult to regulate and direct the outcomes of a complex socio-technical approach through the application of rules and principles.

Well-functioning socio-technical systems are resilient in that they can focus on their primary task, even in rapidly changing circumstances. The subsystems and people working in them constantly adjust their actions according to the requirements of the challenges they encounter. A resilient system reacts to emerging situations, and knows how to anticipate changes (Dekker, 2014, 140, 200–201). Considering that the primary task is ethically sound and that performing it does not violate moral norms, remaining resilient and adhering to the primary task amid change is a morally acceptable goal. As AI applications are socio-technical systems, one of the central tasks of AI ethicists is to determine ways to integrate ethical sustainability with the notion of systemic resilience.

What could serve as an approach that accounts for the nature of AI as an active element of complex socio-technical systems and enrich the methodological tools of a more accurate AI ethics? We conclude our article by suggesting ways to strengthen AI ethics as applied ethics by enriching their methodological and practice-oriented approaches.

SECURING SAFETY AND ASSESSING CHANGE AS ETHICALLY RELEVANT ENDEAVORS OF AI ETHICS

Safety research is a field of rigorous academic study that has had a close connection to the empirical reality of hazards and risks, since its emergence approximately a century ago.[6] The research has concentrated on both preventing harmful events from occurring, in the form of inhibiting accidents, and on making processes and actions more reliable, in the form of improving safety. In both approaches, the concept of risk plays a significant role (Dekker, 2014; Reason, 2016).

Preventing detrimental chains of events is not possible without formulating a concept for the types of happenings to be undermined. Reiman and Oedewald (2008; see also Hollnagel, 2014, 39–47) have detected four main phases wherein the nature of accidents has been conceptualized in the study of accidents. First, faults in mechanical machinery are often identified as the cause of mishaps. These observations have led to improved technology, thus making the functions of devices and mechanical systems more precise and regular.

As technology became more reliable after WW II, the role of the human agent and the interaction between machinery and human beings gained more attention. Devastating accidents, especially at nuclear power plants and chemical factories during the 1980s, necessitated a shift in focus, and the part played by organizational factors received attention (Reiman and Oedewald, 2008, 35). During the past decades, a systems approach has been introduced and adapted as a central tool in both safety research and its practical applications. Additionally, the focus has shifted from accidents to characteristics that maintain safety and enable complex organizations to continue their functions and perform primary tasks, even in rapidly changing circumstances, functioning smoothly through occasional crises and even when they encounter a catastrophe (Hollnagel, 2014).

[5]For the historical development of the theory, refer to Bednar and Welch (2016).

[6]Several academic journals focus on understanding and improving safety, e.g., *Journal of Safety Research, Safety Science*, and *Accident Prevention and Safety*.

Along the way, safety researchers have worked in close contact with representatives of industry, companies, workers, and state officials. The practice-oriented mindset has enabled them to develop several tools for analyzing and preventing accidents, as well as methods of learning from them.[7] Analyses of accidents and safety studies do not usually discuss ethics. The focus is on finding practical solutions to real-life problems by attempting to understand the factors that cause errors, accidents, and damage. This is an ethically sound goal and shows that important moral concerns can be integrated into distinct types of activities without the agents naming their approach ethics.

Many of the hazards and risks associated with developing and using AI resemble issues that are relevant to safety research. Safety studies began with linear accident models wherein a leading role was given to the individual agent, which are presuppositions that many subsequent ethical codes have adopted (Hollnagel, 2014). There are already suggestions on how to widen the view to cover the organizational level and analyze AI systems as parts of the surrounding societies (Tschopp, 2019). The next step is to adopt a systemic approach, following the example of safety studies. The systems approach provides an overall view, opening possibilities for handling significant global problems, creating more acute consumer awareness, and determining ways to generate more accurate regulations. Another point of reference is the concept of risk central in safety studies and emphasized by the European Union AI Act. The AI Act, however, does not offer a systemic approach to safety issues; rather, it concentrates on the regulation of high-risk AI systems and the mitigation of the risks they entail in relation to other legal and ethical principles stated for instance in the EU Charter of Fundamental Rights, in a deductive manner (AI Act, 2021).

The aim of AI design and development is to change reality by creating a desired effect on the targeted issues. One such example is IEEE's initiative "Ethically aligned design." It presents a vision for prioritizing human wellbeing with autonomous and intelligent systems, which provides tools to measure and consider the influence that autonomous intelligent systems have on human wellbeing. Recent studies on what these changes and effects are, and how to assess them, could offer sources to improvement in the resources of AI ethics. The theory of change and the impact management project have been developed to help organizations and companies to better keep track of the changes they intend to bring about and the actual effects of their work in practice. Both approaches are designed to correspond with the complexities of the current world, where it is difficult or impossible to design and realize straightforward strategies. In favorable circumstances with the right measures, the desired change can be achieved (Theory of Change, 2021).

Impact management is a narrower and more practice-oriented approach than the theory of change. Both have the same starting point in that it is impossible to predict the outcomes of organizational activities and that their actual effects deviate from the formulation of previously set strategic goals. Even well-planned actions can cause unintended effects, some of which can be detrimental and cause harm to uninvolved parties. The effects can be ecological, social, political, and cultural, unintentionally changing the lives of both human communities and ecosystems (The Impact Management Project, 2021; Theory of Change, 2021).

Impact management concentrates on creating permanent practices for measuring, assessing, and improving the influences of factors relevant to planetary sustainability. It addresses enterprises and investors who wish to act responsibly in relation to environmental, social, and governance risks, and those who wish to contribute globally to sustainable development goals (SDGs) (The Impact Management Project, 2021).

SDGs are expressed as a set of targets and indicators against which businesses and investors can differentiate and communicate their roles, goals, and performance. The role of impact management is to support actors in better understanding and improving their performance in relation to SDGs. As part of the approach, it evaluates both the positive and negative aspects as well as the expected and real outcomes. Moreover, identifying whom the outcomes concern and which factors and agents realized these outcomes are part of the impact evaluation (The Impact Management Project, 2021).

A similar procedure was recently adopted in the context of an AI research conference, where the authors had to include a statement in their submitted paper that addressed the broader ethical and future societal impacts of their research as a precondition of approval (Prunkl et al., 2021). Prunkl et al. pointed out that the implementation of this new practice was not ideal because the task was complex and difficult, the authors lacked guidance on how to evaluate the possible effects of their research, and there may have been pressure to stress the positive impacts. Furthermore, the outcomes of these impact reports are likely to be cognitively biased and of inferior quality. However, the ethical consideration required in the process of reporting anticipated impacts of AI research may be beneficial to the internalization of ethical thinking as a continuing practice (Rességuier and Rodrigues, 2020).

The theory of change is a more theoretical approach, which aims to provide "a comprehensive description and illustration of how and why a desired change is expected to happen in a particular context" (Theory of Change, 2021). Starting from the outcomes of a program, it works backward toward the desired long-term goals, which were the starting point for the program's plan of action. The task is to fill in the gap between the initial goals and outcomes by determining the actual impact of the action plan. The analysis helps to identify the factors that must be in place for the intended effects to occur (Theory of Change, 2021).

One of the central methods of applying the theory of change is to analyze the underlying assumptions between the goal, chosen course of action, and actual outcomes. This helps the involved parties to specify the necessary components for the outcomes to conform to the plans, and to understand the nature of change, which the planned action should execute. Consequently,

[7]For the different approaches in safety research, refer to the normal accident theory approach (Perrow, 1999), the efficiency-thoroughness trade-off principle (Hollnagel, 2009), the high reliability theory approach (Weick, 1987; Roberts, 1989), and (Hollnagel, 2014; Vanderhaegen and Hollnagel, 2015; Resilience Engineering, 2019). For an overview of the conceptual development of safety research and its potential for AI ethics, refer to Hallamaa (2021).

minor changes that often go unnoticed also receive attention, thereby helping the actors discern factors contributing to desired effects or inhibiting them (Weiss, 1995; Theory of Change, 2021).

The term "theory of change" was coined by Carol Weiss in 1995, who challenged the designers of complex community-based initiatives to articulate their assumptions on how they expected their work to effect change. Her idea was that explicating the implicit presuppositions would enable actors to discern the outcomes of their work and to claim credit for them. Weiss' ideas were initially applied to philanthropic organizations, government agencies, and international non-governmental organizations; however, over the years, they have been linked with systems thinking and complexity (Weiss, 1995; Theory of Change, 2021).

HOW TO IMPROVE AI ETHICS

Safety research, the theory of change, and the impact management approach all contain elements that can be used to improve AI ethics. AI ethics could also become a more relevant point of discussion for the industry if it adopted the reality-based practice orientation used in safety studies and the theory of change. Instead of formulations of high morality, it concentrates on the fact that mistakes happen, people err, and accidents occur, using these as motivators to improve machine-assisted human action. Instead of formulating ethical conflicts, AI ethics could be a way to map an effective route between planned strategic goals and actual outcomes.

Many of the ethical problems in designing and using AI could be formulated as questions concerning safety and risks, making an impact, and initiating change. By concentrating on achieving favorable outcomes and avoiding unwanted effects, the discussion is less theoretical; thus, encouraging the discovery of applicable answers to detected problems.

The theory of change and impact management aim to make effective use of resources. Safety studies have avoided an unfruitful conflict between financial values and other values by translating the costs of neglect in safety to tangible economic losses. Before long, failing to responsibly address issues of safety will cost more than investing in it. Safety studies could provide methods to make a similar correlation between commercial interests and an ethically sound AI design.

Safety studies make constant use of technology studies, and the experience of technology experts has had an invaluable effect on discussions concerning safety. Similarly, a more practice-oriented approach could make AI ethics more easily approachable for the designers and developers of AI than trying to educate everyone in ethical discourse. Instead of concentrating on lists of principles and formulating high moral values, AI ethics should focus on the change and influence that AI has on the world. Doing so could make AI ethics less philosophical and bring them closer to practice-oriented approaches. By concentrating on the methods that cause actual changes, AI ethics could improve the design and development of AI, thereby contributing to practical morality.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

The order of authorship reflects the extent of contributions. Both authors contributed to manuscript revision and read and approved the submitted version.

FUNDING

The authors wish to acknowledge the project Ethical AI for the Governance of the Society (ETAIROS, grant #327352), funded by the Strategic Research Council at the Academy of Finland.

ACKNOWLEDGMENTS

We would like to thank Editage (www.editage.com) for English language editing.

REFERENCES

Abbas, A., Senges, M., and Howard, R. (2019). "A hippocratic oath for technologists," in *Next-Generation Ethics: Engineering a Better Society*, ed. A. Abbas (Cambridge: Cambridge University Press), 71–80. doi: 10.1017/9781108616188.006

ACM (2018). *ACM Code of Ethics and Professional Conduct: Affirming Our Obligation to Use Our Skills to Benefit Society*. Available online at: https://www.acm.org/binaries/content/assets/about/acm-code-of-ethics-booklet.pdf (accessed June 9, 2021).

AI Act (2021). *Proposal for A Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. COM/2021/206 Final*. Available online at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (accessed December 22, 2021).

AI HLEG (2019). *High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy Artificial Intelligence*. Brussels: European Commission. Available online at: https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf

Apotti (2021). Available online at: https://www.apotti.fi/en/ (accessed June 8, 2021).

Appelbaum, S. (1997). Socio-technical systems theory: an intervention strategy for organizational development. *Manage. Decision* 35, 452–463. doi: 10.1108/00251749710173823

Árnason, V. (2015). Toward critical bioethics. *Cambridge Q. Healthc. Ethics* 24, 154–164. doi: 10.1017/S0963180114000462

Ashcroft, R. E. (2004). Bioethics and conflicts of interest. Studies in history and philosophy of science. *Part C Stud. Hist. Philos. Biol. Biomed. Sci.* 35, 155–165. doi: 10.1016/j.shpsc.2003.12.011

Beauchamp, T., and Childress, J. F. (2019). *Principles of Biomedical Ethics, 8th Edn*. Oxford: Oxford University Press.

Bednar, P., and Welch, C. (2016). "Enid mumford: the ethics methodology and its legacy," in *Co-Creating Humane and Innovative Organizations: Evolutions in the Practice of Socio-technical System Design*, eds B. J. Mohr and P. van Amelsvoort (Leuven: Global STS-D Network), 274–288.

Boddington, P. (2017). *Towards a Code of Ethics for Artificial Intelligence. Artificial Intelligence: Foundations, Theory, and Algorithms*. Cham: Springer International Publishing AG.

Canca, C. (2020). Operationalizing AI ethics principles. *Commun. ACM* 63, 18–21. doi: 10.1145/3430368

Clouser, K. D., and Kopelman, L. M. (1990). Philosophical critique of bioethics: introduction to the issue. *J. Med. Philos.* 15, 121–124. doi: 10.1093/jmp/15.2.121

Coeckelbergh, M. (2020). *AI Ethics. The MIT Press Essential Knowledge Series*. Cambridge, MA: The MIT Press.

Dekker, S. (2014). *The Field Guide to Understanding 'Human Error,' 3rd Edn*. Boca Raton, FL: Taylor & Francis.

Dignum, V. (2020). "Responsibility and artificial intelligence," in *Oxford handbook of Ethics of AI*, eds M. D. Dubber, F. Pasquale, and S. Das (New York, NY: Oxford University Press), 215–231.

Dittmer, J. (1995). *Applied Ethics. Internet Encyclopedia of Philosophy*. Available online at: https://iep.utm.edu/ap-ethic/ (accessed June 2, 2021).

Doris, J. (2002). *Lack of Character: Personality and Moral Behavior*. Cambridge: Cambridge University Press.

Elliott, C. (2005). The soul of a new machine: bioethicists in the bureaucracy. *Cambridge Q. Healthc. Ethics* 14, 379–384. doi: 10.1017/S0963180105050528

Floridi, L. (2016). Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions. *Philos. Trans. Series A Math. Phys. Eng. Sci.* 374, 1–13. doi: 10.1098/rsta.2016.0112

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4People - An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach.* 28, 689–707. doi: 10.1007/s11023-018-9482-5

Flynn, J. (2021). "Theory and bioethics," in *The Stanford Encyclopedia of Philosophy (Spring 2021 Edition)*, ed E. N. Zalta. Available online at: https://plato.stanford.edu/archives/spr2021/entries/theory-bioethics/

Fossheim, H. (2014). Virtue ethics and everyday strategies. *Revue Internationale Philos.* 267, 65–82. doi: 10.3917/rip.267.0065

Fox, R. C., and Swazey, J. P. (2005). Examining American bioethics: its problems and prospects. *Cambridge Q. Healthc. Ethics* 14, 361–373. doi: 10.1017/S0963180105050504

Global Network of WHO Collaborating Centers for Bioethics (2021). Available online at: https://www.who.int/groups/global-network-of-who-collaborating-centres-for-bioethics/about (accessed December 10, 2021).

Global Summit of National Bioethics Committees (2021). Available online at: https://www.who.int/groups/global-summit-of-national-bioethics-committees/about (accessed December 10, 2021).

Gordon, J.-S. (1995). *Bioethics. Internet Encyclopedia of Philosophy*. Available online at: https://iep.utm.edu/bioethic/ (accessed May 20, 2021).

Hagendorff, T. (2020). The ethics of AI ethics: an evaluation of guidelines. *Minds Mach.* 30, 99–120. doi: 10.1007/s11023-020-09517-8

Hallamaa, J. (1994). *The Prisms of Moral Personhood. The Concept of a Person in Contemporary Anglo-American Ethics*. Schriften der Luther-Agricola Gesellschaft 33. Helsinki: Luther-Agricola-Society.

Hallamaa, J. (2021). "What could safety research contribute to technology design?" in *Culture and Computing. Design Thinking and Cultural Computing. HCII 2021. Lecture Notes in Computer Science*, Vol. 12795, ed M. Rauterberg (Cham: Springer), 56–79.

Harman, G. (1998–1999). Moral philosophy meets social psychology: virtue ethics and the fundamental attribution error. *Proc. Aristotelian Soc.* 99, 315–356. doi: 10.1111/1467-9264.00062

Hollnagel, E. (2009). *The ETTO Principle: Efficiency-Thoroughness Trade-off: Why Things That Go Right Sometimes Go Wrong*. Farnham: Ashgate.

Hollnagel, E. (2014). *Safety-I and Safety-II: The Past and Future of Safety Management*. Boca Raton, FL: CRC Press.

Howard, R. (2019). "Ethical distinctions for building your ethical code," in *Next-Generation Ethics: Engineering a Better Societ*, ed. A. Abbas (Cambridge: Cambridge University Press), 7–16. doi: 10.1017/9781108616188.002

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2017). *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being With Autonomous and Intelligent Systems, Verion 2*. Available online at: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf. (accessed June 10, 2021).

Jobin, A., Ienca, M., and Vayena, E. (2019). Artificial intelligence: the global landscape of ethics guidelines. *Nat. Mach. Intell.* 1, 389–399. doi: 10.1038/s42256-019-0088-2

Jonsen, A. R. (2012). "A history of bioethics as discipline and discourse," in *Bioethics: An Introduction to the History, Methods, and Practice, 3rd Edn.*, eds J. Silbergeld, A. Nancy, A. R. Jonsen, and R. A. Pearlman (London: Jones & Bartlett Learning), 3–16.

Judson, H. F. (1996). *The Eighth Day of Creation: Makers of the Revolution in Biology*. Plainview, NY: Cold Spring Harbor Laboratory Press.

List of Masters Programs in Bioethics (2021). *Wikipedia. Edited Version of 15 October 2021*. Available online at: https://en.wikipedia.org/wiki/List_of_masters_programs_in_bioethics (accessed December 9, 2021).

Littoz-Monnet, A. (2021). Expanding without much ado: international bureaucratic expansion tactics in the case of bioethics. *J. Euro. Public Policy*. 2020, 858–879. doi: 10.1080/13501763.2020.1781231

Mariani, S. (2016). *Coordination of Complex Sociotechnical Systems. Artificial Intelligence: Foundations, Theory, and Algorithms*. Cham: Springer.

McNamara, A., and Smith, J., and Murphy-Hill, E. (2018). "Does ACM's code of ethics change ethical decision making in software development?," in *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering-ESEC/FSE 2018*, eds. G. T. Leavens, A. Garcia and C. S. Păsăreanu (New York, NY: ACM Press), 1–7.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Perspectives. *Nat. Mach. Intell.* 1, 501–507. doi: 10.1038/s42256-019-0114-4

Morley, J., Floridi, L., Kinsey, L., and Elhalal, A. (2019). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Sci. Eng. Ethics*. doi: 10.2139/ssrn.3830348. [Epub ahead of print].

My Tax (2021). Available online at: Available online at: https://www.vero.fi/en/e-file/mytax/ (accessed June 8, 2021).

National Artificial Intelligence Initiative (2020). *National Artificial Intelligence Initiative: Overseeing and Implementing the United States National AI Strategy*. Available online at: https://www.ai.gov/strategic-pillars/advancing-trustworthy-ai/#Research_and_Development_for_Trustworthy_AI (accessed June 9, 2021).

Overview: Brain Death (2019). Available online at: https://www.nhs.uk/conditions/brain-death/ (accessed May 21, 2021).

Perrow, C. (1999). *Normal Accidents: Living with High-risk Technologies*. Princeton, NJ: Princeton University Press.

Powers, T. M., and Ganascia, J.-G. (2020). "The ethics of ethics of AI" in *Oxford handbook of Ethics of AI*, eds. M. D. Dubber, F. Pasquale, and S. Das (New York, NY: Oxford University Press), 28–51.

Prunkl, C. E. A., Ashurst, C., Anderljung, M. (2021). Institutionalizing ethics in AI through broader impact requirements. *Nat. Mach. Intell.* 3, 104–110. doi: 10.1038/s42256-021-00298-y

Rawls, J. (1971). *A Theory of Justice: Original Edition*. London: Harvard University Press.

Reason, J. (2016). *Organizational Accidents Revisited*. Farnham: Ashgate.

Redefining Death (2019). Encyclopedia.com. Available online at: https://www.encyclopedia.com/caregiving/legal-and-political-magazines/redefining-death (accessed May 21, 2021).

Reiman, T., and Oedewald, P. (2008). *Turvallisuuskriittiset organisaatiot: Onnettomuudet, kulttuuri ja johtaminen*. Helsinki: Edita.

Renda, A. (2020). "Europe: toward a policy framework for trustworthy AI," in *Oxford Handbook of Ethics of AI*, eds M. D. Dubber, F. Pasquale, and S. Das (New York, NY: Oxford University Press), 359–374.

Resilience Engineering (2019). Available online at: https://www.resilience-engineering-association.org/blog/2019/11/09/what-is-resilience-engineering/ (accessed June 8, 2021).

Rességuier, A., and Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data Soc.* 7:205395172094254. doi: 10.1177/2053951720942541

Roberts, K. H. (1989). New challenges in organizational research: high reliability organizations. *Org. Environ.* 3, 111–125. doi: 10.1177/108602668900300202

Ropohl, G. (1999). Philosophy of socio-technical systems. *Soc. Philos. Technol. Q. Electron. J.* 4, 186–194. doi: 10.5840/techne19994311

Rosenberg, S. (2017). *Why AI Is Still Waiting for Its Ethics Transplant*. Available online at: https://www.wired.com/story/why-ai-is-still-waiting-for-its-ethics-transplant/ (accessed June 9, 2021).

Stoeklé, H.-C., Deleuze, J.-F., and Vogt, G. (2019). Society, law, morality and bioethics: a systemic point of view. *Ethics Med. Public Health* 10, 22–26. doi: 10.1016/j.jemep.2019.06.005

The Impact Management Project (2021). Available online at: https://impactmanagementproject.com/ (accessed May 28, 2021).

The Montréal Declaration for a Responsible Development of Artificial Intelligence (2018). Inven_T, University of Montreal's Technosocial Innovation Centre. Available online at: https://www.montrealdeclaration-responsibleai.com/ (accessed May 29, 2021).

Theory of Change (2021). Available online at: https://www.theoryofchange.org/what-is-theory-of-change/ (accessed June 1, 2021).

Tschopp, M. (2019). *On Trust in AI: A Systemic Approach*. Available online at: https://www.scip.ch/en/?labs.20180823. (accessed December 29, 2021).

Vanderhaegen, F., and Hollnagel, E. (2015). Safety-I and safety-II, the past and future of safety management. *Cogn. Technol. Work* 17, 461–464. doi: 10.1007/s10111-015-0345-z

Vincent, J. (April 3, 2019). The problem with AI Ethics. Is Big Tech's embrace of AI ethics boards actually helping anyone? *The Verge*. Available online at: https://www.theverge.com/2019/4/3/18293410/ai-artificial-intelligence-ethics-boards-charters-problem-big-tech (accessed June 2, 2021).

Weick, K. E. (1987). Organizational Culture as a Source of High Reliability. *Calif. Manage. Rev.* 2, 112–127. doi: 10.2307/41165243

Weiss, C. H. (1995). "Nothing as practical as good theory: exploring theory-based evaluation for comprehensive community initiatives for children and families," in *New Approaches to Evaluating Community Initiatives: Concepts, Methods and Contexts*, eds J. Connell, A. Kubisch, L. Schorr, and C. Weiss (Washington, DC: Aspen Institute), 65–92.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.