Philosophy & Technology (2022) 35: 61
https://doi.org/10.1007/s13347-022-00557-9

RESEARCH ARTICLE

The Ethics of AI Ethics. A Constructive Critique

Jan-Christoph Heilinger[1]

Received: 13 January 2022 / Accepted: 18 June 2022 / Published online: 7 July 2022

Abstract

The paper presents an ethical analysis and constructive critique of the current practice of AI ethics. It identifies conceptual substantive and procedural challenges and it outlines strategies to address them. The strategies include countering the hype and understanding AI as ubiquitous infrastructure including neglected issues of ethics and justice such as structural background injustices into the scope of AI ethics and making the procedures and fora of AI ethics more inclusive and better informed with regard to philosophical ethics. These measures integrate the perspective of AI justice into AI ethics, strengthening its capacity to provide comprehensive normative orientation and guidance for the development and use of AI that actually improves human lives and living together.

Keywords Artificial intelligence · Ethics · Structural injustice · Trust · AI justice

People around the globe increasingly encounter, use and benefit from AI in their daily lives in one form or another. AI-based applications range from web-based maps and navigation services to digital behavioural technologies such as mobile health apps, from recommender algorithms in online stores to parking aids, and AI-based services in policing, the legal system, etc. Also, less exciting but tedious tasks—such as analysing immense amounts of data in different domains, or determining the next date for a maintenance check of a machine—can, fortunately, increasingly be done by AI-based systems, relieving humans from burdensome work. Once they are well set up, such AI-based systems work quickly and effectively through vast amounts of data that cannot be handled by humans. And they do so more reliably than humans because machines—unlike humans—are not distracted by fatigue, hunger or the like.

AI figures among the most advanced tools humanity has developed to date, yet its potential for future development remains vast. That is why the *High-Level Expert*

✉ Jan-Christoph Heilinger
jc.heilinger@rwth-aachen.de

[1] RWTH Aachen University, Applied Ethics, Theaterplatz 14, 52062 Aachen, Germany

Group on Artificial Intelligence—an influential advisory body to the European Commission on AI and AI ethics—declared almost euphorically that AI is.

> a promising means to increase human flourishing, thereby enhancing individual and societal well-being and the common good, as well as bringing progress and innovation. In particular, AI systems can help to facilitate the achievement of the UN's Sustainable Development Goals, such as promoting gender balance and tackling climate change, rationalising our use of natural resources, enhancing our health, mobility and production processes [...]. (HLEG 2019)

Such all-encompassing praise indicates the enormous hopes and ambitions, including hope for progress in morally and socially relevant dimensions, that inspire and justify the development of AI-based applications: AI promises to contribute to the good, to just and sustainable human lives and living together on our planet.

The aim of the present paper is to subject such claims and the practice of AI ethics itself, which seeks to provide normative orientation and guidance for the development and use of AI, to a critical analysis. To this end, I proceed as follows. Section 1 will briefly introduce the dominating normative and methodological assumptions, and outline the main topics currently discussed in current ethics of AI. Section 2 will develop critical thoughts about the current state of AI ethics, highlighting conceptual, substantive and procedural challenges that affect the practice of AI ethics. Based on this critical analysis, Sect. 3 will draw constructive conclusions and propose several concrete strategies by which AI ethics could and should develop further.

The normative reflections and proposals made in this paper are based on the ideal of *relational equality*. Relational egalitarians take the equal moral value of all human beings as the normative bedrock for ethical practice in all dimensions of human life and seek to secure conditions under which all can relate to and interact with one another on a footing of equality. Equality, in this relational understanding, is a feature of interactions between people and groups of people that can be influenced by different factors (e.g. economic, political, technological). Relational equality is particularly sensitive to asymmetrical relations of unjustifiably unequal influence and domination that often lead to unequal and unjust distributions of benefits and burdens. Relational egalitarianism aspires to realise more egalitarian distributions where possible, conducive to securing sufficiency for all and the common good (Anderson, 1999; Heilinger, 2020; Lippert-Rasmussen, 2018; Wolff, 1998). Bringing a relational egalitarian account of justice to the normative assessment of AI is meant to expand the focus of AI ethics to pay more attention to the social, economic, political and environmental background structures within which AI is developed and used (cf. D'Ignazio and Klein, 2020). Furthermore, looking at the development and use of AI from this perspective helps also to make pre-existing ethical and social issues visible in a new way.[1] The ultimate goal of the critical analysis offered in the

[1] Cf. Abebe et al. (2020) who argue that computing can act 'as a synecdoche – a part that stands in for the larger whole in discourse and critique. Computing can offer us a tractable focus through which to notice anew, and bring renewed attention to, old problems'.

paper thus is to constructively support ongoing efforts to complement and improve the current practice of AI ethics from the mentioned normative perspectives, so that AI's inevitably massive impact on how humans live and live together will be as good as possible.

1 The Practice of AI Ethics

By now, a number of central ethical issues and problems have been identified and established as 'core questions' arising in the context of AI[2]:

- How can we understand, explain and control—if ever—the internal workings within a complex AI system (cf. e.g. Kempt et al., 2022; Mittelstadt et al., 2019)?
- Who bears responsibility in the event of harm resulting from the use of an AI system (cf. e.g. Santoni de Sio & Mecacci, 2021; Sparrow, 2007)?
- How can AI systems be prevented from reflecting existing discrimination, biases and social injustices based on their training data, thereby exacerbating them (cf. e.g. Bender et al., 2021; Benjamin, 2019; Friedman & Nissenbaum, 1996)?
- How can the privacy of people be protected, given that personal data can be collected and analysed so easily by many (cf. e.g. Véliz, 2020; Zuboff, 2015, 2019)?
- How can autonomous human decision-making be protected against undue influence and deskilling resulting from the use of AI (cf. e.g. Lara & Deckers, 2020; Lu, 2016; Mills, 2020; Reiner & Nagel, 2017)?
- How can it be prevented that an uncontrollable, potentially malicious, superintelligence will, one day, put its own goals above those of humans (cf. e.g. Bostrom, 2014)?

Given the significant economic potential, the far-reaching (geo-) political implications and the social and ethical salience of AI-based technologies, it is not astonishing that next to scholars in the fields of applied ethics and computer sciences, corporations and political institutions have also become involved in the ethical debates. The ethical discourse is thus shaped by contributions from the perspective of the economy, politics and academia.[3] Consequently, the numerous documents that are potentially relevant to the ethical debate about AI are quite heterogeneous and include reports and position papers from policy advisory bodies; guidelines and white papers as self-commitments from companies and research institutions; checklists or questionnaires as handouts for designers; and scholarly publications from all disciplines involved, especially computer science and engineering, management and business studies, sociology and STS and philosophy/(applied) ethics.

[2] For an overview cf. Coeckelbergh (2020), Fjeld et al. (2020)

[3] This is not to deny that others also contribute to the ethical debates, in particular individual users and engaged citizens, civil society initiatives and non-profit organisations, journalism and fictional storytelling in film and literature.

Ethical reflections and arguments in scholarly publications as well as in policy documents and tech industry guidelines[4] often proceed as a discussion of (a) risks and opportunities, or of (b) rules and principles, or of (c) visions and ideals of AI, mirroring the three different normative theories that shape the tradition of Western moral philosophy: consequentialism, deontology and virtue ethics.

The aforementioned ethical questions can thus first be raised and discussed by weighing the hopes and promises against potential risks or problematic, harmful *consequences* resulting from designing and using AI-based technologies (cf. e.g. Coeckelbergh, 2020). Second, as already firmly established in other fields of applied ethics such as bioethics and medical ethics, the ethical discussion can work along the lines of formulating *principles* that are expected to provide orientation regarding what, morally, ought to be done—similar to a catalogue of duties (cf. Jobin et al., 2019). Third, the ethical evaluation can proceed with the formulation of *ideals* and a positive vision to establish goals either for a good, 'virtuous' *use* of AI or even a virtuous *AI itself* (cf. Vallor, 2016). And, of course, the three strategies are not mutually exclusive and, in the different contributions to the ethical debate, can often be found to be employed in parallel. To illustrate this claim with the first substantive question in mind—the one addressing issues of understanding, explaining and controlling AI—the consequentialist discussion focusses on weighing potential risks of an AI that can or cannot be controlled, explained or understood; the deontic discussion would established a principled demand that AI must be explainable and implement measures to secure it as much as possible; and the aretic discussion could either call for virtues in those who develop and use AI or stress the ideal or vision of explainable AI and directly demand that AI-based applications be virtuous, e.g. 'trustworthy', themselves (more on this below, in Sect. 2).

AI ethics thus presents itself as a many-voiced debate involving different perspectives that centres, in quite heterogeneous contributions, around a set of canonical questions and deploys a set of normative tools for ethical analysis and decision-making. Yet, from a normative meta-perspective that scrutinises the current standards and the dominating practices in AI ethics, several critical questions arise. These questions range from conceptual issues about determining the exact topic of the debate, over some basic normative assumptions that underlie it, to the procedures of actual ethical decision-making. In the following, I will identify and briefly discuss—with no claim to completeness—several such normatively relevant meta-questions about AI ethics. The idea is to offer an ethical reflection of the ethics of AI, or, in other words, to work towards *an ethics of the ethics of AI.*

2 AI Ethics: an Ethical Critique

When scrutinising AI ethics in its currently dominating form, three dimensions deserve particular attention because of their theoretical and practical impact on the normative discussions and assessments made: the *conceptual* dimension that defines

[4] For an overview cf. Hagendorff (2020)

the key concepts employed to define the topic of the debate; the *substantive* dimension of determining which ethical question can legitimately be raised within AI ethics; and the *procedural* dimension regarding the methods and practices employed to generate ethical assessments.

2.1 The Conceptual Dimension

The notions deployed to define the topic of AI ethics, on the one hand, shape the normative intuitions with which people approach AI. On the other, they also determine what is considered part of AI ethics—and what not. That is why a conceptual critique provides the grounds for and is an integral element of the normative critique of the practice of AI ethics.

2.1.1 Intelligent? Artificial?

The critical appreciation of the merits and flaws of the current ethical debate about AI starts with a closer look at the concept of AI itself. ‘Artificial intelligence’ is first of all a rather broad technical umbrella term for a number of complex computer technologies that—based on huge amounts of data—are very powerful in pattern recognition, classification, decision-making and prediction and thus can perform tasks in different fields that normally require human intelligence. Among computer scientists, however, the term AI is used less frequently than in public and ethical discourse. Instead, computer scientists usually name more precisely the particular AI method they use, such as machine learning, neural networks or deep learning, each of which constitutes a particular subset of the former (Russel & Norvig, 2021). One of the reasons AI is rather rarely mentioned by computer scientists may be their awareness of the fact that AI—at least in its current forms—is, to pick up Kate Crawford’s provocative formulation, *neither intelligent, nor artificial* (Crawford, 2021, 7).

Certainly, AI can process immense amounts of data and perform highly specialised tasks, but this has little to do with a general creative and cross-sectoral intelligence of humans. A chess computer cannot control drones, a skin cancer detection algorithm can neither translate between different languages nor help me exercise regularly, etc. And at the current state of research, it is not yet foreseeable how—and if ever—a *general*, i.e. non-specialised, form of machine ‘intelligence’ can be developed.[5] Furthermore, far from being merely artificial, AI systems are at the same time highly natural: they rely on numerous natural resources, on raw materials and energy resources extracted from the earth, on countless hours of human labour to

[5] This claim can be true even if *talk of* artificial *general* intelligence can be, on occasion, found among developers and computer scientists, cf. https://openai.com/about/: ‘Our mission is to ensure that artificial general intelligence benefits all of humanity’.

build machines and infrastructure, to develop the programs, to collect and clean data and to train algorithms (cf. Crawford, 2021, chs. 1–2).[6]

Thus, a first basic and conceptual criticism with implications for any appreciation of AI ethics is that *the term AI itself can give rise to misleading intuitions*, which may subsequently also influence AI ethics itself by creating an idea of the technology under consideration that does not match with its reality.

2.1.2 AI as Infrastructure

Next to this terminological issue, an additional conceptual question about the scope of AI ethics deserves attention. AI-based applications can be used in a wide variety of areas, ranging from health and mobility, over communication and consumption, to research, warfare and entertainment. That is why Andrew Ng has proposed to compare AI with electricity.[7] As electricity too, AI is increasingly becoming pervasive and ambient in modern societies—its tasks ranging from 'the general (learning, reasoning, perception, and so on) to the specific, such as playing chess, proving mathematical theorems, writing poetry, driving a car, or diagnosing diseases' (Russel & Norvig, 2021, 19). Thus, AI is becoming a ubiquitous part of infrastructure. Given that AI 'is relevant to any intellectual task; it is truly a universal field' (ibid.), one may even wonder about the need for and benefits of identifying AI as a distinct ethical topic.

Unlike other subfields in applied ethics—such as medical ethics, animal ethics, environmental ethics or the ethics of nuclear technology—the ethics of AI seems to *lack a clearly limited and specific subject area*, insofar as AI is using advanced algorithms, mathematics and statistics and modern computer technologies to address a large range of heterogeneous problems. On this understanding, AI no longer constitutes a narrower subject area, in a similar way as *electricity*-supported strategies to solve problems do not constitute a worthwhile subfield of the ethics of electricity.

This, of course, is a quite provocative claim challenging many contributions to a presumably distinctive AI ethics. But the provocation helps to appreciate the need to assess AI-based technologies as an increasingly important element of contemporary societies, and to embed this assessment in broader debates about ethics and justice. Modern societies, however, are shaped by electricity, etc. and importantly also by AI. Yet, *limiting* the ethical assessment narrowly on the AI-based components itself seems problematic for the reasons just provided: Labelling a field in a misleading way comes with the danger of triggering false expectations, fears and hopes, that risk to lastingly distort an entire debate. And narrowing attention to a ubiquitous tool that is becoming part of everyday infrastructure in modern societies comes with the risk of getting the focus of ethical attention wrong. The ethical questions arising in modern societies are certainly inevitably *influenced* (e.g. mirrored, consolidated,

[6] This claim, however, does not conflict with the claim that AI technologies are also 'artificial', but Crawford wants to direct more attention to the 'natural' underpinnings and enabling conditions of the technology.

[7] https://www.wipo.int/wipo_magazine/en/2019/03/article_0001.html

developed, aggravated, etc.) by AI-based technologies. But focusing narrowly on the AI technologies themselves is insufficient to capture what is morally at stake. Instead, the larger social context, the political, economic, cultural dynamics on both a domestic and the global level need to be seen and scrutinised as the field within which ubiquitous AI exercises its influence. As the following section will further illustrate, the conceptual critique prepares for a substantive expansion of the scope of AI ethics.

2.2 The Substantive Dimension

Two substantive criticisms deserve particular attention and call for complementing and extending the dominating set of questions and debates in AI ethics listed above. The first substantive criticism acknowledges the important fact that AI technologies are infrastructure also in the sense that they require massive amounts of material resources, of human labour, and thus give rise—not only through their computational power but also through their material existence—to numerous ethical issues that are connected to the use of AI. The second criticism challenges from an ethical perspective the tendency to rely on AI-based technologies in the first place to address all kinds of problems.

2.2.1 From the Cloud into the Mines

Kate Crawford and Vladan Joler have undertaken a detailed analysis of what they call the ‘Anatomy of an AI System’ (Crawford & Joler, 2019). The ‘anatomical map’ they sketch shows the massive amounts of human labour, data and planetary resources that are required to develop, build, run and ultimately discard AI-based technologies. Their discussion focuses exemplarily on the life cycle of Amazon’s ‘Echo’-device, the company’s smart speaker running the integrated, voice-controlled intelligent personal assistant service called ‘Alexa’. A visually appealing map of the ‘anatomy’ of this AI system debunks the myth of AI as a solely cloud-based, bodiless and thus clean form of intelligence and instead brings to light how the small and elegant Echo-device, offering impressive and convenient cloud-based services, actually presupposes and hides a massive and very down-to-earth background of resource extraction, energy consumption, waste production, exploitative labour, power imbalances and benefit accumulation among some.

The exemplary analysis of one device has shown the ethically weighty fact that AI-based technologies are demanding large amounts of different resources, a fact largely neglected in the AI-ethical literature. Awareness for the environmental and human costs of these systems is increasingly receiving attention in the AI-ethical literature, e.g. with regard to the energy needed for training neural networks (Bender et al., 2021; Strubell et al., 2019); when showing that AI-based technologies cannot only be used to develop strategies for sustainable cities, farming etc. but are in themselves challenges for sustainability (Brevini, 2020; van Wynsberghe, 2021); or when providing a full ‘Atlas of AI’ that exposes the power-relationships, political and economic and planetary dynamics and costs of AI systems generally (Crawford, 2021).

To provide two more details: First, the financial gains made from AI-based systems are distributed in an extremely unequal way as comparing the annual financial gains of, say, Amazon's CEO on the one extreme end and the underpaid labourers working under dangerous and exploitative conditions in the Cobalt mines on the other extreme end, clearly shows.[8] And second, the lasting harm for humans and the environment that results from toxic waste that remains once the devises are not in use any more.[9] Both details are issues of distributive injustice that, from the perspective of relational egalitarianism, is a consequence of existing asymmetries in power and influence that allow some to dominate others.

Some may object that these issues are nothing but the normal *background* which may be regrettable or not, but not directly relevant for a discussion of the ethics of AI technologies. Similarly in medical ethics, the objection could continue, where concern for the environmental costs of running hospitals or the unequal income distribution between the different groups that contribute to the functioning of a health system also does not figure in our engagement with topics of medical ethics. Thus, focussing on the *distinctive* issues that arise through the very use of AI-based applications should be at the centre of AI ethics and nothing else.

A reply to this objection can be twofold. First, medical ethics has in recent years started to direct more attention to the connections between providing health services to patients, e.g. in hospitals, on the one hand, and its environmental impact, e.g. on the climate and on population health, on the other. Attention to context and background is thus emerging in other fields, as well. And second, it is a longstanding mistake to assume that an ethical assessment could limit itself to what is done by some (developers, users, etc.) while ignoring the influence of presumably normal and thus acceptable background conditions, that, at closer reflection, may turn out to be morally unjustifiable and indicative of structural injustice (Young, 2006, 120–121; McKeown, 2021a, 2021b). An adequate understanding of the ethical issues thus requires assessing the social structures within which action takes place, because such social structures can make the positive moral valence of apparently innocent, well-intended and unproblematic actions shift.[10] Only by expanding the focus of moral attention—which will lead to embedding but not to replacing the established questions of AI ethics mentioned above—can an adequate ethical assessment be provided.

[8] The countless fully *unpaid* users who, through interacting with the Echo-device, provide important training opportunities that help develop and improve the product further, matter too. The massive differences in financial gains achieved by the different contributors to the entire infrastructure necessary for the device to function are not only an instance of distributive injustice (where adequate compensation would be required for all those who contribute to the entire cooperative system), but indicative of a more basic relational injustice that can be captured in terms of domination and intentional clustering of advantages for some and of disadvantages for others (Rawls 1999; Young 1990; Young 2011).

[9] Here again, harm caused to the environment is not a regrettable and unintended effect of the business practice; instead, externalising costs are an essential element of accumulating benefits in growth- and profit-oriented corporations (Hickel 2020).

[10] Insofar as AI ethics integrates these issues, it can very well be called 'critical theory', as suggested by Waelen (2022).

A global infrastructure lies in the background of devices running AI-based applications, an infrastructure marked by massive, structural injustice: Persistently and systemically, the advantages and disadvantages are unequally distributed between different groups of actors, as analysis of the production and supply chain, data extraction, disposal and the unequal distribution of power, influence and profits makes abundantly clear. Thus, the consequences and risks of developing and using AI go far beyond the issues discussed in the dominating debates that thus need to be expanded through an inclusion of the voices of all who are, in one way or another, affected by the development, use and disposal of such technologies.

2.2.2 Technosolutionism

A second substantive challenge for the ethics of AI arises from a very fundamental question: Prior to securing the ethical use of AI to address a problem, it needs to be asked and re-considered whether AI is indeed a suitable means to address a problem in the first place (cf. Riley, 2008, ch. 2; Morozov, 2013). The grounds for this challenge lie in the fact that in some cases AI-based technologies are deployed to address problems that could be better addressed by non AI-based interventions. A general preference for prioritising technological over non-technological solutions may lead to choosing an ultimately unsuitable strategy. The worldview—some may say: the ideology—to address diverse human problems *primarily* with the help of technology and engineering, including problems in the fields of politics and society, education, public health and law, can be called 'technological solutionism' (Morozov, 2013).

Technosolutionism is a wide-spread phenomenon, in no way limited to the deployment of AI-based technologies. Examples thus go beyond apps to fix social problems (e.g. an app to secure consent before engaging in sexual activity) and include the medicalisation of social problems (e.g. administering drugs to pupils with ADHD), or geo-engineering such as solar radiation management to reduce the Earth's albedo in order to reduce global warming that results from environmental pollution, etc.[11]

Two related dynamics strengthen the inclination to rely on technological solutions to all kinds of problems. First is a dynamic resulting from the sheer existence of a powerful tool that makes humans want to use it. Holding a hammer in one's hands makes one approach the world in a specific way: the tool itself creates an invitation to use it. The same holds true for advanced medical, technological or other tools for which novel uses are thought up simply as the result of their availability. In such cases, the availability of certain technologies can lead to an incentive to 'reverse engineer' applications for which the technologies would serve as solutions. And

[11] The problem of technosolutionism can be discussed in terms of relational injustice: Examples abound that can show how a dominating group and perspective deploys its preferred (technological) tools to address a problem that, from the perspective of those most affected, could or should be tackled by different means. The fear that sexual self-determination and evolving, ongoing consent or dissent, for example, will not be respected cannot be appeased by an app that, instead, will strengthen the position of a potential perpetrator who can claim that consent was given.

second, modern societies are characterised by a wide-ranging general openness for addressing problems of different kinds through engineering and digital approaches. Digital technologies are often perceived as progressive, effective, modern, fascinating and, because of their technological nature, as politically neutral[12]—and thus are frequently perceived as preferable over alternative, old-fashioned, ideological interventions to address problems. Yet, the preference for technological solutions tends to create a dynamic to narrow down the perception and the analytical understanding of problems to their technological dimension.

The substantive point of criticising technosolutionism, however, is not to deny that digital solutions can be effective and desirable. It stands beyond doubt that AI-based and other advanced technologies do contribute in meaningful ways to addressing many problems and that using them can provide better outcomes, relieve humans from tiresome work and can also help address social, environment, medical and other challenges. Some problems indeed have technological solutions and others can be addressed at least partly with the help of technologies. The substantive point of the criticism is rather that the ethics of AI in particular (as the ethics of technology in general) has to broaden its focus and must not only inquire whether the use of a particular technology is ethical or how it can be made more ethical. Instead, an *ethical* AI ethics will have to also have to settle whether a particular problem ought, morally, to be conceived of as a technological problem and consequently be addressed with the help of (AI-based) technology in the first place. This is because problems of a non-technological nature may call for solutions that lie (at least primarily) outside of the realm of technology: a social or political problem will be better addressed through low-tech or even no-tech solutions, even when an AI-based high-tech strategy could also contribute to addressing some of their aspects.

An ethical AI ethics will thus carefully seek a comprehensive understanding of perceived problems that does not, from the outset, self-limit its attention to the problems' technological side. Avoiding the technosolutionist mindset means to remain open for a problem's social, political, economic and other dimensions and looking for solutions in these domains, as well.

Importantly, the need to avoid a technosolutionist mindset applies also to many problems that become obvious or aggravated through the use of technology. Think, for example, of 'algorithmic bias' when a computer programme generates unfair outcomes that systematically advantage or disadvantage specific social groups (e.g. Buolamwini & Gebru, 2018). An appropriate solution to this problem must not limit itself to a technological solution, because algorithmic bias is not so much a problem of technology but primarily a symptom and thus indicative of a fundamental social problem.[13] Even if *algorithmic* bias will be reduced or eliminated through improved programming and better data input, the problem of *actual* bias, discrimination and

[12] But cf. Winner's seminal paper on the politics of artifacts (Winner 1980).

[13] This holds true also for the outcomes of recommender algorithms, e.g. in social media presenting increasingly radical and extreme content to its users in order to generate attention. Here again, this is no automatism of the algorithm, but reflective of the interest of the service providers to keep people engaged. The providers prioritise their own commercial interests over the social interest.

unfairness in our societies would persist. Actual solutions will only be found on the social level, while technological solutions will even in the best case provide nothing but cosmetic improvements at the symptom level, leaving the origins, structures and underlying social dynamics and structural injustices untouched.

That is why an additional, important risk connected to the pursuit of technological solutions to social problems lies in the potentially very high opportunity costs. If scarce resources (finances, attention, human ingenuity, etc.) are primarily directed towards technological solutions, these very resources cannot be deployed for other strategies to address the respective problems. Thus each decision to 'techno-solve' any problem bears a burden of proof to show that technology does indeed provide the most promising strategy to address the problem.

2.3 The Procedural Dimension

The ethics of AI deserves scrutiny also with regard to its own (institutionalised) practice, its standards and its methods. The main criticism to be raised here is that many committees or fora to deliberate and determine ethical standards for AI-based technology frequently fail, for systemic reasons, to meet the standards of independent, ethically informed, critical deliberation and decision-making and consequently only provide distorted ethical assessments. In the following, I discuss this issue with a particular focus on the European AI landscape, but related arguments about the influence of particular, strategic interests can also be made for other world areas.

2.3.1 Identifying Experts

It is hardly astonishing that large multi-national corporations, standing in harsh competition with one another regarding market shares, growth and profits, have only a limited and rather specific interest in an independent, critical ethical assessment of their practices. Corporate ethics units, in varying degrees, serve different purposes than providing actual ethics assessments, among them: reducing the likelihood of legal and ethical scandals that would create reputational damage; increasing awareness for and compliance with existing moral and legal standards and regulations among the company's employees; showcasing ethical awareness to meet societal expectations or for reputational gains; and reducing the need for further external, political interventions and regulations because of ethical self-commitment and self-control. To this end, many companies have established corporate ethics boards, offer ethics trainings to their employees or fund ethics research relevant to the company's interests inside or outside of the company.[14] Corporate interest in ethics recently even seems to have started to cool down, now that it is increasingly becoming clear how independent ethics raises issues that might conflict with, even run directly

[14] A recent example from Germany is the Facebook-funded chair for Ethics in Artificial Intelligence at the Technical University Munich.

contrary to the primarily financial interests of corporations.[15] Corporate ethics thus has to walk the thin line of providing, as much as possible, independent and critical ethics assessments, while being and remaining on the payroll of the corporation.

Whether or not such dependency makes independent assessment principally impossible cannot be discussed here. Instead, I want to scrutinise another type of ethics committee, namely an ethics expert advisory group to support political decision-making.

Above, I have already quoted the enthusiastic endorsement of AI from the 'Ethics Guidelines for Trustworthy AI', a document published in April 2019 by the High-Level Expert Group on Artificial Intelligence (preceding the publication of 'Policy and Investment Recommendations for Trustworthy AI' in June 2019). This group was established by the European Commission to provide independent ethical advice on the development and use of AI-based technologies and provide a case in point.

In international comparison, the policy document is exemplary. It identifies, at the behest of a political body, numerous ethics issues arising in the context of AI and calls for a systemic approach to address them by assigning responsibility to 'all actors and processes that are part of the [AI-] system's socio-technical context throughout its entire life cycle' (HLEG 2019, 5). With this, the European demand for the development and implementation of 'trustworthy AI' needs to be compared and contrasted at the global level in particular to the US-American digital ambitions to develop AI as a driver of 'disruptive innovation' (cf. e.g. Girasa, 2020) and the Chinese development and use of AI as controlled by and in the interest of the state (cf. e.g. Roberts et al., 2021). But looking at the genesis and content of the European strategy from an ethics perspective exposes both an institutional-procedural problem and a resulting normative-conceptual problem.

The Independent High-Level Expert Group comprised of more than fifty members, mostly from industry with a tiny fraction of ethics experts, only some of which were philosophers.[16] As important as the involvement of the industry in regulatory processes from the very beginning may be, in order to secure acceptance and compliance in the future, such an unbalanced composition of the group—largely excluding also representatives of users and civil society—clearly conflicts with the ambition to provide independent and comprehensive ethical expertise. Yet, the composition of the expert group does not seem coincidental but fully in line with the political preferences of the European Commission and its interest to advance the standing and business opportunities of European corporations on the global market. Given this ambition, however, the primary interest in ethics will hardly have to establish sound ethical standards and regulations for the development and use of AI-based technologies. For this, a larger share of trained ethicists, sociologists and civil and human rights advocates would have been an obvious imperative. Instead, the interest in ethics was, so it seems, motivated by a perceived need to unburden AI-based technologies from its partly uncanny and unethical appearance. In other

[15] Cf. the debate about the role of Timnit Gebru at Google and the circumstances of the termination of her contract, cf. Tiku (2020), Ghaffary (2021).

[16] See list on https://www.aepd.es/sites/default/files/2019-12/ai-definition.pdf.

words, the purpose of this political advisory group might have been to provide 'ethics washing' (Metzinger, 2019).

'Ethics washing'—or 'ethics theatre'—is to engage in ethical discussions and the formulation of ethical self-commitments with the intention to prevent or at least delay effective legal regulations. This interest can primarily be found in the industry, for which regulations are often costly. To this end, AI-developing companies are establishing internal ethics units or funding research activities in AI ethics in public universities. In the present case, the industry's interest was advanced and supported even with the help of the European Commission.

Given the dominance of industry and industry-friendly politics in the field of AI ethics, many resulting ethical guidelines for AI tend to insufficiently include independent (academic) ethics expertise. Even if the European guideline rightly directs ethical attention towards numerous ethics issues, provides important ethical orientation and may be more advanced and concerned with the common good than existing policy documents from other countries or world regions,[17] the described lopsidedness in the group's composition contributes to severe conceptual distortions that turn out to be, from a normative perspective, highly problematic. I will illustrate this claim with regard to the guideline's core concept, 'trustworthy' AI.

2.3.2 Neglecting Expertise and Trusting AI

The document's guiding ethical idea is the ideal of 'trustworthy AI'. Trustworthiness of an AI system is defined as comprising three components: AI 'should be *lawful*, complying with all applicable laws and regulations; it should be *ethical*, ensuring adherence to ethical principles and values; and it should be *robust*, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm' (HLEG 2019, 6). For the authors, the entity worthy of trust thus is primarily the AI system itself and its 'inherent properties', even though it is admitted that also 'the humans behind' AI systems and 'the qualities of the socio-technical systems involving AI systems' are supposed to be worthy of trust—in analogy to human trust in aviation, nuclear power or food safety. But it makes all of a difference to trust in something or to trust in the people and institutions behind something.

The ethics guidelines gloss over this difference and label something, an AI system, as trustworthy that for principled (conceptual) reasons to be explained below cannot and should not be considered worthy of trust (Bryson, 2018). ng so amounts to, as Metzinger, one of the few ethics expert members of the group, critically remarked, a brilliant marketing coup lulling appropriate ethical concern and necessary critical analysis with the help of a euphonic epithet (Metzinger, 2019).

Here is, briefly, the conceptual argument against the possibility of trustworthy AI that deploys the fundamental distinction between trust and reliance also in the realm of AI. Practical reliance is a 'common core' of trust (Hawley, 2019, 2), and the informed rational expectation that people will do certain things is part of

[17] For an overview cf. Jobin et al. (2019).

trusting them. But while trust *involves* reliance, following Baier's seminal contributions, there is wide agreement in philosophy that trust *exceeds* mere reliance (Baier, 1986).

What now is the distinctive feature missing in cases of mere reliance and present in cases of trust? It is that it goes beyond the informed and rational expectation that someone will act in certain ways and includes a *normative* and an *affective* component (Ryan, 2020). The normative component in an interpersonal relationship of trust is that the trusting person assumes the trusted person will have certain normative reasons, i.e. *should* act as expected (Simpson, 2012). Yet, regrettably, even the best normative reasons do not always lead to corresponding action. Trust includes the expectation that the trusted person will indeed act upon them, but leaves open the possibility that she might also betray the trusting person. Correspondingly, a particular affective dimension exists in the relationship of trust that is absent when merely relying on something or someone: If the trusted person fails to act upon the normative reasons, the reactive attitude on the side of the trusting person goes beyond mere disappointment (which would also be appropriate in case a computer malfunctions, or the bad weather spoils my plans for the week-end). Beyond disappointment, the betraying person, because of her agential features that allow her to act upon reasons, will become an appropriate target for moral criticism or blame.

AI cannot act by following reasons, it cannot explain the outcomes of its decisions by giving reasons, and it is not an appropriate target for moral critique or blame. AI is a sophisticated tool, not a moral agent (Véliz, 2021), on which we may rely, but which we must—for conceptual reasons—not trust. Talking about trustworthy AI seems to elevate AI systems to the status of moral agents, which is not only factually wrong, but also morally dubious, because it hides or reduces the moral responsibility of the appropriate potential targets of trust: those who develop and implement AI. These persons and institutions, however, have—for prudential reasons—to *earn* trust before it becomes rational for anyone to trust them.

Talk of trustworthy AI anthropomorphises technology, obscures agency and reduces the perceived urgency of the need to track responsibility and set boundaries for the development and implementation of AI systems. This misconception of AI as potential bearer of trust—uncritically taken up in countless documents about AI ethics—can be explained in at least two different ways: as a result of an intentional strategy to label something 'problematic with a positive attribute; or as the regrettable outcome of the confusion of considering AI as actually 'intelligent' in a sense comparable to human intelligence and free will. *Honi soit qui mal y pense?* Given the perfect alignment of this conceptual confusion with the commercial interests of the AI-developing and AI-using corporation to delay and avoid regulation, suspicion may be in order and support the diagnosis of intentional ethics washing through the European Commission's High-Level Expert Group. For Metzinger the debate is set. The entire 'Trustworthy AI narrative is, in reality, about developing future markets and using ethics debates as elegant public decorations for a large-scale investment strategy' (Metzinger, 2019). In any case, even a more moderate reading vindicates a morally highly problematic conclusion: AI ethics is strongly influenced by various political and economic interests that put ethics under the risk of being instrumentalised.

3 Steps Towards a More Ethical AI Ethics

The preceding pages have, from a normative perspective, identified six challenges for AI ethics: misleading intuitions resulting from the notion ‘artificial intelligence’; a misconception of the topic of AI ethics by failing to recognise that AI is ubiquitous and part of modern societies’ infrastructure; an insufficient consideration of non-AI issues lurking in the background of using AI-based technologies; the persistence of a technosolutionist mindset inclined to address all sorts of problems (tech and non-tech) with the help of AI and other high-tech solutions; a strong and distortive influence of particular (e.g. commercial, political) interests on the overall AI-ethics debate; and the insufficient integration of independent and critical philosophical and ethical expertise.

These issues—labelled as two conceptual, two substantive and two procedural issues—are *ethical* issues, insofar as they point out ethically relevant shortcomings in the existing debate. Making AI ethics more ethical thus requires to address these issues. In the following section, I will take up the preceding critique to outline potential directions and steps towards an improved, ethical AI ethics. Insofar as my proposal broadens the scope of AI ethics and invites increased consideration of the social, economic, political and environmental background structures within which novel AI systems are being developed and used, it can be understood as an argument for embedding AI ethics into AI justice.[18]

3.1 Conceptual Steps

In order to make AI ethics more sure-footed and avoid distortions that result from conceptual unclarity, an explicit conceptual specification of the subject of AI ethics will be necessary. Of course, those working in the field are aware of the challenges mentioned above, but the ethical discourse that involves different types of players and is accompanied by significant public and lay interest can only gain through such clarifications. Being outspoken about the actual capacities and limitations of the different technologies and being realistic in the assessment of future developments—that include also the increasing likelihood of the next, upcoming AI winter (Floridi, 2020)—will provide a much needed factual basis for addressing the pressing ethical issues at hand. In this context, it is also important not to limit the ethical attention to the specifics of AI-based technologies, but acknowledge that AI is becoming part of the ambient infrastructure in modern societies. Such a refocussing will thus include a dual movement of conceptual narrowing (leading e.g. to reduced interest in a presumed singularity or to decreased fear of an evil super-intelligence) and conceptual broadening (leading to analysing AI even more *in its social context*).

Taking such conceptual clarifications seriously would impact in particular on the quality of the ethical discussions. It will contribute to a much needed toning down

[18] In particular, the substantive and procedural steps are committed the ideal of relational equality (see above).

in a partly overexcited debate about a ‘hyped’ topic.[19] A more composed attitude in parts of the debate would be desirable, also in order to support more appropriate decisions in different fields, such as funding decisions for AI research and AI-ethics research. The amount of (public and private) funding made available for AI ethics seems out of proportion, given that other ethical challenges—the climate, the pandemic, but also different questions from the field of (global) social justice, medical ethics, animal ethics, etc.—urgently call for ethical attention, too.

3.2 Substantive Steps

The mentioned challenges also come with substantive implications affecting the type of questions that should be discussed within the field of AI ethics. First, AI ethics should, as a matter of justice, expand its focus beyond the development and use of AI-based applications to include the structural background conditions required for such development and use. Losing sight of the social, cultural, economic and political environment when focussing narrowly on the moral quality of specific acts is a mistake (that, however, can also occur in moral and political philosophy and other fields of applied ethics). In this case, one simply assumes—but unjustifiably so—that the background would constitute something like the normal, and morally uncontroversial baseline that can be excluded from the ethical analysis that concentrates on what some agent is ng here and now. Instead, the unequal and asymmetric relations of privilege, power and influence deserve full attention when it comes to providing normative guidance about the practical development and use of specific technologies.

If more attention is being directed towards the structural context of a respective specific act or decision, the entire supply chain for, the ongoing (natural and human) resource use of and the distribution of different types of costs and benefits arising throughout the entire life cycle of a technology will become part of the moral situation under consideration. Attention will also be directed towards those who are affected by the development and use of novel technologies in an indirect way, e.g. by being excluded from its use and the connected benefits. As a starting point for addressing the structural dimension of the respective technologies could be to ask: Who benefits from the novel development? How? At which costs that are born by whom?[20]

While, of course, not every discussion always has to include the full structural background, increasing awareness for the conditions under which the development, production and use of AI-based technologies occurs would be an important extension and correction of a debate that too often excludes it.

The second substantive implication of the challenges discussed above is to avoid narrowing down, even within the field of AI ethics, the search for solutions to *technological* strategies. In every case where an advanced technology

[19] Cf. e.g. the *Better Images of AI* initiative collecting illustrations of AI that avoid clichéd and misleading visualisations (https://betterimagesofai.org/).

[20] For a longer list of ethically relevant ‘who’-questions, cf. D’Ignazio and Klein (2020), in particular p. 27.

could provide a solution to a problem that is seen as worthy of attention, one has to ask whether a low-tech or even a no-tech solution would might be available and maybe even preferable. When taking into consideration all costs incurred by AI-based technologies (including negative externalities that may only further increase structurally unjust background conditions), low- or no-tech solutions may turn out to be, from a normative perspective, the preferable option.

3.3 Procedural Steps

The challenges discussed above also carry procedural implications for how AI ethics should be practiced. Here again, two demands follow. First, as a matter of justice, it is essential to include in ethical debates worthy of this name the voices of all affected. Limiting participation in and contributions to the debate to the already influential economic or political agents in the field is a severe ethical flaw. From the exploration and development of novel ideas over the discussion of regulatory frameworks to the assessment of the actual, practical outcome of technological changes in societies and communities, the voices of all affected need to be heard. Community-oriented 'design ethics' and 'design justice' provide already important experience and guidance for future improvements in this direction (Costanza-Chock, 2020): Political bodies should rely more on citizen advisory groups, and also corporations should be more inclusive in considering the social impact of their products. Any ethical and responsible development of novel, AI-based technologies and applications cannot occur in corporations alone. Ultimately, the willingness to make genuine efforts to build socially beneficial technologies (instead of building profitable products advancing the financial interests and growth of the company) requires a cultural and attitudinal development and change in much of the established corporate practice, and thus has to be considered as a long term project. However, if AI is really to advance the common good, the pursuit of this project is worthwhile and contributing to it will be an integral element of AI ethics.

A second procedural implication of the mentioned challenges consists in a call to include the best available ethical experiences and knowledge into the AI ethics debates that are, so far, very rarely dominated by trained ethicists. The border between technical and ethical issues cannot be drawn in a way that would allow a functional separation; ethics and technical issues have to be considered jointly from the outset. Also the conceptual tools to discuss ethical challenges and ideals must not be used uncritically lest they generate more confusion than provide orientation. Two parallel strategies can be recommended here: one consists in seeing to it that trained philosophical ethicists learn to (better) communicate with practitioners and policy makers in the field and will be included more and heard better in the relevant debates; the other consists in training and educating not only ethicists but also provide more substantive training about ethics and justice to all those working in the computer sciences, engineering and in politics (cf. Riley, 2008).

4 Conclusion: the Ethics of AI Ethics

AI is quickly evolving and will continue to impact human lives and living together in the future, calling for normative reflection and guidance. This paper has identified conceptual, substantive and procedural challenges for the current practice of AI ethics: a misconception of the topic of AI ethics that fails to recognise AI as ubiquitous and part of modern societies' infrastructure; an insufficient consideration of non-AI issues lurking in the background of AI-based technologies; the persistence of a technosolutionist mindset inclined to address all sorts of problems (tech and non-tech) with the help of AI and other high-tech solutions; a strong and distortive influence of particular (e.g. commercial, political) interests on the overall AI-ethics debate; and the insufficient integration of independent and critical philosophical and ethical expertise. Reflecting, from a normative perspective, upon the practice of AI ethics, the paper outlined recommendations for an ethical AI ethics to address the identified challenges through terminological reform; through a broadening of scope to better include the dimensions of relational ethics and social justice in the context of AI; and through more inclusive deliberative procedures that better represent all affected and better integrate available philosophical and ethical knowledge.

While further work is necessary to substantiate and detail the proposals, the main conclusion of the argument is the following: In order to use AI and AI-based technologies for the common good and to actually improve human lives and the living together of humans generally—and not only for some, at the expense of others—more ethical reflection and specific concern for issues of structural (in-) justice is imperative. A failure to advance in the directions outlined above will condemn AI ethics to provide only an impoverished ethical analysis, which perpetuates or even aggravates existing injustices. Given the expected impact of AI on our lives, it is of prime concern to deploy the best possible, ethical AI ethics to steer the use of AI-based technologies towards the common good.

Funding Open Access funding enabled and organized by Projekt DEAL. The research was supported by the German Federal Ministry of Health (Bundesministerium für Gesundheit, BMG) within the project KIPeriOP (grant number 2520DAT10G).

Declarations

Ethics Approval and Consent to Participate Not applicable (philosophical desk research).

Consent for Publication Not applicable.

Competing Interests The author declares no competing interests.

References

Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., and Robinson, D. G. (2020). Roles for computing in social change. In, Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. *Association for Computing Machinery, Barcelona, Spain*, 252–60. https://doi.org/10.1145/3351095.3372871

Anderson, E. (1999). What Is the Point of Equality? *Ethics, 109*(2), 287–337.

Baier, A. C. (1986). Trust and Antitrust. *Ethics, 96*, 231–260.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). 'On the dangers of stochastic parrots: Can language models be too big?'. *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* 610–23. https://doi.org/10.1145/3442188.3445922

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new jim code*. Polity.

Bostrom, N. (2014). Superintelligence. Paths, Dangers, Strategies. Oxford/New York: Oxford University Press.

Brevini, B. (2020). Black boxes, not green: Mythologizing artificial intelligence and omitting the environment. *Big Data & Society, 7*(2), 1–5.

Bryson, J. (2018). *AI & Global Governance: No One Should Trust AI*. United Nations University.

Buolamwini, J., and Gebru, T. (2018). 'Gender shades: Intersectional accuracy disparities in commercial gender classification'. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency 81*.

Coeckelbergh, M. (2020). *AI Ethics*. MIT Press.

Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need*. MIT Press.

Crawford, K., and Joler, V. (2019). 'Anatomy of an AI system'. https://anatomyof.ai.

Crawford, K. (2021). Atlas of AI. Power, politics, and the planetary costs of artificial intelligence. New Haven/London: Yale University Press.

D'Ignazio, C., & Klein, L. F. (2020). *Data Feminism*. MIT Press.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. C., and Srikumar, M. (2020). 'Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI'. *Berkman Klein Center Research Publication No. 2020-1*. https://doi.org/10.2139/ssrn.3518482

Floridi, L. (2020). AI and its new winter: From myths to realities. *Philosophy & Technology, 33*(1), 1–3. https://doi.org/10.1007/s13347-020-00396-6

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems, 14*(3), 330–347.

Ghaffary, S. (2021). 'Google says it's committed to ethical AI research. Its ethical AI team isn't so sure'. *Vox* June 2.

Girasa, R. (2020). *Artificial intelligence as a disruptive technology. Economic Transformation and Government Regulation*. Palgrave.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines, 30*(1), 99–120.

Hawley, K. (2019). *How to be Trustworthy?* Oxford University Press.

Heilinger, J.-C. (2020). Cosmopolitan Responsibility. Global Injustice, Relational Equality, and Individual Agency, https://doi.org/10.1515/9783110612271. Berlin/Boston: de Gruyter.

Hickel, J. (2020). *Less is More: How Degrowth Will Save the World*. Penguin/Windmill.

HLEG (2019). High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. Brussels: European Commission.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kempt, H., Heilinger, J.-C., Nagel, S.K. (2022). "I'm afraid I can't let you do that, Doctor" – Meaningful Disagreements with AI in Medical Contexts'. *AI & Society*. https://doi.org/10.1007/s00146-022-01418-x

Lara, F., & Deckers, J. (2020). Artificial Intelligence as a Socratic Assistant for Moral Enhancement. *Neuroethics, 13*(3), 275–287. https://doi.org/10.1007/s12152-019-09401-y

Lippert-Rasmussen, K. (2018). *Relational Egalitarianism*. Cambridge University Press.

Lu, J. (2016). Will Medical Technology Deskill Doctors? *International Education Studies, 9*(7), 130–134.

McKeown, M. (2021). Structural injustice. *Philosophy Compass, 16*(7), e12757. https://doi.org/10.1111/phc3.12757

McKeown, M. (2021). Geist aus der Flasche. Ist der Kampf um einen ethischen Einsatz Künstlicher Intelligenz schon verloren? *Forschung & Lehre, 7*, 548–549.

Metzinger, T. (2019). 'Ethics washing made in Europe'. *Der Tagesspiegel* 8 April 2019 (https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html).

Mills, S. (2020). Personalized nudging. *Behavioural Public Policy, 6*(1), 150–159. https://doi.org/10.1017/bpp.2020.7

Mittelstadt, B., Russell, C., and Wachter, S. (2019). 'Explaining Explanations in AI'. FAT'19: Conference on Fairness, Accountability, and Transparency (FAT'19), January 29–31, 2019, Atlanta, GA, USA (https://doi.org/10.1145/3287560.3287574).

Morozov, E. (2013). To save everything, click here. The folly of technological solutionism. New York: Perseus.

Rawls, J. (1999). *A Theory of Justice* (Revised). Harvard University Press.

Reiner, P. B., and Nagel, S. K. (2017). 'Technologies of the extended mind: Defining the issues'. in Illes, J. (ed.), *Neuroethics. Anticipating the future*. Oxford/New York: Oxford University Press. 108–22.

Riley, D. (2008). *Engineering and Social Justice*. San Rafael, CA: Morgan & Claypool.

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2021). The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. *AI & Society, 36*(1), 59–77.

Russel, S. J., & Norvig, P. (2021). *Artificial Intelligence. A Modern Approach (Fourth Edition* (Global). Pearson.

Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics, 26*(5), 2749–2767. https://doi.org/10.1007/s11948-020-00228-y

Santoni de Sio, F., and Mecacci, G. (2021). 'Four responsibility gaps with artificial intelligence: Why they matter and how to address them'. *Philosophy & Technology 34*, 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Simpson, T. W. (2012). What is trust? *Pacific Philosophical Quarterly, 93*(4), 550–569.

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy, 24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Strubell, E., Ganesh, A., and McCallum, A. (2019). 'Energy and Policy Considerations for Deep Learning in NLP'. *arXiv*:1906.02243.

Tiku, N. (2020). 'Google hired Timnit Gebru to be an outspoken critic of unethical AI. Then she was fired for it.'. *The Washington Post* 23 December.

Vallor, S. (2016). *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press.

Véliz, C. (2020). *Privacy is Power. Why and How You Should Take Back Control of Your Data*. Bantam.

Véliz, C. (2021). Moral zombies: Why algorithms are not moral agents. *AI & Society, 36*(2), 487–497.

Waelen, R. (2022). Why AI Ethics Is a Critical Theory. *Philosophy & Technology, 35*(1), 9.

Winner, L. (1980). Do Artifacts Have Politics? *Daedalus, 109*(1), 121–136.

Wolff, J. (1998). Fairness, Respect, and the Egalitarian Ethos. *Philosophy & Public Affairs, 27*(2), 97–122. https://doi.org/10.2307/2672834

van Wynsberghe, A. (2021). 'Sustainable AI: AI for sustainability and the sustainability of AI'. *AI and Ethics 1*, 213–218. https://doi.org/10.1007/s43681-021-00043-6,10.1007/s43681-021-00043-6

Young, I. M. (1990). Five Faces of Oppression. In I. M. Young (Ed.), *Justice and the Politics of Difference* (pp. 39–65). Princeton.

Young, I. M. (2006). Responsibility and Global Justice: A Social Connection Model. *Social Philosophy and Policy, 23*, 102–130.

Young, I. M. (2011). *Responsibility for Justice*. Oxford University Press.

Zuboff, S. (2015). Big other: Surveillance Capitalism and the Prospects of an Information Civilization. *Journal of Information Technology, 30*(1), 75–89. https://doi.org/10.1057/jit.2015.5

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Public Affairs.