AI and Ethics
https://doi.org/10.1007/s43681-021-00122-8

COMMENTARY

Blind spots in AI ethics

Thilo Hagendorff[1]

Received: 11 August 2021 / Accepted: 25 November 2021

Abstract
This paper critically discusses blind spots in AI ethics. AI ethics discourses typically stick to a certain set of topics concerning principles evolving mainly around explainability, fairness, and privacy. All these principles can be framed in a way that enables their operationalization by technical means. However, this requires stripping down the multidimensionality of very complex social constructs to something that is idealized, measurable, and calculable. Consequently, rather conservative, mainstream notions of the mentioned principles are conveyed, whereas critical research, alternative perspectives, and non-ideal approaches are largely neglected. Hence, one part of the paper considers specific blind spots regarding the very topics AI ethics focusses on. The other part, then, critically discusses blind spots regarding to topics that hold significant ethical importance but are hardly or not discussed at all in AI ethics. Here, the paper focuses on negative externalities of AI systems, exemplarily discussing the casualization of clickwork, AI ethics' strict anthropocentrism, and AI's environmental impact. Ultimately, the paper is intended to be a critical commentary on the ongoing development of the field of AI ethics. It makes the case for a rediscovery of the strength of ethics in the AI field, namely its sensitivity to suffering and harms that are caused by and connected to AI technologies.

Keywords AI ethics · Artificial intelligence · Fairness · Privacy · Explainability · Externalities

1 Introduction

In 2016, *Nature* published an article by Crawford and Calo, entitled "There is a blind spot in AI ethics" [1]. The researchers argued for a more fine-grained technology assessment of AI systems, instead of repeatedly discussing the same small set of either obvious, futuristic, or very abstract ethical issues. The statement on blind spots in AI ethics is half a decade old, and claims about the field of AI ethics per se are rather generalizing. However, AI ethics discourses have changed significantly along the way, with new theories, frameworks, differentiations, and guidelines being developed. Nevertheless, in this paper, I once again want to make the case that current AI ethics has severe blind spots and that, again in the words of Crawford, this time from her 2021 book *Atlas of AI*, a "new era of critique" [2] is much-needed. This new era of critique must embrace a more "radical" ethical approach. Being radical, stemming from Latin "rādīx" ("root"), nudges to ask more fundamental questions that do not mainly focus on how to optimize isolated technical artefacts but on a wider context of their deployment. AI ethics tacitly positions its locus of activity within narrow system boundaries and is often limited to questions of a "value sensitive" design and appropriate implementation of AI systems. It often does this instead of asking whether these systems should be developed or applied in certain domains in the first place. This would require broadening the analytic scope, and envisioning social fields at large instead of isolated, small contexts. In technology ethics, this step towards focusing on socio-technical assemblages is common sense and promoted by the actor-network theory, science and technology studies, new materialism, and other major theories [3–9]. AI ethics would be wise to follow suit. This would allow for a more "radical" approach that can critically deliberate the many externalities of AI by considering the full-stack supply chain of this technology.

It is very hard to find papers on AI ethics that question the use of particular AI applications per se or that challenge the mismatch between AI's inherent logic of automation and the genuinely human-centered requirements of many social domains. Involuntarily, this reveals an affirmative bedrock

✉ Thilo Hagendorff
thilo.hagendorff@uni-tuebingen.de

[1] Cluster of Excellence "Machine Learning: New Perspectives for Science", University of Tuebingen, Tübingen, Germany

Published online: 09 December 2021

of the discipline where "red lines" are only mentioned in the most obvious areas like autonomous weapon systems or AI-guided human rights violations. Papers in AI ethics implicitly see AI as an inevitability, as a natural next step or progression of technology. AI is "better" than pre-AI technologies, hence the very task ethics has to accomplish is to ensure "ethical" or "trustworthy AI". However, it is wrong to assume that the goal is ethical AI. Rather, the primary aim from which detailed norms can be derived should be a peaceful, sustainable, and just society. Hence, AI ethics must dare to ask the question where in an ethical society one should use AI and its inherent principle of predictive modeling and classification at all. It is a major weakness of AI ethics to simply condone AI systems being applied in more and more areas of society, merely because it fits existing (dys-)functional logics of companies, schools, police stations, and the like.

Accordingly, AI ethics often reduces its scrutiny to design decisions and frequently disregards the economic or political situatedness of the technological systems in question. The problem is that even under conditions under which all major requirements of AI ethics principles are fulfilled (meaning that AI systems produce fair outputs, are explainable, privacy-preserving, robust, and accountable), AI applications can be used to exacerbate environmental damages, foster social oppression, support unethical business models, and other problematic developments. AI ethics often frames AI applications as isolated technical artefacts or entities that can be optimized by experts who apply technical solutions to technical problems that are depicted as problems of ethical AI. Contrary to this position, this paper argues that AI applications must not be conceived in isolation but within a larger network of social and ecological dependencies and relationships.

With that in mind, it should be noted that AI applications often harm individuals from already marginalized and vulnerable communities [2, 10]. AI is in many cases a technique of rule, an application for monitoring and controlling people that is often in the hands of societal groups that already possess significant power [11]. It can exacerbate social inequalities and sustain existing economic structures in which the privileged "are processed more by people, the masses by machines" [12]. Policing techniques tend to criminalize poverty [13], companies automatically sort customers into opaque reputation silos based on patterns of their daily lives [14], weapons-grade communication techniques are utilized to deform political opinion formation [15, 16], AI applications are built to assist with or conduct torture [17], social media platforms intentionally foster addictive behavior via AI-based information filters [18, 19], and many more. The list must remain arbitrary and fragmentary since examples of AI misuse are plenty. On the other hand, and strikingly, AI applications that are explicitly deployed in contexts of care, nurture, help, welfare, social or ecological responsibility are relatively rare [20]—with the field of medical AI as well as a variety projects in the area of Beneficial AI as potential exceptions where AI is explicitly applied for the social good [21]. This is likely due to AI's inherent logic of statistical prediction and classification that may be of less use in these contexts.

Ultimately, one can pose the question which side prevails, the good or the bad use cases. Obviously, this issue cannot be answered scientifically. However, empirical research on the values that are embedded in AI research offer sobering results [22]. Among highly cited AI papers that were published at top machine learning conferences, values like performance, building on past work, generalization, efficiency, quantitative evidence, novelty, or understanding are prevalent and prioritized in stark disfavor of values of societal needs, justice, diversity, critique, and other ethical principles that are covered extremely seldomly, if at all. The prioritized values seem to be mere technical issues, but are indirectly laden with sociopolitical implications that revolve around power centralization, benefiting already wealthy industries, and disregarding the interests of underprivileged social groups. This can be demonstrated by the selection of performance metrics, which is a value-laden choice in itself, by the preference for large datasets which favor those few organizations that are able to obtain and process them, by the conservative approach to generalization, where the past is assumed to be equal to the future, by the relative equalization of efficiency and scalability, which again favors powerful actors in the information technology field, by focusing on novelty in relation to past work while denigrating rectifying arguments or critique on existing considerations, and many more. All in all, papers in AI research are motivated by the needs of the community itself, not by the needs of society at large. Furthermore, the papers hardly mention risks and expose a significant blindness to potential harms, even when socially contentious applications in areas like surveillance or misinformation are being researched [22]. Notwithstanding this point, industry involvement in machine learning research is on the rise, while conflicting interests between corporations and public institutions are seldomly disclosed [23]. With these aspects in mind, AI ethics should definitively dare to become more "radical" in the aforementioned sense. It should not be a mere heel lifter for technical solutions in fairness, explainability, or privacy, that are evolving around technical values themselves, but go beyond that.

Moreover, most of the major AI ethics initiatives have significant methodological shortcomings [24]. They choose a principled approach and stipulate a list of rules and standards, suggesting that this has a governing effect on AI research and development. However, as more and more papers on AI metaethics show, many approaches in AI ethics, among them the prevalent principled,

deontological approach, fail in many regards. Typically, AI ethics approaches have no reinforcement mechanisms [25, 26], they are often used for mere marketing purposes [27], they are not sensitive to different contexts and situations [28], they are naïve from a moral psychology perspective in not considering effects of bounded ethicality [29], they hardly have any influence on behavioral routines of practitioners [30], they fail to address the technical complexity of AI, for instance by only focusing on supervised machine learning applications and disregarding ethical implications of deep reinforcement learning etc. [31], while at the same time being technologically deterministic [32], they use terms and concepts that are often too abstract to be put into practice [24], etc. To improve the latter shortcoming, AI ethics recently underwent a practical turn, stressing its will to put principles into practice and to finally have a tangible effect. But the typologies and guidelines on how to put AI ethics into practice stand by the principled approach instead of focusing on virtues, for instance [33, 34]. Thus, they stand by the problems of deontology, although at least one significant effect of the practical turn should be that ethics guidelines finally acquired a certain degree of relevance in their potential to guide decision-making routines in AI engineers and researchers. This paper, however, explicitly does not focus on these methodological questions. Instead, it is focused on the selection of topics that are deemed to be important in AI ethics.

AI ethics has steered itself into a situation in which it has somewhat restricted lines of thought regarding the issues it is focusing on. Hence, this paper is restricted to the question of which topics and issues AI ethics actually attends to. During the last decade, roughly 100 AI ethics guidelines have appeared, while the assumption that newer guidelines were inspired by older ones seems feasible, since they are often very similar [26]. Accordingly, meta-analysis on these guidelines found that they comprise a consensus on a certain set of principles [26, 35, 36], namely explainability, fairness, privacy, safety, and accountability. These principles appear reasonable at first glance since they address AI-specific problem areas. However, a topic rarely discussed in this context is that many of these principles stand in tension to or even contradict each other. Fairness necessarily comes at the price of accuracy, privacy compromises quality and efficiency of services, transparency can undermine safety, etc. [37]. In spite of this, these reoccurring principles are the result of the relative "autopoiesis" of the AI ethics discourse, where one code of ethics is composed of or at least inspired by other codes of ethics, thus the codes are often mere echoes of each other. However, this has transported AI ethics into a self-referential state in which it becomes difficult to think outside of the box. This has also caused a significant redundancy in AI ethics. At the same time, topics that have high ethical significance, but are not part of the established discourse, tip over the edge. A common denominator of these topics is that they cannot be solved technically or occur as externalities outside of the "capitalistic centers". Part of the self-constriction of AI ethics is its metrification and the focus on issues that can be idealized, rendered calculable, and that can be remedied by genuine technical solutions. These issues mainly revolve around explainability, fairness, safety, accountability, and privacy.

In other words, the technically oriented AI community develops technical means that have socially relevant ramifications and seem to require technical rectifications. AI ethics then takes up these predetermined issues and enriches them philosophically. In this process, however, AI ethics often loses track of its strength, namely its sensitivity to harms and suffering as well as its productive power to devise positive visions for human flourishing. Strikingly, large parts of AI ethics are more concerned with the quest for post hoc model explanations, privacy preserving data set processing, auditing frameworks for learning algorithms, etc. than AI's enormous ecological footprint, crowdworker exploitation, experiments on live animal's brains to achieve more brain-like neural nets, feminized AI personas, the trickling down of tools that were once used only by intelligence or military agencies to civilian contexts, and many more pressing issues. This way, AI ethics statements corroborate business-as-usual-approaches [38], where at most, it can act as a "bicycle brake on an intercontinental aircraft" [39].

This paper is intended to draw attention to the forgotten issues of AI ethics and to propel the AI ethics discourse forwards. It has two parts. The first part is on specific blind spots with regard to the particular topics AI ethics focusses on. It critically discusses AI ethics' emphasis on issues that are deemed to be calculable or, in other words, the narrow emphasis on idealized, technical solutions for sociotechnical problems. The three examples that are discussed are explainability, fairness, and privacy. AI ethics discourses deliver conceptual contexts for these principles and repeatedly stress their importance, but as of yet fail to think further, meaning to question the conservative meaning and alignment that are given to these principles. The second part discusses blind spots regarding topics which are hardly mentioned in the AI ethics discourse despite their significant ethical importance that clearly exceed the importance of many traditional AI ethics' principles. Here, especially negative externalities of AI systems are scrutinized. Ultimately, this paper is intended to be a critical contribution to the ongoing development of AI ethics. The hope that is connected to it is that it can influence AI ethics in a way that brings back the field's particular strength, namely its sensitivity to tangible harms that are caused by and connected to technologies of AI.

2 In focus: calculable issues

When sifting through guidelines for AI ethics, a certain set of reoccurring principles that builds core requirements for trustworthy AI can be observed. As mentioned in the introduction, at least three meta-studies on AI ethics guidelines [26, 35, 36] confirm that explainability, fairness, privacy, safety, and accountability are mentioned in almost every code of ethics. Strikingly, most of these principles can be operationalized in mathematical terms and can thus be implemented via technical means. For instance, explainability can be provided by model simplifications [40], privacy standards can be satisfied by differential privacy [41], safety can be strengthened by red teaming exercises or technically versed safety audits [42], etc. Other ethical issues that are less calculable or harder to process mathematically are mentioned rather rarely, as meta-studies on AI ethics guidelines reveal [26]. Explainability, fairness, privacy, and related topics are so appealing for AI practitioners since these issues can be solved via proposing metrics that can quantify ethical principles using theories with clear target states and idealized conditions. One could even claim that AI ethics only reinforces existing practices of resolving technical problems in AI research and development that are done anyhow.

In the following, the paper exemplarily takes up three of the arguably most important AI ethics principles, namely explainability, fairness, and privacy, and elaborates on some of the shortcomings of these principles. It ultimately touches on the question whether their central role in AI ethics is justified. Interestingly, the focus on a calculable approach of dealing with ethical issues has led to the adoption of very conservative mainstream notions of fairness, privacy, etc. that regularly undermine critical research and alternative perspectives on these very values. This becomes clear, for instance, regarding the essentialization of race and gender in fairness-aware machine learning [43, 44], the individualistic data protection perspective in AI privacy [45], and many more. Again, the reason for this lies in the necessity of stripping down the multidimensionality of very complex social constructs to something that is measurable and calculable. The following three subchapters are supposed to paradigmatically shed light on some of the reasons why the three mentioned AI ethics principles could profit from critical perspectives. Ultimately, the purpose of the arguments brought forward is not to go into detail, but to provide a short glimpse on why AI ethics should widen its scope and recognize the many shortcomings of perceiving sociotechnical problems through the lens of calculability, ideal theory, and technical solutionism.

2.1 Example 1: explainable AI systems

One of the most discussed principles of AI ethics is the request for interpretability, transparency, or explainability [46–48]. One of the reasons for the importance of explainable AI stressed by researchers is its strong connection to fairness. Explainable machine behavior can underpin efforts to detect biases, it helps to make AI systems more trustworthy due to its potential to increase the confidence that a model will work as expected, it allows the inference of causal relationships in data, etc. In the end, the quest for explainability results from the complexity of machine learning models that simply exceed human capabilities for information processing—which often fuel enthusiastic claims of "superhuman" machines. Complexity reduction, though, can be achieved technically via visualizations, the segmentation of solution space for local explanations, the computation of the sensitivity a feature has upon model outputs, by trying to find simplified models with similar performances, by extracting relevant examples that relate to certain machine behaviors or by text explanations or other symbols that represent a model's functioning [40]. While it should be relatively obvious that explainability can be achieved via technical solutions, this topic has evoked many papers from philosophy [49–52]. But while philosophers can undoubtedly come up with valuable conceptual analyses or term interpretations, they cannot contribute to the core problem, that is making machine learning architectures actually explainable. At best, philosophers can define benchmarks and goals that practitioners then may strive to achieve, but even in this regard, only genuine technical expertise can be the bedrock to map out tangible steps that show whether the goals can be achieved in the first place or not. In practice, however, philosophical conceptual work and the explainable AI community seem in many cases to be detached from each other due to different bodies of knowledge, vocabularies, level of detail, scientific journals, etc. All in all, with regards to discourses on explainability, it seems that practical philosophy or, specifically, AI ethics, just goes along with what AI experts are doing anyway. In that process, AI ethics deems exactly those topics to be of importance that the technical community is dedicated to, but lacks its own compass and sensitivity to weighty ethical issues.

Having that said, it can even be questioned whether explainability is intrinsically valuable at all [53]. An AI system that is explainable in human terms can be seen as mere means to an end, namely, to foster fairness, trust, accountability, or individual control over decision-making processes. But explainability itself may be seen as an "empty" value that has no meaning besides its instrumental value to make the achievement of other values possible. In addition to that, explaining AI means to give a list of facts about an AI system, but it is unclear at what point this list can be deemed to

not require further explanation [54]. On top of that, knowledge about the process that led an AI system to a particular decision-making behavior does not simply transition to a justification for that very behavior that reaches beyond the mere description of a causal process in a technical artefact, comprising for instance a justification of whether a certain type of algorithmic decision making is an appropriate, acceptable solution for a given problem at all.

The argument that explainability is an "empty" value could also hold true in another regard, where it could be argued that fully explainable AI applications would make the use of AI redundant since in situations where humans possess full knowledge on whether given considerations and explanations for a decision are acceptable they could also make the decision themselves [55]. But since humans usually do not possess such knowledge, AI systems that offer explanations of their own decision-making would not be of much help. Hence, in domains where it is paramount that decision-making should not be opaque, black box machine learning should be replaced by "old-fashioned" types of automation or genuine human decision-making. In addition, even perfectly explainable models do not ensure that they are not used for unethical purposes in any way. Very often, explainable AI researchers demand that especially in high-stakes contexts like jurisdiction, criminal prosecution, loan approval, medical diagnosis, or military decisions, the explainability of AI-driven decision-making is of utmost importance. On the flipside, such claims implicitly entail that AI applications can legitimately be used in the mentioned context, if they just fulfill explainability as well as some other common principles. This, however, can support stifling the question whether one should use AI applications in these context at all, since the applications scenarios per se may be unethical.

2.2 Example 2: algorithmic fairness

The quest for fair AI has become so prevalent that sometimes, it dwindles to an unreflected end in itself. This can exemplarily be demonstrated through the discourse on facial recognition systems that have higher error rates in particular demographic groups, especially in females and people of color [56, 57], since these demographic groups are underrepresented in training datasets. But what does it actually mean to criticize missing accuracy in commercial and state-controlled facial recognition systems? It can definitely be problematic when errors in classifiers lead to groundless suspicion, biased social sorting processes, or other misjudgments. But fair facial recognition systems, which are implicitly demanded when criticizing machine biases, remain a problem when they are used to harm people. Tacitly, discourses in AI fairness ensure "inclusiveness" of marginalized demographic groups in algorithmic decision-making, but, depending on the context of application, this only makes it easier for companies and governments to recognize and surveil these groups. Machine bias in facial recognition systems, however, may allow some individuals to gain "temporary advantages by partially obscuring [themselves] from the eyes of the white supremacist state" [58]. Drawing on a somewhat far-fetched argument, it could even be stated that in this context, "machine bias" is just another term of the otherwise celebrated privacy-enhancing technology that is obfuscation [59]. Instead of a more "radical" approach that stresses the need to prohibit harmful surveillance technologies [60, 61], as suggested above, the AI ethics discourse falls into the "framing trap" [62], assuming that these technologies are legitimate as soon as they fulfil fairness criteria.

Taking up an idea from Fazelpour and Lipton [63], one can differentiate between ideal and non-ideal methodological approaches to machine fairness [64]. An ideal approach constructs idealized conditions, presumes clear target states, defines concrete evaluative standards, stipulates requirements for perfect justice, etc. Moreover, ideal fairness aims to treat similar individuals similarly, to have an eye on protected groups, to measure magnitudes of disparity, and the like. On the flipside, non-ideal approaches are less static and abstract, do not claim to possess ideal standards, are more sensitive to the complexities and causes of different types of interwoven injustices, and can better inform policy-makers on how to mitigate concrete cases of unfairness by using incremental strategies. Moreover, whereas algorithmic fairness is framed in comparative terms, according to Fazelpour and Lipton [63], non-ideal approaches can also work with non-comparative injustices, where fairness is determined by considering the deserts and merits of specific individuals instead of comparing outcomes for different individuals against each other. Furthermore, ideal approaches can define particular fairness criteria. Algorithms that satisfy these criteria are deemed to be axiomatically fair, disregarding cases that may have been overlooked and causes of unfairness that are not covered by the necessarily limited definition. In this context, a much-cited paper from Kleinberg et al. [65] shows that no method of algorithmic fairness can satisfy different fairness conditions at once, rendering fairness an unachievable ideal. The paper's insights, that are approved by succeeding papers [66], can simply be reinterpreted as an affirmation of a non-ideal world in which perfect justice is an unachievable ideal. Here, one must not succumb to the naturalistic fallacy and mix the description of actual unfairness with normative arguments about how fairness ought to be achieved or cannot be achieved in a certain sense at all. Non-ideal fairness, though, acknowledges the existence of different and at the same time often vague, imperfect fairness notions without claiming that fairness should not be an ideal that should be fulfilled whenever possible.

Fazelpour and Lipton's paper [63] marks a starting point in sketching out the often narrow and shortened fairness concepts in the quest for unbiased AI. Considering the long history of critiquing theories of justice, it can be noted that humans know other forms of social coexistence worth striving for besides justice—which puts the massive effort for fair machine learning in a completely new perspective. Without a doubt, justice is important and dispositions to foster it can hardly be questioned. But it is not an absolute value. Especially in cases where the goods that need to be distributed are not scarce, distributive justice is not a necessity. Equity comes into play under conditions of scarcity and limitations, when benevolence becomes suppressed and egoistic habitudes can cause individual suffering [67]. Otherwise, when goods are not scarce, it may well be that the call for fair (algorithmic) decision-making tacitly emerges from less accepted motives like envy and suspicion. Moreover, many "goods" cannot be distributed evenly at all, which for instance holds true for relationships of recognition or other kinds of cultural and social capital [68, 69]. Further, considerations of fair machine learning build on rules of reciprocity. Instead of ensuring fairness, the case could also be made for goodwill or even love [70] that transcends the idea of mutual self-interest inherent in these rules. Goodwill is an extension on perspectives of altruism and disinterestedness that must, by its very nature, be excluded from machine learning related fairness calculations. The same holds true for the idea of extending considerations of justice to components of attention to the individual situatedness of persons. Rule-governed processes in fair machine learning should be augmented by ethically versed judgements, sensitivity for individual cases, and complex forms of neutrality [71]. Exemplarily, this can mean making the transition from merely focusing on the avoidance of unfair disadvantages towards particular sociodemographic subgroups to questioning the production of privileges like whiteness or maleness [72]. In this context, it can be noted that while fairness is typically conceptualized via attributes of gender, race, nation or disability, these categories can in some cases be too strict. Hence, it may be a problem to formally encode and treat them as fixed instead of relational, social constructs or phenomena [43]. Furthermore, when claiming that false negatives and false positives must be equally distributed across these protected groups [73], AI fairness researchers must keep in mind that even if the goal of algorithmic fairness is achieved in this regard, on a different level of fairness, affected individuals still have different means to register, contest, and legally appeal to unfair algorithmic decision-making. Balancing errors does not mitigate this, unfairness remains part of the respective sociotechnical assemblage.

2.3 Example 3: privacy preserving machine learning

Privacy is among the principles mentioned most often in AI ethics [26]. Similar to principles on fairness and explainability, it must be assumed that this is due to the fact that privacy issues in the AI field can be solved via technical solutions, at least when privacy discussions do not concern the question of whether it is acceptable to use algorithmic decision-making in certain social contexts at all. Privacy, in theory, is supposed to ensure an individual's autonomy and self-determination, to be a defense against intrusions of government bodies or companies, to allow individuals to be unobserved when engaging in intimate social relations, to facilitate free opinion formation, etc. [74–76]. Since AI tools often rely on large amounts of personal data and are enablers of mass surveillance and personalized marketing as well as nudging techniques, which are notoriously privacy-sensitive fields, the tools are regarded as threats to informational as well as decisional privacy. To counteract these threats, legally established measures like the right for rectification or erasure, informed consent, purpose limitation, data minimization, independent privacy impact assessments, etc. are put to use in addition to privacy preserving techniques. Here, two technical solutions are mainly proposed, namely differential privacy and k-anonymization [41, 77–79]. These techniques imply altering datasets by adding noise or by manipulating them in such a way that complicates the identification of individuals. This, however, does not necessarily provide full privacy or full protection against re-identification. Moreover, making datasets more protective of privacy comes at the price of accuracy. On the flipside, implementing differential privacy can enable access to new data sources that hitherto could not be obtained due to privacy concerns [80]. This way, privacy-enhancing measurements can actually lead to an increase in the amount of collected sensitive behavioral data.

However that may be, AI ethics promotes these technical as well as legal solutions against privacy violations. The solutions' theoretical background, though, is based on classical and individualistic notions of privacy, which in turn are based on the idea of hiding sensitive information, of controlling data, or of restricting access to intimate details that are tied to a particular person [81]. Newer trends in informational privacy research, however, are extending the idea of privacy violations from a merely individualistic data protection perspective towards notions of interdependent, group, collective, or predictive privacy [82–85]. These concepts deal with interconnected settings where an individual's privacy is bound to others' decisions, with precautionary measures that aim at collectively avoiding ethically problematic cases of predictive analytics, with algorithmically assembled groups of individuals, and the like. In short, new

research shows that privacy or data protection is a cause that can only be addressed collectively.

Apart from that, more and more research questions the alleged high value of privacy—which also casts doubt on the massive importance privacy enjoys in AI codes of ethics. Privacy, with its focus on "hiding" identity factettes, personal information, and intimacy, can actually be an adversary of freedom, emancipation, security, and ultimately democracy [86]. Democratic processes are thwarted when political retribution can only be avoided by secrecy. Privacy allows individuals to shelter from political and economic abuse of power, but thus tacitly accepts it as a fact that cannot be changed [87]. Marginalized groups cannot challenge discrimination by remaining in hiding. Diversity is disguised by privacy, whereas non-conformism is actually widespread in modern societies due to networked publics that rely on surveillance techniques [88]. Furthermore, privacy promotes algorithmic discrimination since the less one knows about an individual the more one relies on inaccurate stereotypes and coarse group identities [37]. Moreover, restrictive privacy policies in health data access do not only prevent the removal of taboos regarding diseases and disabilities but also thwart public health researchers from using valuable information that can ultimately save countless lives [89, 90]. Beyond that, privacy promotes opportunism by the way of problematic changes of roles depending on the social context. Last but not least, it can protect detrimental norm violations and undermine legitimate criminal prosecution [91]. In short, researchers even question to what extent privacy and secrecy actually have a record of fostering societal equality and freedom [87]. Privacy avoids confronting prejudices and injustice and helps to avoid conflicts. Hence, social wrongs stay the same. This does not mean that privacy cannot help in mitigating social ills, but it is a workaround. It can alleviate symptoms, but it does not address root causes. It is a relic from the bourgeoisie's upcoming, but it is increasingly superseded by other values that characterize postmodern, networked information societies. Here, publicity rather than privacy is the default condition, and available big data allow for probabilistic inferences on people's most intimate traits even when they want to keep them secret [92–95]. All these considerations point to the necessity of reconsidering whether technical and legal measures for privacy protection are rightly so salient in AI ethics—or whether other topics should become more prominent. Ultimately, AI ethics' privacy principles aim at protecting sensitive personal information [96], be it by manipulating datasets in order to obfuscate them or by protesting against AI-driven inferences. But neither the datasets nor the inferences are the problem. The real issue is unfair social discrimination and intolerance that renders some information "sensitive" since its disclosure would initiate oppressive measures [45]. For AI ethics, at least when taking the "radical" approach seriously by addressing ethical issues at their roots, this means being more careful when protesting against AI-based inferences on sexual orientations, personality traits, health conditions etc. instead of stressing the growing importance of tolerance in post-privacy societies.

3 Out of focus: negative externalities

The previous chapter was concerned with specific blind spots with regard to the topics AI ethics focusses on. This chapter, then, is on blind spots with regard to the topics that are hardly or not discussed in AI ethics at all. These omitted topics are of significant ethical importance, and their common denominator is negative externalities. To describe and critically assess these externalities is, apart from very few exceptions, not part of AI ethics' language game. But what are the externalities about? In short, AI applications are mainly part of an "imperial lifestyle" [97] that is based on exclusiveness and possesses "capitalistic centers" as well as a suppressed "outside" on which it deflects all sorts of costs. As I want to explain in the course of this chapter, AI infrastructures directly as well as indirectly externalize their various costs on low-wage clickworkers, persons affected by ecological damages, exploited mineworkers, animals in laboratories, etc. These externalities of AI are massively underrepresented in the AI ethics discourse. Rather, it focuses on aspects that occur in the "capitalistic centers", like algorithmic discrimination, privacy issues or AI safety.

The term "negative externalities" originally stems from economics. Here, externalities occur when an entity, meaning a person or company, affects the welfare of other persons or companies in a way that is outside the market mechanism and that is not compensated [98, 99]. Hence, externalities cause indirect monetary or other kinds of social costs to individuals. Speaking in non-economic terms, externalities affect noninvolved third parties who did not consent to the effects the actions of particular first parties have on them. The classic example of negative externalities are cases of air, water, or noise pollution caused by private companies [100, 101]. From a psychological perspective, negative externalities emerge as a problem due to cultural boundaries, group affiliations, and ingroup favoritism [102, 103] that engender little empathic concern for harm in people if this harm occurs in social contexts that are "invisible" or outside of one's own circle of cultural perception. From a postcolonialist perspective, negative externalities are the result of colonial mechanisms of power, economics, and culture which emerge from advanced technologies like AI [104] and affect peripheries of metropoles or centers of power in economically developed countries, typically in the Western world or Global North. In the AI field, negative

externalities are also caused by the development as well as application of AI systems and also affect outgroup members, for instance when AI systems are beta-tested in vulnerable communities [105] or unfold in countries with fragile democracies, restricted access to human rights protection, or severe poverty [106]. These externalities, meaning negative effects on unrelated third parties that typically live outside of western capitalist centers are hardly discussed in AI ethics. It comes as no surprise that even in the origins of AI ethics guidelines themselves, there is a severe under-representation of geographic areas such as Africa, Central Asia, or South and Central America [35]. In general, the key debates in the AI discourse are framed by means of Western values, contexts, and concerns. However, it should be one of ethics' core strengths to overcome ingroup perspectives and emphasize outgroup concerns. AI ethics very often fails to do so. In the following, three subchapters paradigmatically shed light on some of the negative externalities that are typically overlooked. These are the ramifications of AI regarding precarious, low wage clickwork and the impact of AI systems on animals as well as ecosystems.

Discussing these topics in the context of AI ethics, though, provokes the critical question of the area of responsibility of the discipline per se. One could argue that AI ethics does not address the blind spots discussed in this chapter for reasons of modularity, since other fields of applied ethics—where AI ethics is a subfield—are already covering them. Obviously, it makes little sense to broaden the branch of AI ethics in a way that it "invades" these other fields. Such an argument, though, may require assuming clear boarders between the fields, which do not exist. In theory, applied ethics follows a certain taxonomy. However, the different fields define themselves mainly via the relation to a particular topic or social system [107], whereas significant overlaps instead of monolithic approaches occur—especially in the field of digital ethics, as a recent scientometric analysis revealed [108]. Information ethics overlaps with Internet ethics which overlaps with ethics of technology which overlaps with robot ethics, and so on. In the same vein, AI ethics builds a fringed cluster of AI-related topics and can intersect, among others, with business, animal, or environmental ethics, if the issue in question just bears a clear connection to AI technologies. Hence, in the following, I will argue that AI ethics could improve by shedding light on hitherto strongly disregarded issues that revolve around negative externalities that are directly connected to the use and development of AI systems, without ceding them to other fields of applied ethics which would be less apt to discuss them. Ultimately, chaining together various sub-disciplines can support ethical decision-making frameworks, but one discipline has to take the lead, whereas AI ethics should do so in the topics that are to be discussed in the following three subchapters.

3.1 Example 4: precarious annotation work

AI technologies tend to be mythologized [109]. And the more they are mythologized, the more a necessity emerges to hide demythologizing factors. These factors revolve around the dependence of today's AI technologies on human participation—despite that they are heralded as the next step in automatization. In many cases, AI harnesses human labor and behavior that is digitized by various tracking methods. This way, AI does not create intelligence, but captures it by tracking human cognitive and behavioral abilities [110]. Without empirically aggregating recordings of human behavior, or, in other words, the "predatory extractive practices" of "data colonialism" [111], many parts of machine learning would not be possible. An extensive infrastructure for "extracting" [2] valuable personal data or "capturing" [110] human behavior in distributed networks builds the bedrock for the computational capacity called AI. This functions via user-generated content, expressed or implicit relations between people, as well as behavioral traces [112]. Here, data are not the "new oil", not a resource to be "mined", but a product of human everyday activities that is capitalized by a few companies. For supervised machine learning, AI's current main method, datasets that are generated continuously as a by-product of digital technologies must be augmented with annotations to render them utilizable. This is where AI creates new kinds of precarious labor.

AI development has two sides. On one hand, the "centers" of technology development, the luxurious, celebrated, non-hierarchical, playful workspaces at Google, Apple, Facebook etc., which are seen as the "birthplaces" of AI technologies [113]. On the other hand, the "margins" of technology development: the hidden, low-status, low-wage labor at clickwork or labeling factories [114]. The clickwork industry is the invisible backbone of many AI technologies, and the growing market for third-party data labeling solutions was worth $1.7B in 2019 and will likely reach $4.1B by 2024 [115]. Despite its size and the ever-increasing amount of workload that is required, the according labor is mostly hidden and will likely not go away any time soon, even though significant efforts are being made to come up with methods for synthetic or fewer labels [116, 117], not least because of the massive cost savings that could be made in case manual clickwork could become automatized or superfluous. However, where the tech industry currently "fails" to replace handiwork with algorithms, clickworkers have to step in. They work on digital "assembly lines" where they are turned into a "computational service" [118]. They take care of repetitive, dull, and exhaustive data preparation and labeling work, transcribing audio files, putting texts into structured databases, marking objects on images, rate search results, moderate abusive content, etc. Basically, working conditions are as bad as the market tolerates [119]. Workers

are typically excluded from minimum wage or other worker protection laws [120]. Occasionally, the labelling is even done by prison inmates [121]. Roughly speaking, the typical pattern of the supply of labor by the Global South for the Global North is perpetuated. Here, recruiters can practice "labor arbitrage", buying labor from where it is cheapest, which results in a "race to the bottom" in wage rates [122]. Moreover, clickworkers often do not know what the purpose of their work is, which can lead to a situation where they unwittingly support the development of military AI, for instance [123].

With very few exceptions [2, 124], AI ethics is mute concerning the "human in the loop", the "backstage" of AI-based automatization. While in business ethics, ethical considerations regarding the various branches of the gig economy are essential part of the discourse [125–127], AI ethics should shed light on a particularly shrouded type of gig economy, namely the data annotation industry that is in many cases, the essential backbone for training data preparation and hence model training in machine learning. There are barely any voices demanding fair and generous wages for clickworkers, a diversity of tasks instead of highly repetitive ones, the provision of task contexts and transparency, the organization of workers via labor unions, the preparation for times when labeling work can be automatized itself, or the augmentation of datasheets for datasets with information about labeling processes, to name just a few examples how things could be improved. Instead, AI ethics problematizes the potential labor displacement of taxi and truck drivers, the displacement of journalists, shop assistants, and the like. In general, much attention is focused on labor displacement in masculinized professions, whereas for instance secretarial labor, which is mainly done by women [128] and which can be automatized to a high degree is less present in public debates [129]. Apart from that, AI ethics very rarely sheds light on the emergence of a growing clickwork industry that, instead of technically replacing human workforces [130], generates new kinds of work. This situation resembles times when "computers were women" [131]. Nowadays, however, the marginalized clickworkers mainly situated in the Global South have become the human backbone of computing.

3.2 Example 5: anthropocentrism in AI ethics

AI ethics is strictly anthropocentric. It is tailored to humans and mostly turns a blind eye on animals. Only a single, very recent study mentions the role of "nonhumans" in AI ethics [132]. The study, however, lacks obvious empirical examples for AI's impact on animals and remains very abstract, for instance by speculating about malevolent artificial general intelligence and its ramifications for animal populations. In addition to that, a few short articles in AI ethics address the question of how autonomous vehicles should behave when encountering animals [133]. In general, however, AI ethics is similar to many other scientific disciplines like sociology or psychology where animals are largely disregarded despite it making perfect sense to include them due to their similarity to humans, their mental capabilities, their ability to suffer, their intrinsic value, their moral status, etc. In fact, AI has a significant impact on animals in many regards—and vice versa. Animal brains are used to inspire model architectures [134] and animal capabilities are harnessed as a benchmark to measure AI performance [135]. Hence, animals are affected by the research on and development of this technology since it partially relies on animal testing. Here, one could again argue that AI ethics is the wrong discipline to discuss these issues, especially since animal ethics itself has a long history of reflecting on animal experiments and animal exploitation [136, 137]. Similar to other fields of applied ethics that overlap and blend into each other, AI ethics is indeed not supposed to conduce discussions of the foundations of animal ethics, but it has an obligation to take ethical issues into account that are directly associated and entangled with research and development of AI systems. In the same vein, AI ethics discusses algorithmic discrimination without shunting it off to social, business, or other fields of applied ethics. It discusses accountability despite its origin in law, autonomous weapon systems without leaving it to military ethics, medical decision-making irrespective its affiliation to medical ethics, and so on. Similarly, AI ethics should, for instance, not ignore animal testing that is solely done for the purpose of developing advanced, increasingly brain-like AI architectures [134].

The path that led to the success of brain-inspired, deep neural networks traces back to Rosenblatt's idea of the "perceptron" [138]. Here, insights into the functionality of neurons, which were heavily dependent on animal experiments, were used to develop new techniques for information processing, where neurons are simulated by computer programs. Perceptrons receive, similar to biological neurons, inputs, where the sum of the inputs determines whether the perceptron meets a threshold value and gives an output in the form of 1 or 0. This principle is the very fundamentum for today's multi-layered neural nets. Especially, the development of convolutional neural networks were originally inspired by single-cell recordings of mammalian visual cortex [139]. Moreover, current research in "AI attention" is inspired by the primate visual system [140]. In this approach, artificial neural networks do not process entire images but focus on certain areas of an image instead. Not least, reinforcement learning, another method where intelligent agents act in a way that maximizes certain rewards, was partly inspired by studies in animal psychology [141]. Moreover, current research aims at mapping the layout of neurons of a cubic millimetre of rat brain, using cutting-edge brain imaging tools. Said tools are for instance infrared lasers scanning the

brain of fixed living animals and microscopes to scrutinize slices of rat and mouse brain, unravelling information on neural circuitries to be able to build more brain-like artificial neural nets [142].

In short, parts of AI research draw upon animals or animal experiments. Advanced neuroimaging or single-cell recording techniques are used in in vivo experiments, visualizing processes in the brain on different levels of magnification, ranging down to single synapses [143]. In vivo methods often mean that brains of living animals are penetrated, which comprises stereotaxis surgery, a "survivable procedure" [144] where animals are held in place via clamps, after which a craniotomy—a surgical removal of a part of the cranial bone to expose the brain—is performed with a drill to access the brain. Despite brain imaging or recording techniques, further "tools" from genetic bioengineering granted more and more insight into the inner "functionality" of mammalian brains. To increase the statistical power of experiments, preferably large sample sizes, meaning a great number of animals such as rodents, are used. Despite being required to anesthetize the animals via gas or pharmacological agents, it must be assumed that the animals subjected to neurobiological experiments suffer severe and prolonged post-surgery pain as well as distress from captivity, behavioral experiments, and various other sources [145]. The ethical implications of such types of experiments are not discussed in AI ethics at all despite the fact that the experiments serve the exact purpose of propelling AI research and to make AI less "artificial" by copying properties of biological brains in artificial neural nets [134, 146, 147]. Here, AI ethics can draw on arguments of a large corpus of research in animal ethics as well as a long tradition of ethical considerations with regard to laboratory animal welfare [148–150] to critically point at the fact that in AI research animals are primarily seen as carriers and suppliers of data who are not directly morally considerable. This perspective, namely that animals possess a moral status, is 'common sense' within animal ethics. Regardless of whether one argues that animals should simply not be harmed [151] or whether one grants them rights [152–154], completely ignoring their interests is considered ethically wrong. AI ethics, however, has nothing to say on that.

Despite AI research and its conjunction with animal experiments, animals play a further role in the AI world as they are also controlled by AI. AI builds the foundation of modern surveillance technologies, where sensors produce too much data for human observers to sift through. These surveillance technologies are not solely directed towards humans, but also towards animals, especially farmed animals. The confinement of billions of farmed animals requires technology that can be employed to monitor, restrict and suppress the animal's agency [155]. This has recently been reinforced by AI technologies [156–161]. Animals are analyzed via camera vision-based systems for herd "management" purposes such as tracking, fattening control, or facial recognition [162]. Advanced closed-circuit television systems are used to automatically detect and track pigs, allowing the automatic detection of potential health problems without human observation. Robots drive through poultry houses and collect "layers", which are floor eggs not placed in the designated nest boxes. Acoustic data analytic tools listen to "poultry" birdcalls and process the audio files so that they can be used as early warning systems to detect diseases. Feeding sensors in combination with predictive analytics are used for "performance" prediction. Support vector machine classifiers analyzing sensor data are used to detect early symptoms of lameness in cows. On cockroach farms, AI systems process sensor data for features like the insects' temperature, food intake and humidity to ensure an optimal growth environment. Robot vehicles driving around in factory farms for "broilers" are supposed to move the animals around to improve the feed conversion ratio. In "processing plants", intelligent robot arms equipped with cameras are used for "poultry deboning", cutting animal bodies apart with greater efficiency than human butchers. Besides AI applications that directly aim at animal control, various AI methods are used to optimize fodder compositions, design vaccines, and to analyze animal's genetics.

In short, animals in factory farms are exposed to an environment that is increasingly surveilled and controlled by the means of AI. In that, factory farming bears resemblance to modern information societies at large. While AI-based surveillance that is directed at humans is heavily discussed in science and mass media, though, animal surveillance is seldomly mentioned [163], if ever, despite its significant ethical implications that range from the commodification of sentient beings to the even greater emotional distances in the human perception of animal suffering. Ultimately, AI ethics turns a blind eye on the role neurobiological animal experiments play in inspiring model architectures as well as the many areas where animals' conduct of life is subject to AI tools. With that said, this paper stresses the importance of overcoming the anthropocentrism inherent in AI ethics to perceive AI-related suffering in all relevant contexts.

3.3 Example 6: AI systems' ecological footprint

Ecosystem services do not only have an unprecedented monetary value for humans [164], but they are the very reason for the possibility of human life on earth. Over-exploitation of ecosystems harms future generations [165], poor and already underprivileged people [166], animals [167], and many more. Whereas the building industry, agroindustry, or transport industry seem to be main drivers of ecosystem destruction and climate change, the information and communication industry also play a tangential role. This holds

especially true for the AI field. The term "AI" bears a linguistic similarity to the term "cloud". The notion of cloud computing suggests that it lacks materiality, that it is invisible and placeless, that data are "stored in the troposphere" [168], where in fact, big data is anything but transcendent or amorphous. It is grounded in fragile physical infrastructures, cables, hardware, routers, server buildings, power grids, cooling systems, satellites, etc., all of whom require natural resources. A similar situation unfolds with AI. The term "artificial intelligence" again suggests something immaterial, a mental quality that has seemingly no physical implications. This could not be farther from the truth. To appreciate this, one must switch from a data level, where the real material complexities of AI systems are far out of sight, to an infrastructural level and to the complete supply chain.

Here, and first of all, it becomes visible that the cloud and AI systems are intrinsically intertwined. The cloud builds the necessary condition of AI, and the material implications of the cloud cling onto AI, too. These implications are far-reaching and manifest themselves in global networks of cable infrastructures, labor division, logistics, distribution, and manifold externalities. These networks comprise lithium, tin, cobalt and other mines that deliver essential minerals for electronic components or batteries that are part of every digital mobile device, smelters and refiners that produce acidic, radioactive, and otherwise harmful waste products, storage systems and warehouses for logistics and transportation operations, energy and water hungry data centers, affiliated cooling systems, diesel powered generators for backup purposes in cases of blackouts, data annotation factories, collection operations for toxic electronic waste consisting of technical devices with a lifespan of a few years, and many more [2, 169]. All these mining, shipping, manufacturing, and garbage incineration operations are heavily destructive, have a high burden on ecosystems, and come at the cost of human lives, child labor, wildlife populations, natural habitats, toxins in the ground, water, and air, public health, political instability and tensions and low wage labor markets. The material conditions that allow AI usage, especially rare earth elements or "conflict minerals", are also triggers for military operations, violence, murder, and migration that surround the already brutal and slavery-like industry of mining [170].

AI systems are not just demanding in terms of material resources, they also require a lot of energy. Electronic machines, in contrast to combustion engines, can in principle be used sustainably by consuming electricity from renewable energy sources. In practice, however, in many countries, only small proportions of electricity are renewable [171]. Accordingly, powering the computational resources that are required to collect large amounts of training data and to train, test, and apply large AI models comes with a significant carbon footprint [172]. Strubell et al. [173] conducted a life cycle assessment of several large AI models and found out that they can emit around three hundred thousand kilograms of carbon dioxide equivalent. The reason for this lies in the many ways machine learning methods go along with a "bigger is better" approach which prioritizes accuracy over efficiency [174] with costly trial and error processes which span from practitioners intuitively setting up model parameters all to neural architecture search and other tuning and automated optimization processes. Ultimately, the information and communication industry, which incorporates the AI field, has a carbon footprint that is bigger than that of the aviation industry [175]. But while there is flight shame, there is no such remorse for AI use, although some AI ethics researchers have tentatively started to develop a critical perspective on the role that AI has in contributing to climate change [174, 176, 177]. Much of this is dependent on where training servers are located, which energy grid is used, how long models are trained, and what hardware accelerators are in use [178]. However, even under perfect conditions where only renewable energy sources are used, it seems likely that AI remains a polluting technology in many industry sectors due to the business purposes it is utilized for. On one hand, AI technologies are heralded as technical solutions to the climate crisis by helping to develop low-emission infrastructures, operate smart grids, help foster sustainable consumption and production, etc. [21, 179]. On the other hand, AI technologies are used to buttress industries and business models that are environmentally harmful—let alone rebound effects in industries that are deemed to be sustainable [180]. In this regard, the tip of the iceberg is the collaboration between the largest AI companies and the fossil fuel industry [181, 182] which does not only comprise the optimization of oil and gas extraction, but goes so far as to actively support climate change deniers [183]. However, the fact that in theory, AI technologies can in sum enable more than inhibit the accomplishment of targets defined by the sustainable development goals [179], may be a reason for hope.

4 Conclusion

Modern AI ethics is a field in the making. It has undergone different phases, from its early beginning that was mainly characterized by the composition of various lists and frameworks of ethical principles to its current state that can be described as a practical turn, whereby principles are to be translated into practice [33, 184]. However, this rather fast and self-critical methodological advancement of the field is contrasted by a relative standstill in terms of the topics that are discussed. AI ethics is tantamount to a certain set of reoccurring issues that are mainly evolving around explainability, fairness, privacy, accountability, safety, and a few

more. The selection of topics, so it seems, is dictated by a dynamic where AI ethics is reacting to those sociotechnical problems for which AI practitioners can formulate technical solutions. Since these solutions, such as fairness checklists, differential privacy, post hoc explainability approaches, etc., are heavily researched in the AI field, AI ethics reacts to these trends by enriching them with conceptual considerations. However, in this process, AI ethics tends to lose its actual strength, namely its sensitivity to harms and suffering as well as its ability to recognize externalities which require overcoming ingroup perspectives and emphasizing outgroup concerns. This paper is a comment on the related shortcomings of AI ethics. It makes the case for a broadening of the topics AI ethics is concerned with as well as for a more critical perspective on the current set of topics that also allows alternative standpoints on explainability, fairness, privacy, and the like.

Ultimately, the widely established, narrow technical definitions of AI ethics principles have the severe disadvantage of evoking the notion that only experts in the field are capable of reasonable AI governance. This eliminates a more inclusive social and political governance approach, or in other words, true democratic oversight. The AI ethics discourse demarcates "the public" from "stakeholders" or experts such as machine learning practitioners, businesspersons, professional ethicists, or software engineers who are those who educate others [32]. Via public engagement, communication offensives, citizen science, youth competitions, science days, or mere advertisement posters, the public is supposed to get a glimpse of the otherwise incomprehensible complexity of AI research. But by maintaining the notion of a purely technical solutionism for ethical issues that are connected to AI technologies and that apparently revolve mainly around explainability, fairness, or privacy, citizens who are not in the field and who will typically not be able to gain sufficient technical insights are subject to a fake inclusion. Rudimentary democratic oversight is simulated, but as long as genuine sociopolitical, multidimensional approaches for AI technology assessments are not considered and as long as AI ethics stays in its restricted tailoring of topics that only comprise design decisions but not broader social systems, AI governance will remain an elitist project. This can likely be to the disfavor of many societal groups, whereas citizen inclusion will remain a mere lip service. The latter will only be able to trust, which may also explain the inflationary use of the term "trustworthy AI". Trust is a mechanism for the reduction of complexity [185]. This mechanism seems to become more important the more inscrutable the veil of programming code and statistical models becomes. Trust enables the suppression of moments of insecurity, masking risks. But in doing so, trust is itself a risky endeavor. The resulting benefits of a trust relation may not counterbalance the disadvantage of the breach of trust. Trusting too much is always imprudent. That is why mistrust can be very important, especially regarding powerful technological artefacts. Mistrust leads to a situation where individuals do not ignore risks, but perceive them as such and react appropriately. Trustworthy AI may thus be the wrong goal to aim at.

Authors' contributions TH is the sole author and wrote the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was supported by the Cluster of Excellence "Machine Learning—New Perspectives for Science" funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—Reference Number EXC 2064/1—Project ID 390727645.

Availability of data and materials Not applicable.

Code availability Not applicable.

Declarations

Conflicts of interest None.

References

1. Crawford, K., Calo, R.: There is a blind spot in AI research. Nature **538**, 311–313 (2016)
2. Crawford, K.: Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press, New Haven (2021)
3. Latour, B.: Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford University Press, New York (2005)
4. Latour, B., Woolgar, S.: Laboratory Life: The Construction of Scientific Facts. Princeton University Press, Princeton (1986)
5. Barad, K.: Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning. Duke University Press, Durham (2007)
6. Joerges, B., Nowotny, H. (eds.): Social Studies of Science and Technology: Looking Back. Kluwer Academic Publishers, Dordrecht (2003)
7. MacKenzie, D., Wajcman, J. (eds.): The Social Shaping of Technology. Open University Press, Buckingham (1999)
8. Jasanoff, S., Markle, G.E., Peterson, J.C., Pinch, T.J. (eds.): Handbook of Science and Technology Studies. SAGE Publications, London (1995)

9. Hackett, E.J., Amsterdamska, O., Lynch, M., Wajcman, J. (eds.): The Handbook of Science and Technology Studies. MIT Press, Cambridge (2008)
10. Eubanks, V.: Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Marting's Press, New York (2018)
11. Cohen, J.E.: The biopolitical public domain: The legal construction of the surveillance economy. Philos. Technol. **31**, 213–233 (2018)
12. O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishers, New York (2016)
13. Pager, D., Shepherd, H.: The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. Annu. Rev. Sociol. **34**, 181–209 (2008)
14. Lyon, D.: Surveillance as social sorting: Computer codes and mobile bodies. In: Lyon, D. (ed.) Surveillance as Social Sorting: Privacy, Risk, and Digital Discrimination, pp. 13–30. Routledge, London (2003)
15. Cadwalladr, C.: The Great Hack: the film that goes behind the scenes of the Facebook data scandal, 2019. https://www.theguardian.com/uk-news/2019/jul/20/the-great-hack-cambridge-analytica-scandal-facebook-netflix (accessed 11 October 2019).
16. Matz, S.C., Kosinski, M., Nave, G., Stillwell, D.: Psychological targeting as an effective approach to digital mass persuasion. Proc. Natl. Acad. Sci. U.S.A. **2**, 1–6 (2017)
17. McAllister, A.: Stranger than science fiction: The rise of A.I. interrogation in the dawn of autonomous robots and the need for an additional protocol to the U.N. convention against torture. Minnesota Law Rev. **101**, 2527–2573 (2017)
18. Kuss, D.J., Griffiths, M.D.: Social networking sites and addiction: Ten lessons learned. Int. J. Environ. Res. Public Health **14**, 2 (2017)
19. Hagendorff, T.: Jenseits der puren Datenökonomie - Social-Media-Plattformen besser designen. In: Ochs, C., Friedewald, M., Hess, T., Lamla, J. (eds.) Die Zukunft der Datenökonomie, pp. 327–342. Springer, Wiesbaden (2019)
20. Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J.C., Sellitto, M., Shoham, Y., Clark, J., Perrault, R., Index, T.A.I.: Annual report: AI index steering committee. Stanford Kalifornien **2021**, 1–222 (2021)
21. Chui, M., Harryson, M., Manyika, J., Roberts, R., Chung, R., van Heteren, A., Nel, P.: Notes from the AI Frontier: Applying AI for Social Good. McKinsey Global Institute, McKinsey&Company, 2018, pp. 1–52.
22. Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., Bao, M.: The values encoded in machine learning research, arXiv (2021) 1–28.
23. Hagendorff, T., Meding, K.: Ethical considerations and statistical analysis of industry involvement in machine learning research. AI & Soc. J. Knowle. Cult. Commun. **2**, 1–11 (2021)
24. Mittelstadt, B.: Principles alone cannot guarantee ethical AI, Nature. Machine Intelligence **1**, 501–507 (2019)
25. Rességuier, A., Rodrigues, R.: AI ethics should not remain toothless! A call to bring back the teeth of ethics. Big Data Soc. **7**, 1–5 (2020)
26. Hagendorff, T.: The ethics of AI ethics: An evaluation of guidelines. Mind. Mach. **30**, 457–461 (2020)
27. Wagner, B.: Ethics as an Escape from Regulation: From ethics-washing to ethics-shopping? In: Hildebrandt, M. (ed.) Bein Profiled: Cogitas ergo sum, pp. 84–89. Amsterdam University Press, Amsterdam (2018)
28. Lauer, D.: You cannot have AI ethics without ethics, AI Ethics (2020) 1–5.
29. Hagendorff, T.: AI virtues: The missing link in putting AI ethics into practice, arXiv (2020) 1–20.
30. McNamara, A., Smith, J., Murphy-Hill, E.: Does ACM's code of ethics change ethical decision making in software development?, in: Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering - ESEC/FSE 2018, ACM Press, New York,, 2018, pp. 1–7.
31. Whittlestone, J., Arulkumaran, K., Crosby, M.: The societal implications of deep reinforcement learning. J. Artif. Intell. Res. **70**, 1003–1030 (2021)
32. Greene, D., Hoffman, A.L., Stark, L.: Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning, Hawaii International Conference on System Sciences (2019) 1–10.
33. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From what to how an overview of AI ethics tools, methods and research to translate principles into practices, science and engineering. Ethics **26**, 2141–2168 (2020)
34. Hallensleben, S., Hustedt, C., Fetic, L., Fleischer, T., Grünke, P., Hagendorff, T., Hauer, M., Hauschke, A., Heesen, J., Herrmann, M., Hillerbrand, R., Hubig, C., Kaminski, A., Krafft, T.D., Loh, W., Otto, P., Puntschuh, M.: From Principles to Practice: An interdisciplinary framework to operationalise AI ethics, Bertelsmann Stiftung, Gütersloh, 2020, pp. 1–56.
35. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**, 389–399 (2019)
36. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., Srikumar, M.: Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication No. 2020–1, SSRN Journal (2020) 1–39.
37. Whittlestone, J., Nyrup, R., Alexandrova, A., Cave, S.: The role and limits of principles in AI Ethics: Towards a Focus on Tensions, 2019, pp. 1–7.
38. Stark, L., Greene, D., Hoffmann, A.L.: Critical perspectives on governance mechanisms for AI/ML systems. In: Roberge, J., Castelle, M. (eds.) The Cultural Life of Machine Learning, pp. 257–280. Springer International Publishing, Cham (2021)
39. Beck, U.: Gegengifte: Die organisierte Unverantwortlichkeit. Suhrkamp, Frankfurt am Main (1988)
40. A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115.
41. Dwork, C.: Differential Privacy: A Survey of Results. In: Agrawal, M., Du, D., Duan, Z., Li, A. (eds.) Theory and Applications of Models of Computation, pp. 1–19. Springer, Berlin (2008)
42. Falco, G., Shneiderman, B., Badger, J., Carrier, R., Dahbura, A., Danks, D., Eling, M., Goodloe, A., Gupta, J., Hart, C., Jirotka, M., Johnson, H., LaPointe, C., Llorens, A.J., Mackworth, A.K., Maple, C., Pálsson, S.E., Pasquale, F., Winfield, A., Yeong, Z.K.: Governing AI safety through independent audits. Nat Mach Intell **3**, 566–571 (2021)
43. Hanna, A., Denton, E., Smart, A., Smith-Loud, J.: Towards a critical race methodology in algorithmic fairness, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, ACM, New York, 2020, pp. 501–512.
44. Gebru, T.: Race and Gender. In: Dubber, M.D., Pasquale, F., Das, S., Powers, T.M., Ganascia, J.-G. (eds.) The Oxford Handbook of Ethics of AI, pp. 251–269. Oxford University Press, Oxford (2020)

45. Hagendorff, T.: From privacy to anti-discrimination in times of machine learning. Ethics Inf. Technol. **33**, 331–343 (2019)
46. Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I.: Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Med. Inform. Decis. Mak. **20**, 1–9 (2020)
47. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning, arXiv (2019) 1–10.
48. Mittelstadt, B., Russell, C., Wachter, S.: Explaining explanations in AI. Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19 (2019) 1–10.
49. Coeckelbergh, M.: Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. Sci. Eng. Ethics **26**, 2051–2068 (2020)
50. Fazi, M.B.: Beyond human: deep learning, explainability and representation. Theory Cult. Soc. **2**, 1–23 (2020)
51. Erasmus, A., Brunet, T.D.P., Fisher, E.: What is interpretability? Philos. Technol. **2**, 1–30 (2020)
52. Rohlfing, K.J., Cimiano, P., Scharlau, I., Matzner, T., Buhl, H.M., Buschmeier, H., Esposito, E., Grimminger, A., Hammer, B., Hab-Umbach, R., Horwath, I., Hullermeier, E., Kern, F., Kopp, S., Thommes, K., Ngomo, A.-C.N., Schulte, C., Wachsmuth, H., Wagner, P., Wrede, B.: Explanation as a social practice: Toward a conceptual framework for the social design of AI systems. IEEE Trans. Cogn. Dev. Syst. **2**, 1–12 (2021)
53. Colaner, N.: Is explainable artificial intelligence intrinsically valuable? AI & Soc. J. Knowl. Cult. Commun. **2**, 1–8 (2021)
54. Krishnan, M.: Against interpretability: A critical examination of the interpretability problem in machine learning. Philos. Technol. **33**, 487–502 (2020)
55. Robbins, S.: A Misdirected Principle with a Catch: Explicability for AI. Mind. Mach. **29**, 495–514 (2019)
56. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. in: Proceedings of Machine Learning Research, New York, eighthfirst ed., PMLR, 2018, pp. 1–15.
57. Grush, L.: Google engineer apologizes after Photos app tags two black people as gorillas, 2015. http://www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas (accessed 11 December 2015).
58. N. Hassein, Against Black Inclusion in Facial Recognition, 2017. https://digitaltalkingdrum.com/2017/08/15/against-black-inclusion-in-facial-recognition/ (accessed 2 July 2021).
59. Brunton, F., Nissenbaum, H.: Obfuscation: A User's Guide For Privacy And Protest. The MIT Press, Cambridge (2015)
60. W. Hartzog, Facial Recognition Is the Perfect Tool for Oppression, 2018. https://medium.com/s/story/facial-recognition-is-the-perfect-tool-for-oppression-bc2a08f0fe66 (accessed 7 July 2021).
61. Stark, L.: Facial recognition is the plutonium of AI. XRDS **25**, 50–55 (2019)
62. A.D. Selbst, d. boyd, S.A. Friedler, S. Venkatasubramanian, J. Vertesi, Fairness and Abstraction in Sociotechnical Systems, ACT Conference on Fairness, Accountability, and Transparency (FAT) 1 (2018) 1–17.
63. S. Fazelpour, Z.C. Lipton, Algorithmic Fairness from a Non-ideal Perspective, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, ACM, New York, 2020, pp. 57–63.
64. Valentini, L.: Ideal vs non-ideal theory: A conceptual map. Philos Compass **7**, 654–664 (2012)
65. J.M. Kleinberg, S. Mullainathan, M. Raghavan, Inherent Trade-Offs in the Fair Determination of Risk Scores, arXiv (2016) 1–23.
66. Saravanakumar, K.K.: The impossibility theorem of machine fairness—a causal perspective. arXiv (2021) 1–7.
67. Hume, D.: An Enquiry Concerning the Principles of Morals. Prometheus Books, Amherst (2004)
68. Honneth, A.: Recognition and Justice. Acta Sociol. **47**, 351–364 (2004)
69. Bourdieu, P.: Distinction: A Social Critique of the Judgement of Taste. Harvard University Press, Cambridge (1984)
70. Ricoeur, P.: Love and justice. Philos. Soc. Criticism **21**, 23–39 (1995)
71. Nussbaum, M.C.: Poetic Justice: The Literacy Imagination and Public Life. Beacon Press, Boston (1995)
72. Hoffmann, A.L.: Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. Inf. Commun. Soc. **22**, 900–915 (2019)
73. Hardt, M., Price, E., Srebro, N.: Equality of Opportunity in Supervised Learning, arXiv (2016) 1–22.
74. Westin, A.F.: Privacy and Freedom. Atheneum, New York (1967)
75. Nissenbaum, H.: Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford University Press, Stanford (2010)
76. Rössler, B., Mokrosinska, D. (eds.): Social Dimensions of Privacy: Interdisciplinary Perspectives. Cambridge University Press, Cambridge (2015)
77. C. Dwork, Differential Privacy, in: D. Hutchison, T. Kanade, J. Kittler, J.M. Kleinberg, F. Mattern, J.C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M.Y. Vardi, G. Weikum, M. Bugliesi, B. Preneel, V. Sassone, I. Wegener (Eds.), Automata, Languages and Programming, Springer, Berlin, 2006, pp. 1–12.
78. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy, FNT in theoretical computer. Science **9**, 211–407 (2013)
79. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Tech. Rep. SR **I**, 1–19 (1998)
80. Kearns, M., Roth, A.: The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford University Press, New York (2020)
81. Tavani, H.T.: Informational Privacy: Concepts, Theories, and Controversies. In: Himma, K.E., Tavani, H.T. (eds.) The Handbook of Information and Computer Ethics, pp. 131–164. Wiley, Hoboken (2008)
82. Biczók, G., Chia, P.H.: Interdependent Privacy: Let Me Share Your Data. Springer, Berlin (2013)
83. Yu, P., Grossklags, J.: Towards a model on the factors influencing social app users' valuation of interdependent privacy. Proc. Privacy Enhan. Technol. **2**, 61–81 (2016)
84. Mühlhoff, R.: Predictive privacy: Towards an applied ethics of data analytics. SSRN J. **2**, 1–24 (2021)
85. Mittelstadt, B.: From individual to group privacy in big data analytics. Philos. Technol. **30**, 475–494 (2017)
86. Hagendorff, T., der Das E.: Informationskontrolle: Zur Nutzung digitaler Medien jenseits von Privatheit und Datenschutz, Transcript, Bielefeld, 2017.
87. Belliger, A., Krieger, D.J.: Network Public Governance: On Privacy and the Informational Self, Transcript, Bielefeld, 2018.
88. Seemann, M., Das Neue S.: Strategien für die Welt nach dem digitalen Kontrollverlust, orange-press, Freiburg, 2014.
89. Wartenberg, D., Thompson, W.D.: Privacy versus public health: The impact of current confidentiallity rules. Am. J. Public Health **100**, 407–412 (2010)
90. Lynch, C., Holman, C.D.J., Moorin, R.E.: Use of Western Australian linked hospital morbidity and mortality data to explore theories of compression, expansion and dynamic equilibrium. Aust. Health Rev. **31**, 571–581 (2007)
91. G. Owen, N. Savage, The Tor Dark Net, Centre for International Governance Innovation; Royal Institute of International Affairs, Waterloo, Ontario, London, 2015, pp. 1–9.

92. Kosinski, M.: Facial recognition technology can expose political orientation from naturalistic facial images. Sci. Rep. **11**, 1–7 (2021)
93. Kosinski, M., Wang, Y.: Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. J. Pers. Soc. Psychol. **114**, 246–257 (2018)
94. Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. Proc. Natl. Acad. Sci. U.S.A. **110**, 5802–5805 (2013)
95. Kosinski, M., Matz, S.C., Gosling, S.D., Popov, V., Stillwell, D.: Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. Am. Psychol. **70**, 543–556 (2015)
96. Kaissis, G., Ziller, A., Passerat-Palmbach, J., Ryffel, T., Usynin, D., Trask, A., Lima, I., Mancuso, J., Jungmann, F., Steinborn, M.-M., Saleh, A., Makowski, M., Rueckert, D., Braren, R.: End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nat. Mach. Intell. **3**, 473–484 (2021)
97. Brand, U., Wissen, M.: The imperial mode of living: Everyday life and the ecological crisis of capitalism. Verso Books, Brooklyn (2021)
98. Hamowy, R.: Externalities. In: Hamowy, R. (ed.) The Encyclopedia of Libertarianism. Sage, Thousand Oaks (2008)
99. Pigou, A.C.: The Economics of Welfare. Taylor and Francis, London (2017)
100. Goodstein, E.S., Polasky, S.: Economics and the Environment. John Wiley & Sons Inc, Hoboken (2014)
101. Stern, N.: The economics of climate change. Am. Econ. Rev. **98**, 1–37 (2008)
102. Efferson, C., Lalive, R., Fehr, E.: The coevolution of cultural groups and ingroup favoritism. Science **321**, 1844–1849 (2008)
103. Mullen, B., Hu, L.: Perceptions of ingroup and outgroup variability: A meta-analytic integration. Basic Appl. Soc. Psychol. **10**, 233–252 (1989)
104. Mohamed, S., Png, M.-T., Isaac, W.: Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. Philos. Technol. **33**, 659–684 (2020)
105. Nyabola, N.: Digital Democracy, Analogue Politics. ZED BOOKS LTD, London (2018)
106. Milan, S., Treré, E.: Big data from the South(s): Beyond data universalism. Televis. New Med. **20**, 319–335 (2019)
107. Nida-Rümelin, J. (ed.): Angewandte Ethik: Die Bereichsethiken und ihre theoretische Fundierung. Alfred Kröner Verlag, Stuttgart, Ein Handbuch (2005)
108. Mahieu, R., van Eck, N.J., van Putten, D., Van den Hoven, J.: From dignity to security protocols: A scientometric analysis of digital ethics. Ethics Inf. Technol. **20**, 175–187 (2018)
109. Boyd, D., Crawford, K.: Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Inf. Commun. Soc. **15**, 662–679 (2012)
110. Mühlhoff, R.: Human-aided artificial intelligence: Or, how to run large computations in human brains? Toward a media sociology of machine learning, New Media & Society (2019) 1–17.
111. Couldry, N., Mejias, U.A.: Data colonialism: Rethinking big data's relation to the contemporary subject. Televis. New Media **20**, 336–349 (2019)
112. Olteanu, A., Castillo, C., Diaz, F., Kıcıman, E.: Social data: biases, methodological pitfalls, and ethical boundaries. Front. Big Data **2**, 1–33 (2019)
113. Brynjolfsson, E., McAfee, A.: The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. W. W. Norton & Company, New York (2014)
114. Irani, L.: Justice for data janitors. In: Marcus, S., Zaloom, C. (eds.) Think in Public, pp. 23–40. Columbia University Press, New York (2019)
115. Cognilytica, Data Preparation & Labeling for AI 2020, 2020, pp. 1–37. https://www.cognilytica.com/download/data-preparation-labeling-for-ai-2020-cgr-dlp20/ (accessed 22 June 2021).
116. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for Data: Ground Truth from Computer Games. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision—ECCV 2016, pp. 102–118. Springer International Publishing, Cham (2016)
117. Lucic, M., Tschannen, M., Ritter, M., Zhai, X., Bachem, O., Gelly, S.: High-fidelity image generation with fewer labels, arXiv (2019) 1–23.
118. Irani, L.: The cultural work of microwork. New Media Soc. **17**, 720–739 (2015)
119. Casilli, A.A.: Digital labor studies go global: Toward a digital decolonial turn, international. J. Commun. **11**, 1934–3954 (2017)
120. Horton, J.J., Chilton, L.B.: The Labor Economics of Paid Crowdsourcing, in: D.C. Parkes, C. Dellarocas, M. Tennenholtz (Eds.), Proceedings of the 11th ACM conference on Electronic commerce, ACM, Cambridge, 2010, pp. 209–218.
121. Hao, K.: An AI startup has found a new source of cheap labor for training algorithms: prisoners, 2019. https://www.technologyreview.com/2019/03/29/136262/an-ai-startup-has-found-a-new-source-of-cheap-labor-for-training-algorithms/ (accessed 1 July 2021).
122. Graham, M., Hjorth, I., Lehdonvirta, V.: Digital labour and development: impacts of global digital labour platforms and the gig economy on worker livelihoods. Transfer **23**, 135–162 (2017)
123. Fang, L.: Google hired gig economy workers to improve artificial intelligence in controversial drone-targeting project, 2019. https://theintercept.com/2019/02/04/google-ai-project-maven-figure-eight/ (accessed 13 February 2019).
124. Bederson, B.B., Quinn, A.J.: Web workers, Unite!: Addressing Challenges of Online Laborers, in: Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11, ACM Press, New York, 2011, pp. 97–101.
125. Wood, A.J., Graham, M., Lehdonvirta, V., Hjorth, I.: Good gig bad gig: Autonomy and algorithmic control in the global gig economy, work. Employ. Soc. **33**, 56–75 (2019)
126. Healy, J., Nicholson, D., Pekarek, A.: Should we take the gig economy seriously? Labour Ind. **27**, 232–248 (2017)
127. Prassl, J.: Humans as a Service: The Promise and Perils of Work in the Gig Economy. Oxford University Press, Oxford (2018)
128. S. Ruggles, S. Flood, R. Goeken, J. Grover, E. Meyer, J. Pacas, M. Sobek, IPUMS USA: Version 8.0, 2018.
129. Lingel, J., Crawford, K.: Notes from the desk set. Catalyst **6**, 1–22 (2020)
130. Frey, C.B., Osborne, M.A.: The future of employment: How susceptible are jobs to computerization. Oxford Martin Programme on Technology and Employment, 2013, pp. 1–78.
131. Light, J.S.: When computers were women. Technol. Cult. **40**, 455–483 (1999)
132. Owe, A., Baum, S.D.: Moral consideration of nonhumans in the ethics of artificial intelligence. AI Ethics **2**, 1–12 (2021)
133. Keim, B.: How automated vehicles could save millions of animal lives, 2017. https://www.anthropocenemagazine.org/2017/12/automated-vehicles-and-animals/ (accessed 15 November 2021).
134. Hassabis, D., Kumaran, D., Summerfield, C., Botvinick, M.: Neuroscience-inspired artificial intelligence. Neuron **95**, 245–258 (2017)
135. Crosby, M., Beyret, B., Halina, M.: The animal-AI olympics. Nat. Mach. Intell. **1**, 257 (2019)
136. Gendin, S.: The Use of Animals in Science. In: Regan, T., Singer, P. (eds.) Animal Rights and Human Obligations, pp. 197–208. Prentice-Hall, Englewood Cliffs (1989)
137. Singer, P.: Animal Liberation. HarperCollins Publishers, New York (1975)

138. Rosenblatt, F.: The perceptron: A probabilistic model for information storage and organization in the brain. Psychol. Rev. **65**, 386–408 (1958)
139. Hubel, D.H., Wiesel, T.N.: Receptive fields of single neurones in the cat's striate cortex. J. Physiol. (Lond) **148**, 574–591 (1959)
140. Moore, T., Zirnsak, M.: Neural mechanisms of selective visual attention. Annu. Rev. Psychol. **68**, 47–72 (2017)
141. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (2018)
142. Strickland, E.: AI designers find inspiration in rat brains, 2017. https://spectrum.ieee.org/biomedical/imaging/ai-designers-find-inspiration-in-rat-brains (accessed 30 July 2020).
143. Nishiyama, J., Yasuda, R.: Biochemical computation for spine structural plasticity. Neuron **87**, 63–75 (2015)
144. Carter, M., Shieh, J.C.: Stereotaxic Surgeries and In Vivo Techniques. In: Carter, M., Shieh, J.C. (eds.) Guide to Research Techniques in Neuroscience, pp. 73–90. Elsevier, London (2010)
145. Morgan, K.N., Tromborg, C.T.: Sources of stress in captivity. Appl. Anim. Behav. Sci. **102**, 262–302 (2007)
146. Khaligh-Razavi, S.-M., Kriegeskorte, N.: Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput. Biol. **10**, 1–29 (2014)
147. Sinz, F.H., Pitkow, X., Reimer, J., Bethge, M., Tolias, A.S.: Engineering a less artificial intelligence. Neuron **103**, 967–979 (2019)
148. Prescott, M.J., Lidster, K.: Improving quality of science through better animal welfare: the NC3Rs strategy. Lab Anim. (NY) **46**, 152–156 (2017)
149. Wayne, N.L., Miller, G.A.: Impact of gender, organized athletics, and video gaming on driving skills in novice drivers. PLoS ONE **13**, 1–12 (2018)
150. Russell, W., Burch, R., Hume, C.: The Principles of Humane Experimental Technique. Universities Federation for Animal Welfare, Potters Bar (1992)
151. Thompson, P.B.: Ethics on the frontiers of livestock science. In: Swain, D.L., Charmley, E., Steel, J., Coffey, S. (eds.) Redesigning Animal Agriculture: The Challenge of the 21st Century, pp. 30–45. CABI, Wallingford (2007)
152. Donaldson, S., Kymlicka, W.: Zoopolis: Eine politische Theorie der Tierrechte. Suhrkamp, Berlin (2013)
153. Palmer, C.: Animal Ethics in Context. Columbia University Press, New York (2010)
154. Regan, T.: The Case for Animal Rights. Routledge & Kegan Paul, London (2004)
155. McFarland, S.E., Hediger, R.: Animals and Agency: An Interdisciplinary Exploration. Brill, Leiden (2009)
156. Connolly, A.: Is artificial intelligence right for poultry production, 2019. https://www.wattagnet.com/articles/38540-is-artificial-intelligence-right-for-poultry-production?utm_campaign=The%20Batch&utm_source=hs_email&utm_medium=email&utm_content=83935678&_hsenc=p2ANqtz-82sdH078u2hpqxE2EMrXvdJ6PSklNJ3SUujcJsGu9p3H-9NdRlnsuB-EGezh_fRnxt_8eJG4gpFqYCqgE8sv9_86odyQ&_hsmi=83935678 (accessed 10 August 2020).
157. Stine, L.: French poultry tech startup Tibot Technologies raises €3m seed round for health-boosting robot, 2019. https://agfundernews.com/french-poultry-tech-startup-tibot-technologies-raises-e3m-seed-round-for-health-boosting-robot.html (accessed 10 August 2020).
158. Zhang, L., Gray, H., Ye, X., Collins, L., Allinson, N.: Automatic individual pig detection and tracking in pig farms. Sensors (Basel) **19**, 1–20 (2019)
159. Chong, Z.: AI helps grow 6 billion roaches at China's largest breeding site, 2018. https://www.cnet.com/news/ai-helps-grow-6b-roaches-at-chinas-largest-breeding-facility/ (accessed 10 August 2020).
160. Haladjian, J., Hodaie, Z., Nüske, S., Brügge, B.: Gait anomaly detection in dairy cattle. in: Proceedings of the Fourth International Conference on Animal-Computer Interaction, ACM, New York, 2017, pp. 1–8.
161. Carpio, F., Jukan, A., Sanchez, A.I.M., Amla, N., Kemper, N.: Beyond production indicators. in: Proceedings of the Fourth International Conference on Animal-Computer Interaction, ACM, New York, 2017, pp. 1–11.
162. Hansen, M.F., Smith, M.L., Smith, L.N., Salter, M.G., Baxter, E.M., Farish, M., Grieve, B.: Towards on-farm pig face recognition using convolutional neural networks. Comput. Ind. **98**, 145–152 (2018)
163. Braverman, I.: Zooveillance: Foucault Goes to the Zoo, SS 10 (2012) 119–133.
164. de Groot, R., Brander, L., van der Ploeg, S., Costanza, R., Bernard, F., Braat, L., Christie, M., Crossman, N., Ghermandi, A., Hein, L., Hussain, S., Kumar, P., McVittie, A., Portela, R., Rodriguez, L.C., ten Brink, P., van Beukering, P.: Global estimates of the value of ecosystems and their services in monetary units. Ecosyst. Serv. **1**, 50–61 (2012)
165. Steffen, W., Rockström, J., Richardson, K., Lenton, T.M., Folke, C., Liverman, D., Summerhayes, C.P., Barnosky, A.D., Cornell, S.E., Crucifix, M., Donges, J.F., Fetzer, I., Lade, S.J., Scheffer, M., Winkelmann, R., Schellnhuber, H.J.: Trajectories of the earth system in the anthropocene. Proc. Natl. Acad. Sci. U.S.A. **115**, 8252–8259 (2018)
166. King, A.D., Harrington, L.J.: The inequality of climate change from 1.5 to 2°C of global warming. Geophys. Res. Lett. **45**, 5030–5033 (2018)
167. Román-Palacios, C., Wiens, J.J.: Recent responses to climate change reveal the drivers of species extinction and survival. Proc. Natl. Acad. Sci. U.S.A. **117**, 4211–4217 (2020)
168. Portmess, L., Tower, S.: Data barns, ambient intelligence and cloud computing: The tacit epistemology and linguistic representation of Big Data. Ethics Inf. Technol. **17**, 1–9 (2015)
169. Joler, V., Crawford, K.: Anatomy of an AI system, 2018. https://anatomyof.ai/ (accessed 6 February 2019).
170. Spohr, M., Wolfrum, R., Danz, J., Renner, S.: Human rights risks in Minin: A baseline study. 2016.
171. World Bank, Renewable energy consumption (% of total final energy consumption): Sustainable Energy for All (SE4ALL) database from the SE4ALL Global Tracking Framework, 2021. https://data.worldbank.org/indicator/EG.FEC.RNEW.ZS?end=2015&start=1990&view=chart (accessed 12 July 2021).
172. Dhar, P.: The carbon impact of artificial intelligence. Nat. Mach. Intell. **2**, 423–425 (2020)
173. Strubell, E., Ganesh, A., McCallum, A.: Energy and policy considerations for deep learning in NLP. arXiv (2019) 1–6.
174. Schwartz, R., Dodge, J., Smith, N.A., Etzioni, O.: Green AI. arXiv (2019) 1–12.
175. Belkhir, L., Elmeligi, A.: Assessing ICT global emissions footprint: Trends to 2040 & recommendations. J. Clean. Prod. **177**, 448–463 (2018)
176. Mulligan, C., Elaluf-Calderwood, S.: AI ethics: A framework for measuring embodied carbon in AI systems. AI Ethics **2**, 1–13 (2021)
177. van Wynsberghe, A.: Sustainable AI: AI for sustainability and the sustainability of AI. AI Ethics **2**, 1–6 (2021)
178. Lacoste, A., Luccioni, A., Schmidt, V., Dandres, T.: Quantifying the carbon emissions of machine learning. arXiv (2019) 1–8.
179. Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S.D., Tegmark, M., Fuso Nerini, F.: The role of artificial intelligence in achieving the sustainable development goals. Nat. Commun. **11**, 1–10 (2020)

180. Greening, L.A., Greene, D.L., Difiglio, C.: Energy efficiency and consumption—the rebound effect—a survey. Energy Policy **28**, 389–401 (2000)
181. Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, E., Sánchez, A.N., Raji, D., Rankin, J.L., Richardson, R., Schultz, J., West, S.M., Whittaker, M.: AI Now 2019 Report, 2019. https://ainowinstitute.org/AI_Now_2019_Report.pdf (accessed 18 December 2019).
182. Merchant, B.: Amazon Is Aggressively Pursuing Big Oil as It Stalls Out on Clean Energy, 2019. https://gizmodo.com/amazon-is-aggressively-pursuing-big-oil-as-it-stalls-ou-1833875828 (accessed 12 July 2021).
183. Kirchgaessner, S.: Revealed: Google made large contributions to climate change deniers, 2019. https://amp.theguardian.com/environment/2019/oct/11/google-contributions-climate-change-deniers (accessed 21 July 2021).
184. Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mokander, J., Floridi, L.: Ethics as a service: A pragmatic operationalisation of AI ethics. Mind. Mach. **31**, 239–256 (2021)
185. Luhmann, N.: Trust and Power. Polity, Cambridge (2017)