Patterns

CellPress
OPEN ACCESS

Review

A high-level overview of AI ethics

Emre Kazim[1,*] and Adriano Soares Koshiyama[1,*]
[1]Computer Science, University College London, Gower St, London WC1E 6EA, UK
*Correspondence: e.kazim@ucl.ac.uk (E.K.), adriano.koshiyama.15@ucl.ac.uk (A.S.K.)
https://doi.org/10.1016/j.patter.2021.100314

THE BIGGER PICTURE The development and deployment of AI in business and society is rapid and ubiquitous. This disruptive technology is revolutionizing entire industries and presents significant opportunity. However, triggered by high-profile cases of harm (e.g., *Facebook-Cambridge Analytica*), concern is rising. This has spurred stakeholders to respond (governments, NGOs, academia, industry) with ethics-focused research, principles, and frameworks. This literature has flourished, often referred to as "Trustworthy AI". This review embraces *inherent interdisciplinarity* in the field by providing a high-level introduction to AI ethics drawing upon philosophy, law, and computer science. Readers will explore key terms and central themes (from AI to translating ethics into engineering practice). This serves as a point of departure to the literature as well as future debates—including the interrelation of *data ethics and AI*; the *legal status of algorithms*, *economic and political* impacts; and *nature-centric AI*.

SUMMARY

Artificial intelligence (AI) ethics is a field that has emerged as a response to the growing concern regarding the impact of AI. It can be read as a nascent field and as a subset of the wider field of digital ethics, which addresses concerns raised by the development and deployment of new digital technologies, such as AI, big data analytics, and blockchain technologies. The principle aim of this article is to provide a high-level conceptual discussion of the field by way of introducing basic concepts and sketching approaches and central themes in AI ethics. The first part introduces concepts by noting what is being referred to by "AI" and "ethics", etc.; the second part explores some predecessors to AI ethics, namely engineering ethics, philosophy of technology, and science and technology studies; the third part discusses three current approaches to AI ethics namely, principles, processes, and ethical consciousness; and finally, the fourth part discusses central themes in translating ethics in to engineering practice. We conclude by summarizing and noting the inherent interdisciplinary future directions and debates in AI ethics.

INTRODUCTION

Artificial intelligence (AI) ethics is a field that has emerged as a response to the growing concern regarding the impact of AI. Indeed, there have been an increasing number of high-profile cases of harm that has resulted either because of the misuse of the technology (e.g., psychometric voter manipulation,[1,2] facial recognition surveillance, mass data collection without consent, etc.), or as a result of the technology having design flaws (e.g., bias in cases of recidivism,[3] loan rejection, and medical misdiagnosis, etc.).[4] We read AI ethics as a nascent field and as a subset of the wider field of digital ethics, which addresses concerns raised by the development and deployment of new digital technologies, such as AI, big data analytics, and blockchain technologies. There is a growing and evolving literature that has significantly increased in the past number of years (2017–) and this literature *continues to evolve* with reviews from a computer science perspective[5] and from the philosophical and humanities perspective.[6–8] The principle aim of this article is to provide a high-level overview in this evolving space that serves as an introduction to the field. We do this by introducing basic concepts and sketching approaches and central themes in AI ethics. We recognize that this is an attempt to contribute to the field by drawing on our expertise in the humanities and engineering, by including literature from both disciplines and attempting to present them in a manner that is legible by all parties. As such, the article is neither to be read as a piece of moral philosophy, nor pure engineering; rather, we hope that this article is read as an attempt to action calls for interdisciplinarity and to stimulate interdisciplinary thought in the field.

The overview is structured into four parts; the first part introduces basic concepts and terms, i.e., what is meant by key terms, such as "AI" and "ethics"; the second part explores some predecessors to AI ethics, namely engineering ethics, philosophy of technology and science and technology studies; the third part discusses three current approaches to AI ethics namely, principles, processes, and ethical consciousness; and finally, the fourth part discusses central themes in translating

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

ethics in to engineering practice. We conclude by summary via overview and by noting future directions and debates.

CONCEPTS AND TERMS

In this section we introduce concepts and key terms. As a point of departure, we begin by noting areas covered in "digital ethics" and then "AI ethics." This ordering follows our belief that AI ethics is a subdiscipline of the broader umbrella of digital ethics. Following this, we unpack these by exploring how the term "digital" is being used in digital ethics and then expand on how the term "AI" is being used (similar to how AI ethics falls under digital ethics, AI is shown to fall under "digital"). Furthermore, we explicate how the term "ethics" is being used in this context and expand upon the dominant ethical philosophies that we believe AI ethics may draw upon. This section closes with an exploration of "human centric AI," which we take to be the emerging overarching value framework of AI ethics.[9]

Before turning to the concepts and terms, it is important to make clear that the following is not to be read as a series of definitive definitions but rather as introductions and points of departure to think about these concepts that underpin AI ethics. Indeed, at a basic level, simply registering to data scientists and other practitioners and contributors to the field that there are foundational ideas from philosophy that AI ethics appeals to would be a positive contribution. As such, readers should engage critically with the below concepts and key terms that are introduced.

Digital and AI ethics

Drawing on a conception of ethics that encompasses broader social and political themes, we read digital ethics as covering the psychological, social (including environmental), and political impact of emerging digital technologies. The psychological refers to the likes of agency (moral self-determination), cognitive shifts, and selfhood; where the social refers to identity, belonging, and communities, as well as environmental issues; and, where the political refers to legal/jurisdictional, democratic (including accountability), and the economic realm. Furthermore, this can be thought of in terms of the scope of the impact, i.e., the psychological represents impact on the individual, the social the impact on the collective and environment, and the political the impact on the organizing structures of society. Thus, it is also clear that digital ethics is a highly interdisciplinary field, requiring expertise that spans the sciences (computer science) and the humanities (philosophy, law, sociology, psychology, etc.).

Turning to AI ethics, we offer a similar treatment. This is *the psychological*, *social*, *and political impact of AI.* This flows from our take on digital ethics presented above but is specified for AI. In the case of AI the psychological refers to the likes of mental autonomy, protection from undue manipulation, and the right to know when one is interacting with a non-human agent; the social refers to the likes of issues of justice and fairness (both procedural and substantive), as well as environmental concerns; and the political refers to impacts on democratic processes and the economy. Like digital ethics, AI ethics is thus highly interdisciplinary and is rapidly evolving.[6,10,11]

Below we expand by exploring the key constitutive terms, namely "digital" and "ethics."

Digital

In the context of digital ethics, "digital" is a reference to emerging technologies that are based on developments in computer science over the past decade. Indeed, they can be thought of more broadly in terms of the emerging socially disruptive technologies grouped by the OECD.[12] In our view, the key novel technologies we believe are indicative of this epoch are:

- *Blockchain:* decentralized record of digital transactions
- *Internet of things (IoT):* any device with an on/off switch that is connected to the internet
- *Big data analytics:* behavioral insight drawn from large information pool
- *AI/machine learning (ML) and associated algorithms:* automation of decision making, and performance of tasks that would normally require human intelligence

Importantly, these technologies are to be thought of together and as interrelated. For example, AI/ML has been developed and discussed since the mid-twentieth century; however, it has grown in importance and application through advances in computational power and the emergence of large datasets. As such, big data analytics is a product of AI and big data. Other examples are Federated Learning, which utilizes AI and blockchain,[13] and the evolution toward Smart Cities, which utilize AI, IoT, and big data analytics.[14] We concede that the above examples periodize the novel technologies to developments in the past decade or so, when others may extend the period to include early-stage AI (cf. symbolic AI, etc.) and the advent of the internet.

The current epoch is often referred to as the "fourth industrial revolution."[11] Where the first, second, and third industrial revolutions were characterized, respectively, by the use of water and steam power in the mechanization of production (circa 1750–1820), the use of electricity to power mass production (circa 1870–1920), and the use of electronics and information technologies in the mass automation of production and processing (circa 1950–). The fourth is a development upon the third (circa 1990–), and it is characterized by a fusion of technologies that blur the digital, physical, and biological spheres (e.g., cyberspace, virtual and augmented reality, body-machine interface, and robotics).[15]

Indeed, we can think of two additional themes namely (1) ubiquitous adoption of these technologies and (2) futurism. Where the former is a reference to the increasing use and normalization of such technologies in everyday life, government service provision, and industry. The latter is a reference to the philosophical/science fictionesque discussions that are emerging as a result of these changes (e.g., debates around the "singularity," transhumanism, and posthumanism, often presented in utopian/dystopian terms).[16] As such, the notion of digital ethics can be expanded and expressed in terms of the *impacts* of new digital technologies, through analysis of potential opportunities and risks in contemporary and future contexts (i.e., *it is an applied ethics*). Indeed, it is useful to point out that this is a reference to ethics in the broader impact sense, rather than the ethics of scholasticism.

AI

As the concern in this overview is with AI ethics, in this subsection we expand upon how we are using the term "AI"—listed as

one of the new digital technologies driving the fourth industrial revolution. The foundational term here is "algorithm," which we read in the broadest terms as *a set of rules or processes that aims to solve a problem or task.* In the digital realm, algorithms can be expressed in a computer through programming language. To understand this, we must first note that broadly there are two classes of AI algorithms, which might be termed: *static algorithms*—traditional programs that perform a fixed sequence of actions, usually classified as knowledge-based systems;[17,18] and *dynamic algorithms*—that learn and evolve by interacting with the environment, usually classified as ML algorithms.[19,20]

We can think of this as an "AI continuum" of epistemological models,[21] with the current most successful of which is

> *ML algorithms*—a type of program with the ability to learn without explicit programming, and can change when exposed to a new environment or information

Traditionally, ML can be broadly subdivided into:

> *Supervised learning*: a program is trained on available and processed data, where specified inputs are used to predict outputs
> *Unsupervised learning*: the goal of a program is uncover a hidden structure in the data, thereby "discovering" previously unknown patterns
> *Reinforcement learning*: the goal of the program is to make decisions, achieved through an iterative trial and error process that is mediated by a reward/penalization mechanism for decisions chosen in the development

ML applications in financial services can provide examples of these: Suppose a database of financial reports is available, if some of them have been historically labeled as positive and negative, we can leverage this to automatically tag future documents. This can be accomplished by training an algorithm in a supervised fashion. If these documents were unstructured, and spotting relations or topics is the goal (political events, economic data, etc.), an algorithm trained in an unsupervised manner can help uncover these hidden structures. Also, these documents can characterize the current state of the capital markets. Using that, an algorithm can decide which actions should be taken in order to maximize profits, hedge against certain risks, etc. By interacting and gaining feedback from the environment (markets), an algorithm can reinforce some behaviors so as to avoid future losses or inaccurate decisions.[22,23]

In addition to the above-mentioned subdivision of ML there are further and disruptive forms of more advanced ML systems that are making the resolution of previous problems cheaper, faster, and more scalable, like deep learning,[24] adversarial learning,[25] and transfer and meta learning.[26,27]

Importantly, we recognize that AI is used in a more general and lay vernacular, encompassing a more popular (non-engineering) use of the phrase. In this form, AI is used more in terms of general automation. We note this in recognition that those who participate in the AI ethics debate and literature often do so with the more common use of AI in mind. Indeed, we anticipate that delimiting what exactly is being referred to AI may become a point of contention in the future.

Ethics

Ethics is a broad discipline with considerable scope and plurality of understanding. Although there are calls and an increasing literature encompassing ethical perspectives of non-Western traditions in AI ethics literature (for example, the IEEE's *Ethically Aligned Design* [2017] calls for incorporation of non-Western ethical systems and highlights some of these, including ethics originating in Japan and Africa),[28] the predominant discourse of AI ethics is found within the Western European and North American contexts, which is also to be read in terms of our access to the English language literature. As such, while bearing in mind the increasing challenge to ethical frameworks representing solely the canon of literature that we write and research within, the "ethical" in AI ethics represents key concepts from this philosophical canon. For example, we broadly understand *ethics as the rational and systematic study of the standards of what is right and wrong*, *and morality as the commonly used term for notions of good and bad in more common use of the English language.*[29]

In addition to ethics and morality we include the theme of law into our broader notion of AI ethics. We do this because we maintain the philosophical position that law and ethics are highly interdependent and because we note that key themes of governance, accountability, and transparency (see section 'Major Themes in AI Ethics' 5) draw heavily on jurisprudence. We take law to be the *codified rules and guidelines in a particular jurisdiction.* Importantly, the law is enforceable, i.e., there is a coercive core (usually by the executive branch of the government). In philosophical ethics we think of this in terms of external constraints, which compares to the internal states described by notions of virtue.

These concepts are highly related and in natural language use of these terms are often interchangeable and synonymous in common parlance. *As such*, *the conceptual discussions we have offered and will offer below*, *should be understood as notions that allow AI ethics discussions to be structured rather than stable and inflexible concepts that the debate must be forced to fit within.*

Philosophy of ethics

The scope of philosophical ethics is vast, with various scholastic schools and traditions, each with its own community and considerable internal plurality (e.g., existentialism, utilitarianism, naturalism, egoism/hedonism, deontological/rights ethics, etc.). Two dominant approaches, which can be read as underpinning common law and continental law, are "utilitarianism" (often referred to as consequentialism) and rights-based ethics. An addition to these, and less reflected in the context of contemporary law, is virtue ethics. These are the three dominant ethical theories in academic philosophy of ethics. These three are thereby explored below.

1 *Utilitarianism:* formulation of principles by considering the consequences of actions that would result from those rules, where the maximization of pleasure/minimization of displeasure is sought. There are numerous interpretations of utilitarianism (e.g., "act" and "rule" utilitarianism) as well as questions regarding how the terms "pleasure" and "displeasure" are to be understood and quantified. Putting these scholastic concerns aside, the operative concept is that ethics is about weighing the consequences

of actions. One domain where this approach to ethics dominates is that of the justification of government policy and decision making (e.g., policy regarding healthcare is often justified by appeal to maximization of health outcomes for citizens (at population level), a claim that competes with dignitarian ethics of healthcare); economic policy is often justified through maximization of Gross Domestic Product, etc.

2 *Rights:* entitlements by virtue of belonging to a class. Here, the ethical framework is such that a series of "rights," i.e., that which is referred to as entitlements in the definition above, are conferred to a person simply by belonging to a class. Where class is understood as a generic category of identity.[29] Two central examples of rights are human and civil rights. Human rights are rights entitled to anyone in the class "human," and civil rights are rights entitled/conferred upon any citizen of a particular jurisdiction. Whereas human rights are considered inviolable and fundamental simply by virtue of being a human, civil rights are conferred to members of the political community.

3 *Virtue:* development of the character of an individual and actions that result as a consequence of good character. Virtue ethics (also known as natural ethics) is a classical position that is rooted in pre-enlightenment Aristotelianism. It is an approach to ethics that emphasizes character development—it is closely associated with "perfectionism," where a person develops over time toward an idealized notion of the perfect Self (often described in the religious terms of becoming godlike or being in union with the divine). Good character is understood in terms of values (read, virtues), such as honesty, self-control, integrity, courage, generosity, and fairness.

Human-centric AI

In all three ethical approaches, the central subject of concern is the human being, i.e., "persons." As such, it is necessary to offer a working definition of human, which we take to be a *rational animal*. There are numerous ways in which this definition can be understood; however, for the present purpose it is sufficient for us to emphasize that the class "human" is principally defined in terms of possession of the rational faculty. Thus, humans share all other characteristics with animals (movement, reproduction, etc.) but are differentiated into a separate class by reason. Reason itself requires fleshing out and can be thought of in broader terms, which include, "freedom", "volition," "intentionality," and "agency". These terms themselves are hotly debated within the philosophical literature; however, for our purposes, we can read them all as referring to *reason* as an ability to meaningfully make choices, i.e., agency and autonomy.

As such, the human is defined as an "agent" and "meaningful choice" is understood as self-conscious decision making. Following from this we can define "dignity" as respect for the moral status of human beings as rational agents making meaningful choices, i.e., existing autonomously. In the context of AI ethics, autonomy can be subdivided into mental and physical autonomy, where mental autonomy concerns respect for a person's deliberative faculties and processes (for example, the right not to be manipulated consciously or subconsciously), and where physical autonomy concerns respect for a person's body and choices over their own body.[29]

Furthermore, this aligns with a human centric approach to AI: *the development and deployment of AI systems that respect human dignity and autonomy.* Indeed, human centric AI can be thought of in positive terms, i.e., that automated systems should be developed and deployed for the betterment of humankind, to advance well-being, human dignity and human flourishing. When a system reflects this overarching value framework then it can be thought of as *trustworthy*, i.e., Trustworthy AI—as discussed by the European Commission's "Ethics Guidelines for Trustworthy AI" (2019).[30] In sections below we expand upon themes that fall under trustworthiness in the context of AI systems.

Conclusion

In the above we have explored philosophical ethics in the context of AI ethics; according to our view "the psychological, social, and political impact of AI," can now be understood as a judgment and assessment of these domains through the filter of the ethical approaches and terminology introduced above. *This is a study of normativity*, which is the evaluation and justification of the good.[31] In contrast to this is an anthropological approach to ethics, which is characterized not by evaluation and rational justification, but instead through study and observation of what people think and how they behave. It is to learn about ethical behavior in the world as it is.[32,33] Although we often think of law as the codification of the ethics of the popular will, most countries do not use direct referenda to legislate (cf. Swiss "Popular Initiatives"). As such there is a complex relationship between law, ethics, and morality (as discussed above), and therefore moving directly from anthropological ethics to codification of law is not standard practice. Instead, anthropological ethics can be read in terms of gauging respect for democracy, which will have consequences concerning trust in government and democracy, rather than in the straightforward determination of right and wrong or legality and illegality. One interesting example in the AI ethics literature is to postulate a "moral Turing test" where a system can be thought of as ethical if it can convince someone interacting with it that it is reasonably moral, such as how regular human interaction would accommodate ethical pluralism in human interactions.[34]

PREDECESSORS TO AI ETHICS

AI ethics is an emergent field that is still in its nascent phase and continues to evolve. However, there are a number of disciplines that have long traditions and literature from which AI ethics draws and can be seen as, in various ways, a continuation of. The three bodies of literature we believe are most relevant are (1) engineering ethics, (2) philosophy of technology, and (3) science and technology studies. Before we discuss each in turn below it is important to note that there are other bodies of literature, which we are not discussing, that many will reasonably argue are crucial predecessor/streams that feed into AI ethics.[6] Perhaps the most important is the ethics of robotics.[24,35,36] We certainly recognize this and our selection of the three chosen themes are not to be read as exhaustive. The reason we have chosen the fields is because the three span the extremes of interdisciplinarity—with, on the one hand the philosophy of

technology as the ''abstract'' non-empirical pole, and, on the other hand, engineering ethics as practitioner, highly applied, other pole (and science and technology studies straddling the middle). In addition to these predecessor disciplines, we note that the umbrella of ''responsible innovation'' as a new approach toward innovation is a strong contemporary current. This is demonstrated by the fact that social and ethical aspects are explicitly taken into account and economic, socio-cultural, and environmental aspects are balanced.[37,38]

Engineering ethics

Engineering ethics can be thought of in terms of the values and ethical systems relevant to the practice of engineering. Engineering is a term that refers to structuring, design, and building. It is perhaps the most inherently practical discipline within the broadly construed sciences (often referred to as an applied science). As such, it is natural that there already exists a body of literature discussing and debating the social and environmental impact of engineering. Although engineering encompasses numerous subdisciplines (such as civil, mechanical, computer, and chemical engineering), all of which have specific societal impacts, the field has matured to the point where non-subdiscipline-specific, general community-based engineering codes have emerged.

For example, the UK-based *Royal Academy of Engineers*, which is a membership by nomination and selection community of fellows, has a published ''Engineering Ethics'' guideline ''Code of Practice'' (2020).[39] The guideline is divided into two parts: first, a ''Statement of Ethical Principles'' (namely: (1) honesty and integrity; (2) respect for life, law, the environment, and public good; (3) accuracy and rigor; and (4) leadership and communication); and second, ''Engineering Ethics in Practice,'' which is the fleshing out of the statement of ethical principles through granular real-world case studies. There are other similar examples of this (see IEEE, ''Code of Ethics'' [2020]).[40]

It is noteworthy that engineering ethics is predominantly driven by self-assembled communities and associations, who develop their own standards, are process orientated (e.g., case study exploration, etc.), and typically go beyond legal compliance.

Philosophy of technology

The philosophy of technology can be thought of in terms of an *investigation into the nature of technology and how it impacts the individual*, *society*, *and the political.* There are numerous branches of philosophy that feed into digital ethics. This includes the philosophy of ethics, as discussed above, but also includes political and social philosophy. The most relevant philosophical predecessor is the literature referred to collectively as the ''philosophy of technology.''

The philosophy of technology emerged circa the 1920s and can be read as continuing until today (major figures are Martin Heidegger [d. 1976], Herbert Marcuse [d. 1979], and Jurgen Habermas). It is differentiated from the philosophy of science, which has a longer legacy in the history of philosophy and is concerned with method and knowledge. Contrastingly, philosophy of technology emerges as a result of technological innovation (where technology ''*tekhnē*'' means art or craft); indeed, it concerns the applications and uses of discoveries in science. The principal focus is in the appraisal of how technology affects the human condition and whether the technology is neutral or value laden (e.g., exploration of whether nuclear technology is inherently good or bad, or whether it is dependent on the deployment of the technology in particular contexts).[41] From a historical perspective, the philosophy of technology can be read as a response to the overtly optimistic attitudes of the enlightenment and ''positivism,'' as well as the post-WWII world, which had an ever-present nuclear threat and a 1960s counterculture that was typified by ''social conscious.'' Indeed, it challenges the idea that there is a necessary connection between scientific discovery and scientific progress, and that this progress includes, and is extended to, society. The literature is typically negative, highlighting the dangers, risks, and loss of meaning through adoption of new technologies and increased technocratization with respect to the ordering principles of society. Key themes are automation, alienation, destruction, and loss of connection to nature, uniformity, shallow consumption, and excessive rationalization. However, it must also be mentioned that there are more positive philosophy of technology views in the contemporary debate (e.g., Verbeek's mediation theory [2015]).[42]

In addition to noting the general negative critique of technology in the philosophy of technology, it is also noteworthy that the perspective of the engineers and scientists themselves, i.e., the practitioners, are missing. This ''negativity'' also ignores empirical evidence that those who have a less negative view could cite, namely that technologies, such as vaccinations and other medical equipment (e.g., pacemakers, etc.) have tangibly and demonstrably decreased mortality rates. This ''turn to evidence'' is discussed in the next subsection.

Science and technology studies

Science and technology studies can be thought of in terms of an investigation into the effect of culture, society, and politics on scientific research/activity and technological innovation, and into the effect of scientific research/activity and technological innovation on culture, society, and politics.

The philosophy of technology can be criticized from several points. Two of these are that it is (1) typified by moral panic and (2) that it is non-empirical. Considering moral panic, namely *the phenomena of an acute reaction to a shocking/dramatic event*, *typified by knee-jerk reactionaryism*—the negative critique and commentary is read as excessive and blind to the benefits that such technologies have conferred to humanity. Indeed, it is read as highly politicized, with the philosophy of technology being instrumentalized as a polemical force of political intervention rather than as a considered investigation into the nature and impact of technological innovation.

The accusation that the philosophy of technology is non-empirical is motivated by the evolution and development of the social ''sciences,'' where sciences are placed in quotation marks due to the increased incorporation and methodological approach of the empirical (natural) sciences in the humanities (such as sociology and anthropology). Increasingly sophisticated mechanisms of empirical investigation—that survey, test hypothesis, and observe through data analysis (cf. natural sciences)—were brought to questions that the philosophers of technology were commentating on.

This empirical turn gave birth to science and technologies studies (circa 1980s–).[43] Major figures include Bruno Latour,

Andrew Light, and Donna Haraway. Science and technology studies problematized the overwhelmingly negative critique of technology by pointing out that technology can be designed and used differently and this can lead to radically different social outcomes (e.g., technological developments associated with the internet can be instrumentalized for mass totalitarian surveillance or instrumentalized to facilitate radical anonymity; similarly, nuclear technology can be used to make devastating weapons or to produce a steady and reliable source of energy). In other words, technologies can be used for good and for bad. Evaluation of the value system that embodies and is present in the deployment of technologies, is done through empirical investigation. Questions can be investigated empirically regarding who is building the technology, who is using it, and how it has actually impacted society (and particular groups within society). The notion that there is a social constructionism/contingency to technology is central to science and technology studies.[44,45] An example of this sociological/empirical turn can be given with respect to the impact and evaluation of mass consumption (characterized as shallow consumption by traditional philosophers of technology): surveys/studies conducted showed that people enjoy mass culture, cinema, music, etc., and that they continue to do so after being presented with the arguments against the ''culture industry.''[46] This raised a counter critique directed toward the philosophers of technology, namely that they fail to respect the aesthetic tastes and autonomy of non-philosophers and depict people as a gullible mass, i.e., patronization.

Conclusion

In the previous section we presented a number of definitions, in this section we have presented some predecessor disciples that we believe the field of AI ethics can draw from. These definitions and predecessor disciplines will inform what is meant and what is being drawn upon, in the burgeoning field of AI ethics. In the section below we sketch the current landscape of AI ethics through a high-level analysis of the main approaches. These are (1) the principles approach, (2) ethical-by-design/processes approach, and (3) ethical consciousness approach.

THREE APPROACHES: PRINCIPLES, PROCESSES, AND ETHICAL CONSCIOUSNESS

In the above, we defined AI ethics in terms of impact analysis: this can be read as a response to the increasing number of high-profile cases of harm that has resulted either because of the *misuse* of the technology (e.g., psychometric voter manipulation,[1,2] facial recognition surveillance, mass data collection without consent) or as a result of the technology having *design flaws* (e.g., bias in cases of recidivism,[3] loan rejection, and medical misdiagnosis).[4] As a result, the two main approaches to AI ethics has been a principles approach, which can be read broadly as an attempt to guide and structure the uses of the technology (thereby mitigating the risk of misuse) and an ethical-by-design approach, which seeks to mitigate the harms that result from design flaws. Below, we explore these two approaches and also a third, denoted as ''ethical consciousness,'' which draws from the business ethics literature and concerns a need to institute particular structures and shifts in cultures, attitudes, and norms of those who use, develop, and deploy AI systems.

Principles

The most vocal response to the harms and risks of new digital technologies, in particular AI, has been to call for guidelines that inform and direct the use and development of these technologies. Attempts to delimit these principles fall into three main categories (1) abstract first-principles, (2) development and application of legislative standards and norms, and (3) to make analogy with bio/medical ethics. Below, we sketch these.

Abstract first-principles

The first, principles, approach is to articulate a number of statements, typically the expression of a set of values, and present these as guidance and standards with which the AI systems (and other novel digital technologies) can be developed and deployed. Such statements of principles have been produced by all relevant stakeholders, namely academia,[47] industry (e.g., *Google: Artificial Intelligence Principles*, 2020),[48] NGOs (*The Asilomar AI Principles*, 2017;[49] *The Montreal Declaration for Responsible AI*, 2017),[50] and government (UK House of Lords Select Committee on Communications, 2018)[51]—although the majority of have been in the context AI, data ethics is heavily drawn upon.

The principle approach is plagued by several problems (here the examples cited are from the UK House of Lords report ''Regulating in a Digital World'' [2018]). First, the principles are mostly vague and thereby difficult to interpret (e.g., the accountability principle ''individuals and organizations need to be held to account,'' does not provide an expansive explanation of how this would look in practice).[52,53] Second, the principles are incongruent, i.e., within the same set of principles, the individual principles contradict one another (e.g., both the principle of ''openness'' and the principle of ''privacy'' are asserted together, where respect for one is likely to come at the cost of implementation and respect for the other). Third, there is considerable lack of clarity regarding terminology/concepts (e.g., principles of accountability, transparency, and openness are complimentary and, in some respects, synonymous expressions). Finally, although there is overlap between the various statements of principles, there is also a clear lack of consensus. Indeed, to date (April 2020) there are over 80 AI statements of principles.[5] From an engineering perspective, these problems make it difficult to translate the principles into practice.

Legislation

The most direct way in which to approach AI ethics is to ensure that the technologies are developed and deployed in a lawful manner. Indeed, legal compliance is a clear and objective standard by which to judge and evaluate ethics (where lawfulness can be read as a necessary but insufficient condition of ethics). However, straightforward recourse to the law is not possible. This is because there are a number of nuanced concerns. First, the question is posed as to whether it is necessary to create new laws or to update and ensure application of existing ones (see European Commission [2020]).[54] Indeed, one approach that can be taken is to allow a body of case law to emerge and derive standards and, if need be, legislate accordingly based on this, i.e., a ''bottom-up'' approach (see Committee on Standards in Public Life [2020]).[55] Second, there is the question of

whether legislation is appropriate at all, with other options possible, such as self-regulation and/or a standards body.[56,57] Third, there is the issue of jurisdiction, where technological development and deployment are obscured through internationalism, which does not respect or easily lend to jurisdictional oversight (e.g., nation states and international unions [European Union, African Union, etc.]). This also allays itself the problem of enforcing the law. Fourth, in the common law tradition (typically found in the anglophone), based on case and precedent, the statutes accommodate ambiguity and contradiction (relying on judicial discernment). This is a problem with respect to automated systems, where increasingly there is a call to automate regulatory compliance via expressing laws in codes/protocols (the ambiguity and contradiction accommodated by common law does not translate in this context). Regarding this specific concern, continental law (exemplified by EU's legislative agenda) typified through a top-down (first-principles) legal philosophy is less likely to have this problem. Finally, and more specifically, in the context of AI there is the question regarding the legal status of an algorithm (e.g., will algorithms follow how companies have rights and obligations, and will AI systems have artificial personhood status),[58] and how the questions of legal culpability, which rely on judgments of intent, can be formulated in the context of AI systems.[59]

Bio/medical ethics

Bio/medical ethics can be turned to for inspiration when approaching and developing AI ethics due to the fact that it is well established, robust, has accountability mechanisms and is an example of the ethics of an applied science with significant social impact[60–64]; this is not to be confused with applications of new technologies in medicine)[65] or the ethics of using AI in medicine.[66] However, there are important disanalogies between bio/medical ethics and any AI ethics scheme that may emerge. In the context of AI systems there is: first, no common aims and fiduciary duties (legal or ethical relationships of trust), i.e., the relationship between the doctor and the patient is disanalogous to the relationship between the engineer/company and the public); second, there is no professional history and norms, i.e., the technologies are still at an early stage and the field of AI ethics is still contested (as discussed above); and finally, no robust legal and professional accountability mechanisms is present, i.e., in medicine a doctor can be "struck off" and/or a license withdrawn.[67]

Processes

A second approach to AI ethics is to address risk and harm that can result because of design issues and lack of appropriate governance.

Ethical-by-design

An ethical-by-design approach is a commitment to building systems ethically and in the hope that harm can be prevented. There are several approaches to ethical-by-design. First, via co-design, which is a reference to interdisciplinarity in the design processes. The idea here is that AI engineers may not be best placed to understand and discern the ethical dimension and potential impact of the technology and, as such, experts from anthropology, sociology, philosophy, psychology, law, etc., i.e., "ethicists," can be integrated into the team at the development stage. Second, by having clear principles, laws, standards, and guidelines with which to structure and judge design. As noted above, there is considerable ambiguity and lack of consensus in the field of digital regulation (perhaps with the exception of GDPR) and standards, thus making it difficult to establish best practices in the domain of translating principles into engineering practice. Finally, and as a corollary to the previous point, from a design perspective, implementation of ethical principles will need to be balanced and "traded-off." For example, in the design, emphasis on transparency and openness may come at the cost of privacy. These judgments need to be justified, articulated, and expressed in different forms in various contexts. With this practical dimension and necessity of trade-offs in mind, several practical manuals have been published, all of which note these trade-offs and discuss ethical evaluation (including interdisciplinarity) in the design, development, and deployment phases, respectively.[28]

Governance

Within the processes approach to AI ethics questions of governance are emerging. More generally, with respect to novel digital technologies governance can be divided into two broad streams, namely technical and non-technical (with ethical-by-design falling into the former category). Technical governance concerns systems and processes that render the activity of the technology itself accountable and transparent—this includes justifying what design choices are made and ensuring the system is accessible. Non-technical governance concerns systems and processes that focus on allocating decision makers, providing appropriate training and education (in the context of new digital technologies, such as AI, education and training will require continuous updating), and keeping the human-in-the-loop with respect to how automated decisions are used while respecting human rights (often referred to in the context of human centric AI).[9]

Falling under governance is the growing literature on auditing and impact assessments. Auditing and impact assessments involve the creation of metrics for tracing and tracking decisions, making the technologies accessible for verification and accountability.[68] The most well-established form of this is Data Protection Impact Assessments (DPIA)[69,70] To date more general "data ethics" canvases and process frameworks have emerged,[71] and impact assessments specific to AI are being called for and developed.[54,72]

Ethical consciousness

Ethical consciousness refers to a person or institution or cultural norm, that has a disposition that is motivated by a moral awareness rather than, say, exclusively a concern of economics (pay and profit), or legality (responsibility, culpability, and compliance). In other words, this is a desire to "do the right thing". Ethical consciousness can be read as coming out of business ethics,[10] which is an applied ethics within the commercial environment. Sharing many of the themes from the previous section, it encompasses the integration of codes of conduct and compliance; however, it also expands to consider reputational issues, (corporate) social responsibility, and, most relevant to the development of ethical consciousness, concerns for institutional philosophy and culture. Drawing particularly on the latter, ethical consciousness can be stated in terms of societal and culture shifts in the awareness of citizens, technology developers and deployers, policy makers, and leaders of industry, in the ethical dimensions of new digital technologies. Such a shift will be

facilitated through an increase in digital literacy, particularly important for meaningful human intervention and issues of consent.[73]

Conclusion

The above approaches can be thought of in terms of (1) the theoretical and abstract, (2) the practical and process, and (3) culture and society, and in turn can be thought of as all necessary in the development of a mature AI ethics, i.e., one that is truly reflective of the nuances and the inherent complexity found in all forms of applied ethics. This also challenges any attempt to silo questions and responsibilities; AI ethics is not exclusively in the purview of philosophers or lawyers or sociologists or engineers, etc., rather it is inherently interdisciplinary. As such, the call for interdisciplinarity must be followed with development of methods and structures by which they can be fulfilled. It is likely that this aligns with the call for training and education (noted above in the context of discussions regarding governance); as such an increase in "literacy" is crucial in informing all relevant stakeholders and society at large and will be facilitated through a holistic education and training agenda (touching upon and integrating ethics, policy, and engineering).

MAJOR THEMES IN AI ETHICS

There are many terms and phrases that have emerged within the AI ethics literature. For example, a comprehensive 2019 review of AI ethics guidelines found eleven "ethical principles" namely (1) transparency; (2) justice, fairness, equity; (3) non-maleficence; (4) responsibility and accountability; (5) privacy; (6) beneficence; (7) freedom and autonomy; (8) trust; (9) dignity; (10) sustainability; and (11) solidarity. These principles were identified through frequency of the terms (and their synonyms) in the literature (the terms above are listed in order of prevalence). However, as noted above, there is overlap and these terms require considerable disambiguation.[5] For the purposes of this overview we draw on the growing engineering expertise that overlaps with the ethics principle space. Indeed, in the section below, we identify and explore six themes that we believe encompasses the attempt to bridge the need to implement ethics into engineering and systems. These are, human agency and oversight, safety, privacy, transparency, fairness, and accountability. Drawing on the European Commission's "Ethics Guidelines for Trustworthy AI" (2019),[30] we note that these themes can be read as falling under the umbrella of "Trustworthy AI" (introduced above in Concepts and terms).

Human well-being

This theme is grounded on the ethical principle of respect for human dignity (see section Concepts and terms) and includes psychological, social, and environmental well-being. Here, key themes are:

- *Impact on human agency:* this touches on the impact on individuals and, in particular, mental autonomy. For example, consider whether a system directly or indirectly diminishes the deliberative/rational capacity of humans (e.g., cognitive shifts in attention spans). Another issue is consent, where respect for human agency would entail meaningful and informed consent, including the right to withdraw consent and be presented with consent mechanisms that are explicable in the context of an average user.
- *Societal impact:* the societal refers to identity, belonging, and communities, and includes the political legal/jurisdictional, democratic, and economic impacts. Citizen rights fall here, where issues of fairness—both procedural and substantive, and bias should address. With respect to the economic, concerns include, fair competition as well as setting the framework for competition.[74] Finally, the environmental impact should be considered, including issues of sustainability.

The above may be read as a re-statement of the ethical imperative that grounds human centric AI, i.e., that automated systems should be developed and deployed for the betterment of humankind, to advance well-being (or at least not adversely affect it), human dignity, and human flourishing. Indeed, although the discussion is presented in terms of mitigating risks and potential harms, it is important to bear in mind the considerable benefits of AI for people and society.[75]

Safety

This theme is based on the ethical principle of preventing harm, where harm is defined in terms of adverse effects on human well-being, i.e., the psychological, social, and environmental human well-being.

Here, the approach is one of identifying risks and then mitigating for them: crucially the approach is preventative. Key themes here are:

- *Robustness:* systems should be robust against adversarial attacks, i.e., hacking. Here, resilience is important and that there are measures to stop/resist exploitation of a system (e.g., data poisoning, model leakage).
- *Malicious use:* a system may have been developed for one use and then be appropriated and/or modified for another, malicious, use, i.e., dual use (e.g., the weaponization of delivery drones).
- *Reliability and reproducibility:* reliability concerns the system working within the framework of why it was developed and deployed, whereas reproducibility concerns consistent behavior when given the same set of inputs and under the same conditions. In the context of robustness this is important because a system that is unreliable and does not reproduce results will lead to untrustworthiness in the system.
- *Fallback plans and unknown risks:* a concern for robustness is to address known risks (such as those cited above, i.e., security, malicious use, reliability, and reproducibility) and unknown risks. With respect to the former, safeguards can be put in that specifically monitor and track usage and/or metrics of known risks and put in place stops or other mechanisms that would mitigate this risk. With respect to the latter, it is not possible to fully anticipate risk and, as such, mechanisms can be put into place to mitigate this (e.g., fallback mechanisms, automatic stops—statistical or rule-based, metrics, periodic request for human operators to continue operating, etc.).

Privacy

This theme is based on the public and political demand to respect a human's personal information. This relies on a distinction between the private-personal and the public-political/communal sphere, where the former is seen as demanding a higher level of respect for privacy than the latter.[76] Informed consent is crucial here, where people are informed and updated regarding the storage and use of their data. Furthermore, there are debates concerning the value derived from personal data and the distribution of financial benefits derived thereof (e.g., mass data-driven business models). In addition, privacy has emerged as a political concern, with mass surveillance and personal data being used to target and engage in recommendation and manipulation (for both political and economic ends).

- *Data stewardship:* the management of data spans many stages, including collection, pre-processing, tracking providence, analysis, publication of results, re-use, and recycling, all while maintaining security and, where appropriate, anonymizing. Stewardship is the management of this multi-layered process and has come to be a discipline and set of skills in its own right. Part of this remit is data protection, which is crucial to preserving privacy with respect to who has access to the data (in particular personal data).
- *Data minimization:* within the context of privacy and data protection, a generalized principle to use only the amount of data that is needed is referred to as data minimization. Here, three dimensions are identified, namely (1) adequacy: where the data are sufficient to fulfill a stated purpose; (2) relevant: where the data have a justifiable link to the stated purpose; and (3) necessary: where the data are limited and no more than what is needed is held (and, where appropriate, deleted when no-longer in use for the stated purpose).

Transparency

This theme is based on the principle of openness, which is crucial to establishing trust and accountability. Transparency can be thought of with respect to what decisions are being made regarding how the AI system is used and with respect to how the system comes to its decisions. The former touches on governance (which is also expanded upon in the accountability section below), whereas the latter concerns explainability of automated decision makers. As such key themes are:

- *Explainability:* being able to explain how a system has come to a decision (cf. black box problem) and making that decision explicable to various stakeholders, i.e., explicability will depend on the technical knowledge of the person, what role they play in the development and deployment of the system, and what kind of end-user they are. Furthermore, there are a host of technical requirements and tools that may be grouped under explainability, namely, accuracy, traceability, tracking, general (global/model), and specific (local/data point) explanations.
- *Communication:* in addition to explicability, i.e., communication, of automated decisions, there is also the concern for communicating the capabilities and purposes of the system to those both directly and indirectly impacted.[77] One crucial dimension is that, in cases where a system may be mimicking human subjectivity (e.g., a chatbot), it should be communicated clearly that the user is interacting with an AI system.

Fairness

This theme is based on the ethical principle of human equality. Fairness falls under debates about justice and is hotly contended. A central question is what definition(s) of fairness/justice to commit to, i.e., there are mutually exclusive theories of fairness, such as corrective, distributive, procedural, substantive, comparative, etc. The question is also raised as to the scope or remit within which notions of fairness/justice are being discussed, i.e., fairness in the context of political communities (citizenship rights) and/or universal human concerns, and, if appropriate, how to define demographics, i.e., gender, nationality, race, socio-economic background, etc. Key themes here are:

- *Bias:* here, bias refers to preferential or discriminatory treatment of persons or groups. Concerns touch upon bias in (historical) datasets, intentional exploitation of people (e.g., customers/regional pricing), and quality of service provision. This also includes a distinction between fairness in terms of treatment and fairness in terms of impact.[78]
- *Accessibility:* although much of the discourse in AI ethics concerns mitigation of harms, it is also clear that there are significant benefits that people and society will gain from these systems. As such, it is paramount that all people, to the greatest extent possible, have equal access to these technologies; aside from affordability, designs should be user-friendly (e.g., toward different demographics, cultural and linguistic groups, and, in particular, those with disabilities), i.e., there is not a one-size-fits-all approach.
- *Participation:* communication, in an accessible and explicable vernacular, to users will facilitate meaningful engagement of wider society with AI systems. This will also facilitate learning and develop a more holistic approach to consent. Participation also includes, soliciting the views of stakeholders during the development of the system. This expands to diversity (in option and background) in hiring and the interdisciplinary teams involved in governance and development.

Accountability

Ethical AI is a branch of applied ethics and, as such, is inherently concerned with how AI systems impact human beings. How the systems are developed, the processes, logic of decision making, the allocation of duties with respect to who makes decisions, and how and to what extent, where impacts, risks, and harms gauged and measured. All of this falls under the remit of accountability, and, as the previous sentence indicates accountability relates to knowing who had made decisions, how those decisions were made, and what systems or tools were put in place to measure and track, i.e., governance. Finally, accountability is central for the possibility of redress and assigning legal liability. Key themes here are:

- *Keeping it human:* Crucial to accountability is ensuring that there are robust human oversight mechanisms, this is based on the principle, and current legal standing (see section Legislation), that humans are ultimately accountable and thereby responsible for harms that may result from AI systems. Within the literature, there is a growing discourse regarding keeping the "human-in-the-loop,"[79] which is discussed in two ways, firstly as ensuring that there is human intervention in the decision processes of automated systems and secondly in terms of human oversight regarding automated decisions, i.e., in the context of "decision support systems." With respect to the latter, a "semi-automated decision" scheme can be thought of, where a system generates results, directions, and recommendations (e.g., whether to hire someone, or reject a loan application) and is followed by human review in order to affirm or reject the recommendation (Information Commissioner's Office, 2020).[80,81] In these cases, checks should be in place to assure that the human review does not become a rubber stamp exercise rendering the decisions effectively solely automated. Ensuring human responsibility also entails mapping of duties and risks to responsibilities and roles within an institution.
- *Algorithmic impact assessments:* this relates to direct mechanisms by which to assess and thereby put in place measures to mitigate potential harms of AI systems. We can divide this into two approaches, namely (1) impact assessments and (2) auditing of technology.[82]
  1. Impact assessments can range from assessments of fundamental rights, psychological and social well-being (e.g., social cohesion), citizen rights, democracy, economic, and environmental impacts.
  2. Auditing can be directed to the technology itself, and focus on fairness (e.g., tracking bias metrics), explainability (e.g., providing global and local explanations of models and individual decisions), and robustness (e.g., testing how resilient a system is to hacking).

It is important that these are conducted in such a way as to facilitate inspection (perhaps even independently conducted). Moreover, clear documentation is necessary. This includes documenting any trade-offs and the methodology and logic behind trade-off choices.

CONCLUSION

In this high-level overview and introduction, we have offered basic conceptual overviews of terms, such as AI and ethics. Following this, we explored some predecessors to AI ethics, namely engineering ethics, philosophy of technology, and science and technology studies. We then discussed three current approaches to AI ethics, namely, principles, processes, and ethical consciousness. Turning to translating AI ethics into engineering practice we surveyed the themes of human centric AI, safety, transparency, fairness, and privacy.

We believe that AI ethics will develop such that it will become clear that the field is inherently interdisciplinary. For example, some themes in AI ethics that we see as *necessarily* requiring interdisciplinarity are:

- *Data ethics and AI*: given the substantial literature, practice, and regulation around data ethics, we anticipate that the relationship between data ethics and AI will increase in importance. This includes whether the two are compatible, whether one is prioritized over the over, i.e., will AI ethics "sit" on top of data ethics or will data ethics have to be reconsidered and reformulated in light of increased AI adoption? We anticipate that this debate will be both conceptually important and have significant regulatory/practical consequences .[83]
- *Legal status of algorithms*: raised 4.1.2earlier, the legal status of an algorithm with respect to responsibilities and obligations of those developing and deploying them, is likely to raise a number of complex questions regarding the nature of legal culpability and even questions of agency and personhood. We anticipate that this question will increase in complexity and importance the more AI systems are embedded in people's daily lives and in proportion to the function of these systems (e.g., sectors, such as medicine, may require nuances that other sectors, such as entertainment, will not).[84]
- *Economic impacts*: we believe that the relationship between AI and the economy will become a major theme of AI ethics. In addition to the current discussion of automation and the loss of labor, which allay into questions, such as universal basic income, etc., there are broader questions regarding taxation of AI systems, national and international procurement standards and strategies, and the strategic importance of AI in national budgets.
- *Political impacts:* beyond concerns for misuse of AI systems in the democratic systems (e.g., voter manipulation), debates about how AI impacts the structure of the state, the very notion of a nation (with clear juridical remit), and trust in government communication, management, and service provision, will become central themes within AI ethics.
- *Nature centric AI*: we believe that concerns regarding the natural environment and climate will also feature strongly in AI ethics. Here, there is the basic ethical debate regarding the energy resources that AI requires and whether this is justified. However, beyond this there is the very conception of ethics that includes the environment as centrally as it does human concerns, i.e., nature centric, where "nature" includes humans, animals, and the natural environment.

In this overview, readers will have been introduced to a plethora of concepts and array of ideas from a range of disciplines and literature—as noted in Concepts and Terms2, readers should take a critical view to our assertions and use them as a point of departure for further thought and exploration. In this vein, we recognize AI ethics as a nascent field that is open and plural in its various standpoints. We view this plurality positively and, as such, welcome the multiple frameworks and perspectives that are currently present. However, with a longer-term view in mind, while respecting this plurality, we hope that common terms and key concepts will be shared. This article should be read with this in mind.

ACKNOWLEDGMENTS

The authors would like to acknowledge Cisco Research Centre for their research grant (2020-222054 3696).

AUTHOR CONTRIBUTIONS

The authors, Emre Kazim and Adriano Soares Koshiyama, have contributed to the article via their respective specialisms. Emre has primarily contributed the sections on ethics and predecessors to AI ethics, and Adriano to sections providing an overview of AI/ML and section Major thmes in AI ethics 5. However, both authors have iterated throughout the text.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Krotzek, J.L. (2019). Inside the voter's mind: the effect of psychometric microtargeting on feelings toward and propensity to vote for a candidate. Int. J. Commun. *13*, 3609–3629.
2. Costa, E., and Halpern, D. (2019). The behavioural science of online harm and manipulation, and what to do about it: an exploratory paper to spark ideas and debate. In The Behavioural Insights Team Report, pp. 1–82.
3. Chouldechova, A. (2016). Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. arXiv, 1–6.
4. Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., et al. (2018). AI Fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv, 1810.01943.
5. Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. Nat. Mach. Intell. *1*, 389–399.
6. Müller, V.C. (2020). Ethics of artificial intelligence and robotics. In The Stanford Encyclopedia of Philosophy, E.N. Zalta, ed. https://plato.stanford.edu/archives/win2020/entries/ethics-ai/.
7. Coeckelbergh, M. (2020). AI Ethics (MIT Press).
8. Gordon, J.-S., and Nyholm, S. (2021). Ethics of Artificial Intelligence (Internet Encyclopedia of Philosophy).
9. Lukowicz, P. (2019). The challenge of human centric AI. Digitale Welt *3*, 9–10.
10. Kriebitz, A., and Lütge, C. (2020). Artificial intelligence and human rights: a business ethical assessment. Bus. Hum. Rights J. *5*, 84–104.
11. Floridi, L. (2014). The 4th Revolution (Oxford University Press).
12. OECD, K. (2018). OECD Science, Technology and Innovation Outlook 2018 (OECD Publishing).
13. Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., and Van Overveldt, T. (2019). Towards federated learning at scale: system design. arXiv, 1902.01046.
14. Burange, A.W., and Misalkar, H.D. (2015). Review of Internet of Things in development of smart cities with data management & privacy. In 2015 International Conference on Advances in Computer Engineering and Applications (IEEE), pp. 189–195.
15. Phillips, A.M., and Mian, I.S. (2019). Governance and assessment of future spaces: a discussion of some issues raised by the possibilities of human–machine mergers. Development *62*, 66–80.
16. Porter, A. (2017). Bioethics and transhumanism. J. Med. Philos. *42*, 237–260, Oxford University Press.
17. G. Schreiber, B. Wielinga, and J. Breuker, eds. (1993). KADS: A Principled Approach to Knowledge-Based System Development, *Vol. 11* (Academic Press).
18. Giarratano, J.C., and Riley, G. (1998). Expert Systems (PWS publishing co).
19. Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Springer Science & Business Media).
20. Sutton, R.S., and Barto, A.G. (2018). Reinforcement Learning: An Introduction (MIT press).
21. Russell, S.J., and Norvig, P. (2016). Artificial Intelligence: A Modern Approach (Pearson Education Limited).
22. Treleaven, P. (2015). Financial Regulation of Fintech Journal of Financial Perspectives, *Vol. 3*. https://ssrn.com/abstract=3084015.
23. Barber, D. (2012). Bayesian Reasoning and Machine Learning (Cambridge University Press).
24. Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning (MIT press).
25. Huang, L., Joseph, A.D., Nelson, B., Rubinstein, B.I., and Tygar, J.D. (2011). Adversarial machine learning. In Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence (ACM), pp. 43–58.
26. Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M.W., Pfau, D., Schaul, T., and De Freitas, N. (2016). Learning to learn by gradient descent by gradient descent. In Advances in Neural Information Processing Systems, pp. 3981–3989.
27. Devlin, J., Chang, M.W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. arXiv, 1810.04805.
28. Ethically Aligned Design (2017). A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems (IEEE), Version 2.
29. Kazim, E. (2017). Kant on Conscience (Brill).
30. Ethics Guidelines for Trustworthy AI: High-level Expert Group on Artificial Intelligence (8th April 2019) European Commission.
31. Hanna, R., and Kazim, E. (2021). Philosophical foundations for digital ethics and AI ethics: a dignitarian approach. AI and Ethics, 1–19.
32. Awad, E., Dsouza, S., Kim, R., Schulz, K., Henrich, J., Shariff, A., and Rahwan, I. (2018). The moral machine experiment. Nature *563*, 59–64.
33. Arvan, M. (2018). Mental Time-Travel, Semantic Flexibility, and AI Ethics (AI & SOCIETY), pp. 1–20.
34. Allen, C., Varner, G., and Zinser, J. (2000). Prolegomena to any future artificial moral agent. J. Exp. Theor. Artif. Intell. *12*, 251–261.
35. Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., et al. (2018). A general reinforcement learning algorithm that masters chess, Shogi, and Go through self-play. Science *362*, 1140–1144. https://doi.org/10.1126/science.aar6404.
36. Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., et al. (2016). Artificial Intelligence and Life in 2030, One Hundred Year Study on Artificial Intelligence: Report of the 2015–2016 Study Panel (Stanford University).
37. von Schomberg R. (2011). European Commission. Towards responsible research and innovation in the information and communication technologies and security technologies fields. Available at SSRN 2436399. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2436399.
38. von Schomberg, L., and Blok, V. (2019). Technology in the Age of Innovation: Responsible Innovation as a New Subdomain within the Philosophy of Technology (Philosophy & Technology), pp. 1–15.
39. (2019). Engineering ethics (2020) Royal Academy of Engineering. https://www.raeng.org.uk/policy/engineering-ethics/ethics.
40. (2019). IEEE code of ethics. (2020). IEEE. https://www.ieee.org/about/corporate/governance/p7-8.html.
41. Balabanian, N. (2006). On the presumed neutrality of technology. IEEE Technol. Soc. Mag. *25*, 15–25.
42. Verbeek, P.P. (2015). COVER STORY beyond interaction: a short introduction to mediation theory. Interactions *22*, 26–31.

43. M. Franssen, P.E. Vermaas, P. Kroes, and A.W.M. Meijers, eds. (2016). Philosophy of Technology after the Empirical Turn (Springer).

44. Winner, L. (1980). Do artifacts have politics? Daedalus *109*, 121–136.

45. Zittrain, J.L. (2006). The generative internet. Harvard Law Journal *119*. https://doi.org/10.1145/1435417.1435426.

46. Curran, J., and Gurevitch, M. (1996). Mass Media and Society, 2nd Edition (Edward Arnold), pp. 305–324.

47. Floridi, L., and Cowls, J. (2019). A unified framework of five principles for AI in society. Harv. Data Sci. Rev. (1.1), 1–15. https://doi.org/10.1162/99608f92.8cd550d1.

48. (2020). Google: artificial intelligence principles. https://ai.google/principles/.

49. (2017). The Asilomar AI principles. https://futureoflife.org/ai-principles/?cn-reloaded=1.

50. (2017). The Montreal Declaration for Responsible AI. https://www.montrealdeclaration-responsibleai.com/.

51. Regulating in a Digital World. (2019). House of Lords Select Committee on Communications. https://publications.parliament.uk/pa/ld201719/ldselect/ldcomuni/299/299.pdf.

52. Verma, S., and Rubin, J. (2018). Fairness definitions explained. In 2018 IEEE/ACM International Workshop on Software Fairness (FairWare) (IEEE), pp. 1–7.

53. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R.S. (2011). Fairness through awareness. arXiv, CoRR abs/1104.3913 (2011). preprint arXiv:1104.3913.

54. White Paper on Artificial Intelligence (2020). A European Approach to Excellence and Trust (European Commission).

55. Artificial Intelligence and Public Standards: report. (2020). Committee on Standards in Public Life, Chair, Lord Evans of Weardale KCB DL.

56. Simon, J.P. (2019). Artificial Intelligence: Scope, Players, Markets and Geography (Digital Policy, Regulation and Governance).

57. Lauterbach, A. (2019). Artificial Intelligence and Policy: Quo Vadis? (Digital Policy, Regulation and Governance).

58. Treleaven, P., Barnett, J., and Koshiyama, A. (2019). Algorithms: law and regulation. Computer *52*, 32–40.

59. Vetrò, A., Santangelo, A., Beretta, E., and De Martin, J.C. (2019). AI: From Rational Agents to Socially Responsible Agents (Digital Policy, Regulation and Governance).

60. Mittelstadt, B., and Floridi, L. (2016). The ethics of big data: current and foreseeable issues in biomedical contexts. Sci. Eng. Ethics *22*, 303–341.

61. Mittelstadt, B. (2017). Designing the health-related internet of things: ethical principles and guidelines. Information *8*, 77.

62. Christine, H. (2019). Making the ethical practical: ideas for applying a digital health ethical framework for charities. DataKind.org. https://www.datakind.org/blog/making-the-ethical-practical-ideas-for-applying-a-digital-health-ethical-framework-for-charities.

63. Panesar, A. (2021). Machine learning and AI ethics. In Machine Learning and AI for Healthcare (Apress), pp. 207–247. https://doi.org/10.1007/978-1-4842-6537-6_8.

64. Arnold, M.H. (2021). Teasing out artificial intelligence in medicine: an ethical critique of artificial intelligence and machine learning in medicine. J. Bioeth. Inq. https://doi.org/10.1007/s11673-020-10080-1.

65. Char, D.S., Shah, N.H., and Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. N. Engl. J. Med. *378*, 981.

66. Lamanna, C., and Byrne, L. (2018). Should artificial intelligence augment medical decision making? The case for an autonomy algorithm. AMA J. Ethics *20*, 902–910.

67. Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nat. Machine Intelligence, 1–7.

68. Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, L., Pavey, G., et al. (2021). Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms (SSRN).

69. Guide to the General Data Protection Regulation (GDPR). (2019). Information Commissioner's Office. https://ico.org.uk/media/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr-1-0.pdf.

70. Kazim, E., and Koshiyama, A. (2020). The interrelation between data and AI ethics in the context of impact assessments. AI Ethics. https://doi.org/10.1007/s43681-020-00029-w.

71. Data Ethics Canvas (2020). Open data institute. https://theodi.org/article/data-ethics-canvas/.

72. Jordan, S., Fazelpour, S., Koshiyama, A., Kueper, J., DeChant, C., Leong, B., et al. (2019). Creating a Tool to Reproducibly Estimate the Ethical Impact of Artificial Intelligence (AI PULSE Papers).

73. Miller, C., and Coldicutt, R. (2019). People, power and technology: the tech workers' view. Doteveryone. https://doteveryone.org.uk/report/workersview.

74. Khan, A.Q., Khwaja, A.I., and Olken, B.A. (2016). Tax farming redux: experimental evidence on performance pay for tax collectors. Q. J. Econ. *131*, 219–271.

75. Taddeo, M., and Floridi, L. (2018). How AI can be a force for good. Science *361*, 751–752.

76. Mittelstadt, B. (2017). From individual to group privacy in big data analytics. Philos. Technol. *30*, 475–494.

77. Binns, R. (2018). Algorithmic accountability and public reason. Philos. Technol. *31*, 543–556.

78. Lipton, Z., McAuley, J., and Chouldechova, A. (2018). Does mitigating ML's impact disparity require treatment disparity? In Advances in Neural Information Processing Systems (NeurIPS), pp. 8125–8135.

79. Wang, G. (2019). Humans in the Loop: The Design of Interactive AI Systems*64* (J. Artificial Intelligence Res), pp. 243–252.

80. Guidance on the AI auditing framework: Draft guidance for consultation. (2020) Information Commissioner's Office.

81. Kazim, E., Denny, D.M.T., and Koshiyama, A. (2021). AI auditing and impact assessment: according to the UK information commissioner's office. AI and Ethics, 1–10.

82. Examining the Black Box: Tools for assessing algorithmic systems. (2020). Ada-Lovelace Institute & DataKind UK. https://www.adalovelaceinstitute.org/examining-the-black-box-tools-for-assessing-algorithmic-systems/.

83. Kazim, E., and Koshiyama, A. (2020). The interrelation between data and AI ethics in the context of impact assessments (AI Ethics). https://doi.org/10.1007/s43681-020-00029-w.

84. Zekos, G.I. (2021). AI and legal issues. In Economics and Law of Artificial Intelligence (Springer), pp. 401–430.