AI and Ethics (2026) 6:528
https://doi.org/10.1007/s43681-026-01326-6

RESEARCH

What is AI ethics, and what is it good for? Clarifying debates on AI ethics

Astrid Schomäcker[1,2] · Timo Speith[1,2] · Luca Deck[2,3] · Thorsten Helfer[4] · Hanwei Zhang[5] · Lena Kästner[1,2]

Received: 15 April 2026 / Accepted: 11 August 2026 / Published online: 18 September 2026
© The Author(s) 2026

Abstract

The rapid integration of artificial intelligence (AI) into various aspects of society has raised pressing ethical concerns, from privacy and fairness to autonomy and safety. In response, the field of AI ethics has grown significantly, with a particular focus on developing guidelines to support the responsible development and use of AI. However, this trend has faced criticism, with concerns about "ethics washing" and skepticism regarding the overall value of AI ethics. In this paper, we argue that such criticisms often rely on an incomplete understanding of AI ethics. To address this, we propose an understanding of AI ethics that distinguishes between diverse *instances* of AI ethics (viz., research, education, practice, and governance instruments) and the *purposes* they serve (viz., knowledge, communication, and application). By disentangling these elements, we can illuminate different ways of doing AI ethics and provide a structured way to assess their value. Our arguments highlight that critiques of specific instances and purposes of AI ethics do not justify dismissing AI ethics altogether. We conclude by demonstrating how our proposed distinctions can reconcile existing defenses of AI ethics and guide future discourse, ensuring that the field remains both robust and relevant in addressing the ethical challenges posed by AI.

Keywords AI ethics · AI ethics guidelines · AI regulation · Critical perspectives on AI

✉ Astrid Schomäcker
astrid.schomaecker@uni-bayreuth.de

Timo Speith
timo.speith@uni-bayreuth.de

Luca Deck
luca.deck@uni-bayreuth.de

Thorsten Helfer
helfer@cispa.de

Hanwei Zhang
zhang@depend.uni-saarland.de

Lena Kästner
lena.kaestner@uni-bayreuth.de

1 Department of Philosophy, University of Bayreuth, Bayreuth, Germany

2 AI Responsibility Centre, University of Bayreuth, Bayreuth, Germany

3 Fraunhofer Institute for Applied Information Technology, Sankt Augustin, Germany

4 Helmholtz Center for Information Security, Saarbrücken, Germany

5 Department of Computer Science, Saarland University, Saarbrücken, Germany

1 Introduction

AI has become ubiquitous. It is driving our cars, writing our emails, deciding about our loans, predicting our behavior, and much more. Many of these applications pose risks to our ethical values, including privacy, safety, fairness, and human autonomy. As a society, we need to determine how to benefit from AI without accepting too many risks. To this end, a trend has emerged to assist programmers, companies, or even nations by formulating AI ethics guidelines that provide guidance on the morally responsible development and use of AI [1, 2].

However, this development has not been met with approval alone. In particular, the publication of AI ethics guidelines by companies has raised suspicions of "ethics washing" [3], which is comparable to "greenwashing," i.e., the use of labels for environmentally friendly production for economic rather than environmental reasons [4]. A related concern is that voluntary AI ethics guidelines might be used to make legal regulation of AI appear less pressing [5, 6].

Unfortunately, from these critical views of AI ethics *guidelines*, critiques have evolved that read as if they are criticizing AI ethics *in general*. Munn, for instance,

dismisses AI ethics altogether as "useless" [7]. Moving beyond academic debates, Andreessen's Techno-Optimist Manifesto has garnered significant attention for arguing that ethical challenges should be left to the free market and that tech ethics is "the enemy" [8].

While several authors set out to defend AI ethics in general and AI ethics guidelines specifically (e.g., [5, 9]), no coherent picture about the *nature* and *value* of AI ethics has yet emerged. Indeed, some defenses of AI ethics even appear to contradict each other. To remedy this situation, we believe, it is useful to understand AI ethics as a *systematic engagement with moral norms concerning AI systems* that can be achieved in different ways and for different ends. Therefore, this paper describes several important distinctions, both between different things that can be called "AI ethics" and between different *purposes* AI ethics might serve. While these distinctions are already implicit in many discussions of AI ethics, making them explicit sheds light on the complex nature of AI ethics and enables a structured assessment of its value.

Our arguments build on the observation that the term "AI ethics" can be understood in various ways. Crucially, we think, "AI ethics" may be understood narrowly—as a set of *moral norms* concerning (the use of) AI—or broadly—as different *ways of dealing with* such norms. Thus, the seemingly simple question "Is AI ethics useful or not?" is ambiguous: is it referring to a specific set of norms, or rather to the ways of dealing with these norms? After all, objecting that a *specific norm* proposed by an ethics board is not useful (e.g., because it is too vague to be implemented) is different from arguing that the *ethics board* itself is not useful (e.g., because developers do not follow its recommendations). AI ethics in the narrow sense has received a lot of attention, not only within the plethora of AI ethics guidelines [2], but also in philosophical discussions of AI ethics [10]. We think, however, that AI ethics in the broader sense deserves equal attention. There are many ways of dealing with moral norms regarding AI;[1] and, importantly, choosing a different way may lead to vastly different effects. Hence, discussing the usefulness of AI ethics requires clarity about the specific way of dealing with moral norms that is meant by "AI ethics" in a given context. We call these different ways of dealing with moral norms "instances of AI ethics." Thus, we believe it is time to ask, "Which instance of AI ethics is useful for what purpose?"

[1] Given the increasing ubiquity of AI, many moral issues might soon be, in some sense, "regarding AI" without thereby necessarily being considered an issue of AI ethics [11, 12]. Drawing the boundaries of AI ethics in the narrow sense, however, is not the goal of this paper, and we thus rely on an intuitive understanding of which moral norms are relevant.

The goal of our paper is to provide an answer to this question. Taking a bird's eye view of the contemporary discussions about AI ethics, we provide an overview of paradigmatic *instances* of AI ethics, examine what different *purposes* they might serve, and systematically relate the different instances and purposes.[2] We believe that the conceptual distinctions we provide deliver several important contributions to the discussion on the value of AI ethics. First, they help illuminate the nature of AI ethics by mapping the field and sharpening the concept of "AI ethics." Second, they show that being critical of some instance of AI ethics does not imply a reason to reject AI ethics altogether. As such, they help to defend the value of AI ethics. Third, our distinctions allow us to systematize the existing criticisms and defenses of AI ethics and resolve apparent disagreements between them. As such, these distinctions provide a valuable tool for systematically assessing criticisms of AI ethics, paving the way for a more constructive debate on AI ethics.

Finally, using our distinctions, we can showhow the purposes and instances interact with one another, which can guide considerations on how to best combine different instances of AI ethics to ensure that AI benefits society as a whole.

We start by introducing the current state of the debate (Sect. 2). Then, we distill the conceptual distinctions that underlie our arguments from current practices and debates about AI ethics. We carve out different general *purposes* that AI ethics may reasonably serve, namely *knowledge*, *communication* and *application* and discuss their relationship among each other (Sect. 3). Afterwards, we disentangle different *instances* of AI ethics. They include *research* (e.g., in a philosophical discipline), *education* (e.g., in academic teaching), *practice* (e.g., the integration of specific features into AI systems), and *governance instruments* (e.g., the conduct of ethical review boards). With these categories in mind, we will systematically examine the relationships between different instances and purposes (Sect. 4). Before we conclude, we show how these distinctions help structure the debate on the value of AI ethics (Sect. 5) and discuss some limitations and potential objections (Sect. 6).

2 Background and related work

To elucidate the need for conceptual clarity regarding the different instances and purposes of AI ethics, it is valuable to understand why AI ethics has recently come under attack,

[2] We do not claim that the instances and purposes we describe are unique to AI ethics. However, comparing AI ethics to other areas of applied ethics is beyond the scope of this paper. For some insights, see [13].

how it has previously been defended, and why other systematizations of AI ethics do not provide sufficient conceptual clarity to structure this debate.

Let us begin with the criticisms of AI ethics, which were provoked by the rise of AI ethics guidelines. Some authors have criticized the focus on principles in AI ethics guidelines (e.g., [13–15]). Mittelstadt, for example, warns that a principled approach to AI ethics faces technical and normative barriers that impede its translation into practice [13]. He argues that while principles in AI ethics might resemble those in medical ethics, AI development lacks the necessary structures to make these principles effective. Some critics have gone so far as to see these worries as a reason to abandon AI ethics altogether. For instance, Munn argues that AI ethics produces meaningless, isolated, and toothless principles and warns that the turn to AI ethics might be "a dangerous distraction." [7, p. 869]. Similarly, Hagendorff analyzes a large number of guidelines and concludes that "AI ethics is failing in many cases" [14, p. 113]. Denying the general value of AI ethics, tech mogul Andreessen calls tech ethics "the enemy" [8].

Several authors have defended AI ethics against this criticism. Most of them seem to agree, both among themselves and with the critics, that problems arise when AI ethics is narrowly focused on identifying ethical principles and codifying them into guidelines. However, their defenses reveal no consensus on what AI ethics is, nor on whether it is already effective or requires improvements to become so. Among the articles that argue that AI ethics is already effective, Rességuier and Rodrigues claim that ethics is "primarily a form of attention" [5, p. 2] and that it is (already) a powerful tool to detect changes in the world that require action. Principles are therefore "an end of ethics, not ethics itself" [italics removed, 5, p. 3]. Lundgren claims that "'AI ethics' is a term used in various disciplines to denote the study of ethical aspects of AI" [9, p. 1013] and thus argues that Munn's "critique is misguided because most of [it] is about practical enactment of policy" [9, p. 1014]. Additionally, he defends ethical guidelines against criticism by showing that many guidelines are more practice-oriented than Munn's criticism might suggest (see also [16]).

Others seem to assume that AI ethics needs improvement to be more useful. These approaches focus heavily on application contexts and often do not make explicit what they mean by "AI ethics". Morley et al. present tools to translate the "what" of ethical principles into a practical "how," which they consider to be "the second phase of AI ethics" [17, p. 2147]. This use of the term "AI ethics" is in tension with both [5] and [9], as it neither seems to refer to a form of attention nor to the study of ethical aspects of AI. Hagendorff argues that "[i]nstead of focusing only on principles, AI ethics can put a stronger focus on virtues or, in other words, on character dispositions in AI practitioners in order to effectively put itself into practice" [18, p. 2]. Again, he does not define "AI ethics" and his use of the term is neither coherent with [5] nor with [9].

The above examples unveil a lack of clarity about the nature and expected value of AI ethics within the debate. Given this lack of clarity, however, progress in utilizing AI ethics in the most effective ways is hampered. Against this background, we believe that some systematization is required to organize the various replies and enable readers to understand the relationship between different positions.

There are already various approaches to the systematization of AI ethics in the literature, including reviews, overview articles, and frameworks. They typically fall into one of three categories: First, there are reviews of AI ethics guidelines (e.g., [2, 16, 19]). These reviews typically focus on identifying principles demanded across guidelines, or, more generally, on identifying similarities and differences in the conceptualization of AI ethics across organizations and countries.

Second, there are introductory books and overview articles on the normative questions that arise regarding AI (e.g., [10, 20]). Such overviews usually discuss various aspects (e.g., opacity or bias) or areas of AI use that raise ethical debates (e.g., voice assistants [21] or autonomous weapon systems), or discuss which normative questions should be considered AI ethics (e.g., [11]).

Third, some frameworks aim to improve AI ethics, e.g., by identifying overlooked areas (e.g., [22, 23]) or by listing individual solutions for various issues (e.g., [12]). A notable example of such work is [23], which organizes research papers on AI ethics along three dimensions that together determine a paper's potential real-world impact.

While these systematizations have their value, they all focus on *one specific* way of dealing with moral norms regarding AI (i.e., only on guidelines or only on research questions). However, as we will argue below, organizing the debate about the value of AI ethics requires a look at *different* ways of dealing with moral norms regarding AI and understanding their relationship to different purposes that AI ethics can serve. Our distinctions provide clarity and thus help to systematically assess the value of AI ethics in its various different forms and identify ways to productively combine the different instances.

3 Purposes of AI ethics

Before we discuss the relationship between our instances and purposes of AI ethics, we introduce the three purposes and their relationship to one another (see the right side of

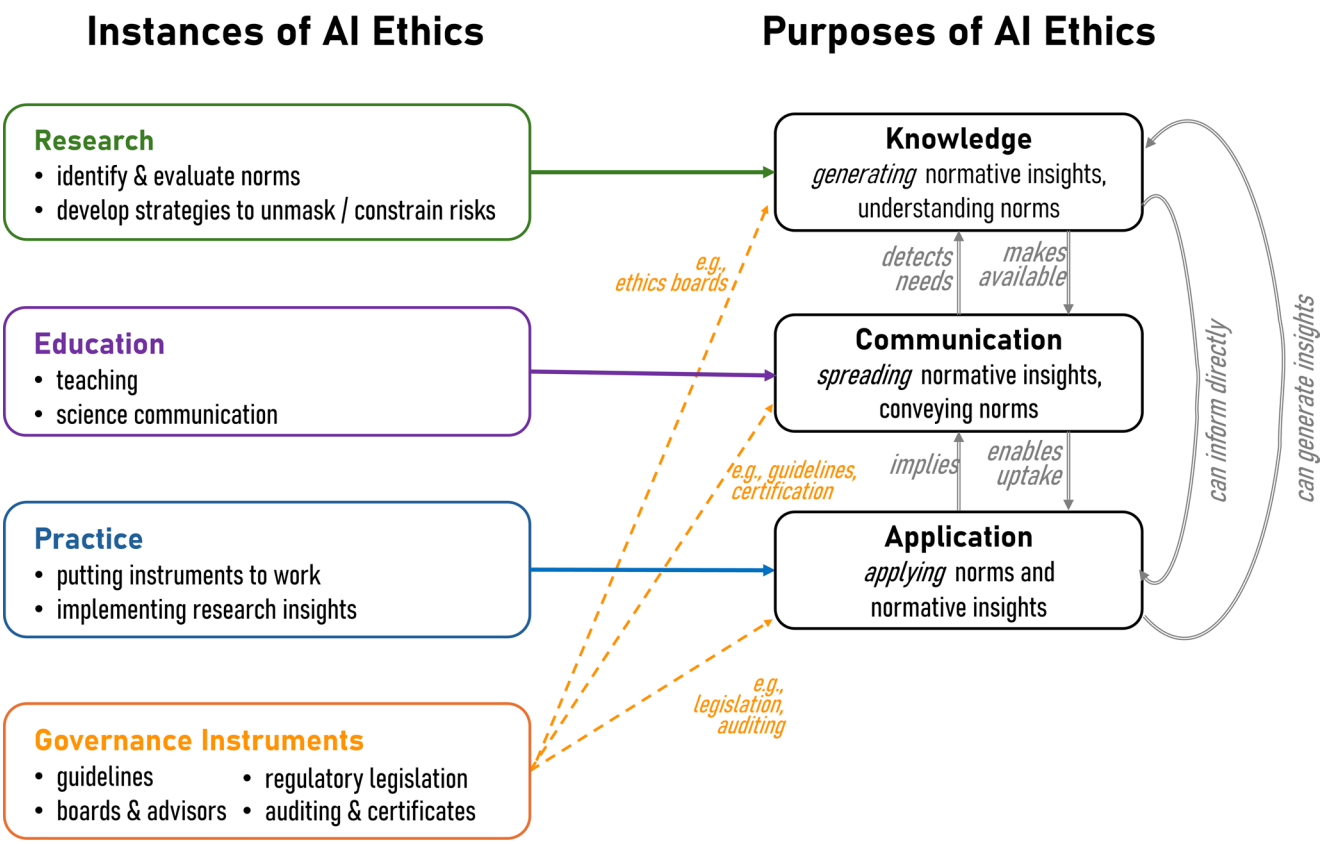

Fig. 1 Relationships between instances and purposes of AI ethics at a glance. Thick lines between instances and purposes signify conceptual relationships. Dashed lines indicate that different governance instruments can be used towards different purposes (supplementing other instances). Grey arrows indicate relationships between the different purposes; via these relationships, all instances can eventually contribute to all purposes

Fig. 1 for visualization)[3]: They are *application*, *knowledge* and *communication*.

3.1 Application

When we think of the criticism of AI ethics discussed in Sect. 2, much of it seems to assume that the purpose of AI ethics is what we call *application*: that is, to influence the development and deployment of AI to ensure it conforms with moral norms. By this, we mean that AI ethics is expected to shape how AI is designed (e.g., by requiring that data collection follows certain protocols) or to prevent undesired impacts of AI (e.g., by shutting down systems that do not comply with certain norms). While at first sight this might seem to be the main purpose of AI ethics, we argue that there are at least two further purposes of AI ethics that are related to application but that one can strive for independently.[4]

3.2 Knowledge

First, we can strive to increase our knowledge and understanding of normative aspects of AI, that is, engage with moral norms regarding AI towards an *epistemic purpose*. Such aspects can include, for instance, which ethical issues arise from AI, which moral norms hold regarding these issues,[5] which ethical beliefs about AI are held by different

[3] We believe that our characterization of instances and purposes, respectively, is largely exhaustive. However, even if further instances or purposes might be identified, the instances and purposes we discuss here are conceptually distinct and stand in systematic relationships to one another.

[4] To those invested in the philosophy of purposes: We stay neutral on whether there is a singular overarching purpose in everything we do, as discussed in some philosophical debates (classically [24, ch.1]), and about the metaphysical nature of purposes, including their exact relationship to individual people's intentions and the right manner to individualize purposes. Neither of these questions has bearing on the main messages of the current paper.

[5] In this paper, we stay neutral concerning different realist and anti-realist views of moral norms as discussed in metaethics, and we do not discuss what it takes to *know* that something is morally right or wrong. Nevertheless, we take it to be evident that gaining such knowledge is often the purpose of different instances of AI ethics.

people, or how norms like fairness can be operationalized for computing.

There is a close relationship between the epistemic purpose and the application purpose. AI developers cannot ensure that AI complies with certain norms if they are not aware of these norms. Thus, a proper fulfillment of the application purpose requires that the epistemic purpose has previously been fulfilled by some instance of AI ethics. Importantly, striving for the epistemic purpose can be considered valuable and successful even without the knowledge being directly put into application.

3.3 Communication

Second, some instances of AI ethics are primarily used for *communication*. By that, we mean any intention to inform someone else about an aspect related to ethical issues regarding AI. This involves, e.g., communicating that a person, group, or other entity adheres to or approves of a certain set of AI norms.

Some forms of communication can be similarly seen as a prerequisite of application as knowledge, and knowledge is a prerequisite to communication, insofar as communication is a transfer of knowledge from one person to another. In many cases, the experts that know about moral norms (e.g., ethicists) are not the people who can ensure their application (e.g., AI developers). Thus, those who know about moral norms need to communicate their knowledge to those who should apply them. Again, an instance can be successful at communicating about moral norms, and hence valuable, even when it does not directly ensure application. Other forms of communicating about moral norms are about transferring knowledge about the speaker rather than the norms, i.e. to communicate to others which moral norms the speaker endorses.

Two aspects about these purposes are important for our arguments: They are clearly conceptually distinct and they have different conditions for when they are fulfilled. We also believe that they are likely complete in the sense that they are superordinate categories that subsume many finer-grained purposes. For example, the purpose of adhering to fairness norms and requirements falls under the *application* purpose; the purpose of increasing awareness of the environmental impact of AI development falls under the *communication* purpose.

Based on these purposes, our understanding of AI ethics can be characterized as follows:

> AI ethics is the systematic engagement with moral norms concerning AI systems, aimed at understanding these norms, communicating them, or applying them to AI systems.

4 Prototypical instances of AI ethics and their relationship to the purposes

We now discuss four paradigmatic ways of dealing with moral norms regarding AI, for all of which we claim that they could reasonably be called "AI ethics" and which we call *instances*. Our four instances are: *AI ethics as research*, *AI ethics as education*, *AI ethics as a practice* and *AI ethics as governance instruments*. They all stand in different relationships to the three purposes introduced above. The first three instances constitute the prototypical way in which one of the purposes is achieved and their success is judged based on their success at fulfilling that purpose: research primarily aims at producing knowledge about moral norms; education primarily aims at communicating this knowledge; and practice primarily aims at using this knowledge for implementation in the design, development, or deployment of AI systems. This is a conceptual relationship insofar as understanding what, for instance, the concept "research" means includes understanding that research aims at knowledge. Regarding the other combinations of instances and purposes, there is no such conceptual relationship. Nevertheless, most instances can still be used with the intention to fulfill the other purposes; e.g., research can be conducted in the hope to reach application. We will discuss this in detail below.

The last instance, *AI ethics as governance instruments*, which includes ethics guidelines, ethics boards, ethical auditing, and similar approaches, serves as an umbrella term for several approaches that are used by companies, government bodies, and NGOs that deal with the ethical use of AI. We include them as they fit our understanding of AI ethics as described above and are regularly called "AI ethics" (e.g., in the debate on the value of AI ethics we introduced in Sect. 2). However, they are not conceptually related to one of the purposes. For them, it is less straightforward to see what value they are expected to have and how successful they are at achieving their purpose.

Importantly, these four instances are not mutually exclusive. Individual ways of dealing with moral norms regarding AI can fall into several of these instances; for example, a conference can be both a research event and an educational event. Nevertheless, the distinctions are useful to highlight various possible ways of dealing with moral norms regarding AI.

Let us look at these instances and their relationships to the purposes in turn. For our discussion, we consider two questions regarding these relationships to be interesting. First, is it fair to assume that an instance is used with the intention to achieve a specific purpose? This question is important as it seems unreasonable to criticise an instance for not achieving a purpose it was never meant to achieve. Second, do

we have reasons to believe that the instance can or cannot achieve that purpose? For both questions, our assessment is based on philosophical considerations, research experience in ethics and computer science, and explicit statements found in the literature. Our discussion proceeds as follows: First, we describe the instance, then we explicate how it relates to the prototypical purpose, and then we discuss the relationship to the other purposes. Figure 1 provides a systematic overview of the relationships between instances and purposes of AI ethics.

4.1 Research

First, we want to look at AI ethics as a *research* endeavor. As such, AI ethics may help to identify and evaluate relevant norms for the development and use of AI systems. In addition, it may seek to illuminate the nature of AI ethics (as in this paper), delve into projects addressing specific topics such as fairness and non-discrimination in AI systems [25, 26], or aim to develop strategies to constrain potential artificial superintelligence [27, 28]. Given these broad topics, AI ethics as a research endeavor can be undertaken in research institutions, as well as in companies and government institutions. Often, though not always, the results of research will be published in academic journals or discussed at scientific events.

Existing research projects on AI ethics can be situated on a spectrum from purely conceptual projects to those explicitly aimed at influencing implementation decisions. On one end of that spectrum, we find research serving primarily to explore; this work is often conceptual in nature and is usually conducted out of scientific curiosity (e.g., [29]). For example, researchers interested in AI ethics might study which ethical questions are raised by AI, to what extent ethical frameworks can capture these issues, or how ethical theories must be revised in light of insights gained from their application to AI. On the other end of the spectrum, we find work aiming to effectively guide practices (e.g., [30–32]). AI ethics of this kind aims to identify normative principles that AI developers and deployers should adhere to when deploying AI systems in the real world, lessons from political philosophy to questions on algorithmic fairness [32].

Given the close relationship between research and the striving for knowledge, we consider it fair to assume that most researchers aim to fulfill the *epistemic purpose*, and sometimes that is their only goal. The amount of knowledge about moral norms that is generated in professional research practices provides strong reasons to believe that research can fulfill the epistemic purpose.

Some activities within research also have a *communicative purpose*, e.g., writing research papers or presenting results at conferences. In these activities, researchers communicate their insights to one another to further epistemic progress. When a research project aims to fulfill the communication purpose in a way that enables application, the goal should be to communicate insights to, e.g., AI developers. The important role of research papers and conferences in research in general suggests that, within these communities, these activities can fulfill the purpose. However, when communication is targeted at lay people, misunderstandings seem more likely and it is important to adapt the content and presentation for the target audience. Furthermore, when companies employ AI ethics researchers, this can be interpreted as an act of communication about the company, signaling that the company is committed to the ethical use of AI. Such communication is more likely to be successful if the company is overall seen as trustworthy and their actions are in line with their claims.

At least some AI ethics research endeavors are directly aimed at *application*. This is most obvious for research on the side of the spectrum that is developments in AI. But even research on very general, conceptual questions might be done in the hope that the insights that are gained will eventually influence application. In reality, however, there often seems to be a disconnect between ethics research and AI development. This is likely due to the fact that for research to fulfill the application purpose, several factors need to be fulfilled: The research needs to attract practitioners' attention, the insights need to be communicated in a way that is accessible to them and motivates them to adhere to the suggested ethical principles, and preferably, the insights should be easily operationalizable. In many cases, at least one of these factors is missing.

4.2 Education

Our second instance of AI ethics is AI ethics *education*. Nowadays, courses on ethics and technology are ubiquitous and tailored to a variety of audiences. The most common courses are in high school or at higher education institutions, teaching anyone (from developers to the general public) who engages with AI systems at any stage of its lifecycle [33, 34]. These courses often address existing norms and/or inform about relevant ethical issues surrounding AI use, development, and research to increase awareness of ethical issues.[6]

Given the close relationship between education and communication, we can assume that the standard purpose of offering AI ethics education is to *communicate*. Teaching AI ethics communicates the moral problems or moral norms identified in ethics research as well as the methods to acquire further ethical insights to the students. These

[6] For a list of worldwide courses, syllabi, and their analysis, see [35].

students might be (future) developers, policymakers, or users of AI, some of which might have a chance to ensure the application of the norms. High-quality AI ethics classes should be well-equipped to fulfill this communicative purpose, though problems can occur when students are not motivated to learn or when ideas need to be greatly simplified and students cannot grasp the full picture.

Sometimes, AI ethics education also seems to be offered to communicate something beyond the content of the classes: An institution that offers AI ethics courses might want to communicate to other institutions that AI ethics education is relevant and important, and that they should follow suit. Whether AI ethics as education is successful at communicating these aspects likely depends on how trusted the institution offering the courses is, how much of a role model it is for others, and how it acts otherwise.

For the participants of AI ethics classes, the main purpose seems to be *epistemic*. Many students, especially when participation is voluntary, will attend such classes to answer questions they have about AI ethics. When teaching is done well, this purpose should often be fulfilled. Furthermore, teaching AI ethics can have an epistemic purpose for the teachers as well. It is quite uncontroversial that the process of preparing material to teach provides learning gains [36].[7] Therefore, teaching AI ethics likely fulfills the epistemic purpose for both students and teachers.

Lastly, education can be intended for *application* when the goal is to enable developers, users, and policymakers to make better choices when dealing with AI. At least since the 1970s, it has been demanded that computer scientists be educated in the ethical issues surrounding their field [38], and with the rise of AI, demands for a general education in AI ethics have become louder [39]. It is an empirical question whether education can fulfill this purpose, i.e., whether AI ethics teaching leads to more ethical AI. Arguably, success will depend on a number of factors, including teacher training, student motivation, and the resistance developers will encounter in practice. For a guide on best practices within such courses, see, e.g., [40, pp. 6–8].

4.3 Practice

AI ethics, as the practice of following moral and ethical norms, consists of the development, deployment, and use of real-world AI systems, where the norms in question may be considered non-functional requirements [41, 42] on a given AI system. Implementing norms or values in practice requires different strategies depending on where the norms come from and how they are described. In some cases, it is clear how a norm should be implemented. In particular, governance instruments often require compliance with unambiguous standards. For example, the EU AI Act stipulates the necessity of a stop button for high-risk AI systems (*wrt.* Article 14-4 (e)), along with requirements for a comprehensive technique report and evaluations.

In most cases, however, developers are required to design systems that align with non-binding guidelines or values, such as privacy and fairness. In such cases, practitioners must assess ethical risks and address relevant norms throughout the AI lifecycle [33, 34, 43, 44]. These norms are diverse, complex, and often interdependent. Therefore, the process requires careful consideration of which norms to follow, to what extent, and under what conditions. Using fairness as an exemplary norm, a plethora of (conflicting) formal fairness metrics is available [45, 46], requiring developers to make normative and economic tradeoff decisions. This process is neither technically nor normatively straightforward and usually requires a range of programming tools as well as normative assumptions [47]. As best practices, toolboxes, and services tailored to facilitate the implementation of AI ethics are burgeoning (e.g., [48–51]), the implementation of ethical norms becomes more convenient for practitioners.

The prototypical purpose of AI ethics as *practice* is the *application* of moral norms to AI. Those who apply moral norms in practice are the ones who are directly involved in the conception, development, deployment, etc. of AI systems and thus have the ability to shape the nature of such systems. On the one hand, practitioners are striving to match normative requirements within the development of a specific product. On the other hand, practitioners are continuously developing broadly applicable toolkits to operationalize norms such as fairness (e.g., [49, 50]). However, effectively fulfilling the application purpose faces several gaps in the process of translating high-level principles into concrete instantiations [13, 52], which show the importance of fulfilling the other purposes as a prerequisite of the application purpose. First, the work environment has to foster awareness of the importance of norms and allow to engage in discourse about the norms to be applied. Second, there needs to be consensus about the selection of norms and their interpretation in the specific context. Both of these gaps can be filled by better communication. Third, appropriate tools need to be identified and employed to effectively translate the selected norms into practice, which requires knowledge about the relationship between moral norms and specific use cases. And finally, it should provide robust processes to evaluate the successful operationalization [53]. Closing these gaps is an ongoing challenge [52].

Given that knowledge on how to operationalize moral norms or principles is often lacking, it can be part of AI

[7] It is not as clear whether the teaching process itself is as beneficial for learning as material preparation since teaching contexts can vary immensely [37]. In any case, many educators who teach AI ethics will have to do both.

ethics as a practice to aim for the *epistemic* purpose. Often, practitioners need to pick up certain norms and explore how they might be implemented in AI systems. Practitioners usually have the relevant expertise to determine what a useful operationalization should look like, which is important for fulfilling this purpose. However, their operationalization might fail to track the relevant norms if they lack an understanding of the ethical issues at hand. Additionally, practice might also identify novel ethical challenges and consequences emerging from specific AI applications. For example, the challenges of algorithmic fairness have raised awareness and re-contextualized ethical challenges such as intersectionality [54, 55] and equality of opportunity [32, 56]. AI artifacts may even give rise to novel ethical challenges, such as representational harms in computer vision and generative AI [57, 58]. In both cases, feedback loops from practice to research are highly valuable. Researchers scrutinize whether norms are successfully operationalized in real-world applications, and practitioners provide crucial hands-on perspectives that might contribute to a better understanding of novel or existing norms.

An aspect of practice with a *communicative* purpose is documenting data characteristics, design decisions, and model evaluations [53]. Loi et al. [59] argue for "design publicity," which subsumes value, translation, and performance transparency as key tenets to communicate the rationales behind the selection of norms, how they are considered in the AI system, and a verification thereof. Such communicative efforts should fulfill the communication purpose at least within the community of AI practitioners, but can also inform other stakeholders such as NGOs and affected parties.

4.4 Governance instruments

We now turn to the last instance of AI ethics that we want to discuss: AI ethics as governance instruments. Several different approaches can be subsumed in this category. The probably most discussed governance instruments are ethics guidelines published by various bodies or consortia. They tend to describe the norms and principles that the authors consider important for (some kinds of) AI and provide suggestions for best practices. Beyond that, governance instruments include ethics boards or ethics advisors, ethical auditing, and regulatory legislation. Examples of the latter include the General Data Protection Regulation (GDPR) [60], the Digital Services Act (DSA) [61], and the EU AI Act [62].

Why do we still need governance instruments if there is, for every purpose we introduced, an instance that is aimed at that purpose and generally suited to fulfill it? One could hope that a proper interplay between research, education, and application would make governance instruments superfluous, because research and education would provide generally shared, clearly operationalizable norms that AI should adhere to and those working in practice would be motivated to adhere to them merely by recognizing their moral value. However, there are three main problems. First, different people, including different ethics researchers, often disagree about what is morally right. Second, ethics research often provides general principles ("AI should be fair") that still need to be operationalized for a specific application context. Third, the relationship between morality and motivation is complex. While most humans are motivated from early childhood on to act pro-socially and thus in line with the moral norms of their society [63], they can disagree with the moral beliefs of others and they can act against what they themselves consider to be morally right. Thus, one person (e.g. an ethicist or politician) holding a moral belief and communicating it to another person (e.g. an AI developer) does not guarantee that the other person will adopt that moral belief and apply it in practice.

Against this backdrop, we conceive of governance instruments as complementing other instances either to determine and communicate which moral norms a specific group of people endorses, to find and communicate operationalizations of general moral norms, or to provide external motivation to comply with a specific set of norms. External motivation can be provided by combining considerations about moral norms with *enforcement mechanisms*. Such enforcement mechanisms can be of different strengths, with the legal system often being considered the strongest. Given that the different governance instruments differ starkly in their nature, we will take ethics boards, guidelines, ethical auditing, and legislation as examples and examine them in turn.

4.4.1 Ethics boards and ethics advisors

AI ethics boards or AI ethics committees have been installed by several private companies (such as IBM[8]) and public institutions (such as universities). Ethics boards can be composed of members with various backgrounds from philosophy or sociology to law to computer science or engineering (see [64]). These boards can play different roles within the organizations they belong to: They may develop guidelines (as described below), evaluate specific projects or products (e.g., through an audit), or advise practitioners on the ethical dimensions of their developments [65]. Sometimes, but not always, ethics boards are given the authority to intervene in the development of AI. Instead of full ethics boards, some institutions have individual ethics advisors [65]. Ethics

[8] https://www.ibm.com/impact/ai-ethics.

Boards can also exist outside institutions that are actively involved in the development of AI, such as the "Data Ethics Commission" installed by the German Federal Government.

For ethics boards, different assumptions seem plausible about which purpose they are meant to serve. First, we can assume that the purpose of some ethics boards is to determine the norms that AI in a specific organization should adhere to. This is the *epistemic purpose*. Whether ethics boards can fulfill this epistemic purpose will likely depend on several factors, including the ethical and technical expertise of the members of the ethics board [66]. Ethics boards within companies usually have at least one epistemic advantage compared to academic researchers: Their access to details about new developments enables them to address specific ethical questions.

When in an advisory role, one of the purposes of AI ethics boards or advisors seems to be to *communicate* the knowledge of ethical experts to the people who shape AI. We consider it safe to assume that ethics boards will, in most cases, be able to communicate their normative beliefs successfully, as long as their statements are perceived by the people they intend to reach. If the purpose is to communicate something about the organization the ethics board is a part of, e.g., the ethical values that it holds, fulfilling the communicative purpose will also depend on the organization's overall reputation.

Another way to understand the purpose of ethics boards is in the *application* of the specified norms in practice. Aside from cases of intentional ethics washing (see Sect. 5), we can assume that ethics boards are installed in the hope of ensuring that AI products are built according to the specified ethical principles (as expressed, e.g., by Google's responsible AI team[9]). Their chances of success likely depend on the exact nature of the ethics board. Ethics boards or ethics advisors in direct contact with developers are in a good position to find operationalizable versions of various normative requirements. By contrast, ethics committees that develop general principles or guidelines often articulate normative requirements that are not immediately operationalizable. As for motivation, a lot depends on the ethics board's institutional role. Ethics boards with a mere advisory role will likely affect application less than a board that has the power to intervene in the development and deployment of a product.

4.4.2 Guidelines

In recent years, numerous AI ethics guidelines have been published. The challenges that arise from quickly evolving AI have motivated public and private institutions, including the European Commission, companies like Google and SAP, and non-profits like the Association of Computing Machinery (ACM) to develop guidelines on the ethical development and use of AI (see [2]). Such guidelines describe the normative principles and moral values that should serve as the basis for the development and use of AI systems, as envisioned by those issuing the guidelines.

Regarding guidelines, their primary purpose could be seen either in the communication of norms or in the application of norms. The publishing of a guideline is an act of *communication*. A guideline developed by a company communicates which ethical principles the company endorses. Guidelines published by organizations like the ACM communicate to others which norms and values these organizations endorse. When guidelines are written clearly and read by the target audience, we would expect them to fulfill that purpose. If the purpose is to communicate something about the publishing entity, e.g., that the related company places a high value on ethical standards in general or on the standards described in the guidelines specifically, success will also depend on the company's overall reputation.

Additionally, many might hope that the publication of guidelines will lead directly to the *application* of the specified norms. As with ethics boards, fulfilling that purpose requires at least that the norms are operationalizable and that AI practitioners are motivated to implement them. However, a lot of criticism has focused on the fact that many guidelines contain principles like "AI should be fair," which are hard if not impossible to operationalize because of the numerous different ways in which such a claim can be understood [45, 47]. Regarding motivation, the hope that AI practitioners are intrinsically motivated to develop products in line with any guidelines might be futile when competing motivations or pressures are at play. Within companies, the relevant motivation can be provided by embedding the guideline in appropriate structures, e.g., by installing an ethics board and giving it the authority to oversee compliance. This would mean adding an *enforcement mechanism* to the guideline.

In principle, it is also conceivable that bringing together a group of different experts to write a guideline has an *epistemic* purpose. Especially when people with different backgrounds come together to write a guideline, they contribute different perspectives, and we can thus assume that new insights about the normative beliefs of these different stakeholders will be gained.

4.4.3 Ethical auditing and ethical certification

A further instrument that deals with ethical issues surrounding AI is ethical audits. An ethical audit, similar to a financial or technical audit, is a structured process to make an

[9] https://research.google/teams/responsible-ai/

AI's ethical properties transparent or to ensure that an AI adheres to a specific ethical guideline [67]. Ideally, it should be conducted by an independent body that is not involved in the AI's development. In principle, this independent body can be within the organization (e.g., an ethics board) or outside it. External auditors can, e.g., be other companies or even government bodies. The outcome of an external audit can be an ethical certification.

The main purpose of requiring or offering ethical audits and certificates is likely the *application* of certain norms. Audits and certificates can increase motivation to adhere to ethical norms, especially when audits are mandatory or when certificates promise higher economic revenue. It has to be noted that this is a different motivation than the intrinsic motivation to adhere to ethical norms, with the effect that norms might be adhered to rather by the letter than in spirit. Another issue regarding fulfillment is that audits and certificates must rely on previously agreed-upon principles (e.g., a guideline) and thus any issues with operationalization within the guideline will affect the quality of the audit.

For those who undergo the auditing process or aim to acquire the certificate, another purpose can be to *communicate* the ethical properties of their AI products. With audits and certificates, the likelihood that such a communication about the company or the product is successful depends on the reputation and trustworthiness of the issuing entity (and thus less on the reputation of the entity that is audited than for other instances). They thus provide more objective information about the moral norms a company adheres to. There is, however, some potential for miscommunication when people do not know the criteria for passing the audit or receiving the certificate [68].

In order to fulfill the communicative purpose, some aspects of the auditing or certification process need to fulfill the *epistemic* purpose, i.e., they need to determine the ethical properties of an AI. Here, success mainly depends on whether the auditing mechanism can track the properties it is intended to uncover.

4.4.4 Legislation

By our conception of AI ethics, legislation also falls under AI ethics. While the exact relationship between legal norms and ethical norms is debated [69], ideally, ethical considerations should at least be the basis of any development of AI regulation.[10] In contrast to guidelines, laws are issued by the government and are binding for all parties working on AI within the respective country or region. However, the processes that lead to new legislation are often much slower than those leading to new guidelines, so legislation is usually unable to respond quickly and flexibly to new developments in AI. Whether or not one agrees that the law should be considered a form of ethics, it is interesting to compare it to other governance instruments.

The most obvious purpose of legislation seems to be *application*. Legislation combines the moral norms it describes with strong enforcement mechanisms, i.e. sanctions for a failure to adhere to the norm. Therefore, legislation can be seen as the most likely governance instrument to fulfill this purpose, as the sanctions provide strong motivation for most developers. This does not mean, however, that the law cannot be broken, and that people cannot attempt to find legal loopholes that allow them to continue to act against the norms (e.g., by moving to other countries). Additionally, it has to be noted that law-making has the same issues with operationalizability as any other attempt to provide general principles.

Some people also see a *communicative* purpose in legislation. For example, Dragoş Tudorache, co-rapporteur for the European Parliament, described the EU AI act as a means to communicate to other lawmakers in the world how certain ethical issues can be approached (see [70]). When perceived by the intended recipients, this purpose should usually be fulfilled.

Some practices within the process of developing new legislation have an *epistemic* purpose: understanding the relationship between new risks stemming from AI and existing legislation, and identifying new norms that can and should be incorporated into legislation. Though not perfect, the legislative process is built to fulfill that purpose.

[10] To many, legislation might rather appear to be an alternative to AI ethics (see, e.g., the description of ethics as "soft law" [2, p. 396]). However, it is a way of dealing with moral norms regarding AI and thus fits our conception. We also think that providing a definition that allows to distinguish between legislation and other forms of governance that are readily described as AI ethics (e.g., an AI ethics board with the power to stop development) without *ad hoc* excluding legislation would be a complex task that does not add to the goal of our paper. Various conditions that might provide a plausible distinction, e.g., differences in the relationship to moral facts, voluntary adherence, or potential consequences, would invite counterarguments. Consider, e.g., the suggestion that adherence to moral norms is voluntary while adherence to legal norms is not: One can still choose to break the law and live with the consequences. Also acting against widely accepted moral norms is not without consequences; think, e.g., of boycotts of companies which are seen as acting unethically.

5 Our distinctions at work

With our conceptual distinctions in view, let us revisit the criticisms of AI ethics discussed in Sect. 2. As we will show, our distinctions help structure the debate on AI ethics more clearly: we can specify which instances and purposes are being targeted by what criticisms. Thus, we can reveal that none of the criticisms discussed affects AI ethics as a whole (Sect. 5.1), recognize the dangers of discrediting AI ethics (Sect. 5.2), and show how seemingly contradictory replies to criticisms can be reconciled (Sect. 5.3).

5.1 Criticisms of AI ethics

5.1.1 Ethics Washing

A first criticism of AI ethics that can be clarified with the help of our distinctions is the claim that AI ethics is used for "ethics washing" [3, 66, 71, 72]. For example, big tech companies have been criticized for employing ethics departments that have no influence on the products that these companies develop [7]. The worry is that companies do this merely to appear more ethical. One goal of this strategy might be to convince legislators that no binding regulation for AI is needed, as the companies already pay sufficient attention to the ethical implications of their products. If this strategy is successful and political actors refrain from more regulation, this could lead to a problematic shift of power: rather than democratically elected leaders, big tech companies decide how AI shapes our society (see [73]).

Using our distinctions, we can clarify that this criticism is typically directed at the use of AI ethics for a particular variant of the *communication purpose*. More specifically, ethics washing occurs when companies employ governance instruments (e.g., guidelines and ethics boards), research (e.g., from "Responsible AI" groups), or education (e.g., company internal ethics training) primarily to communicate something about themselves, e.g., that they place a high value on building AI ethically. Often this communication is insincere, i.e., the emphasis on ethics is not accompanied by a genuine commitment to prioritizing ethical considerations over competing economic goals. This may involve describing moral norms through a lens biased toward the company's benefits or simply failing to act in accordance with those norms when they conflict with other corporate objectives. What is problematic, then, are (i) the deceptive elements of the communication and (ii) the likely mismatch between the moral beliefs these actors genuinely hold, those they communicate, and those they ultimately put into practice.

Ethics washing becomes particularly problematic when consumers take such claims at face value and fail to examine whether a company actually lives up to its moral claims (e.g., whether a company that claims to value privacy has implemented appropriate measures to protect its customers' data). At a broader level, successful ethics washing may also contribute to a lack of regulation when governments believe that big tech companies are sufficiently guided by their moral compass to ensure ethical AI, and thus assume that further enforcement mechanisms, like legal regulation, are unnecessary.

Though ethics washing as described above can be an issue, this does not imply that all AI ethics is ethics washing. For one thing, even the kind of communication *sometimes* utilized for ethics washing (i.e., communicating the moral principles a company endorses) can be valuable if it is done earnestly. For instance, it can support consumers in choosing companies whose moral commitments align with their own values. For another, conveying an endorsement of moral norms is not the only purpose of AI ethics (see Fig. 1), neither within companies nor in society in general, and sincerely working towards the other purposes remains valuable. Certain approaches are not likely to be used for ethics washing. For instance, state-funded AI ethics research in philosophy departments aims to gain unbiased insights into moral norms, using approaches like Rawls' "veil of ignorance" to balance the needs of all involved stakeholders [74]. AI ethics boards in NGOs or AI ethics guidelines published by state actors are parts of a democratic process in which moral norms are identified and communicated and can be important precursors to legislation. These instances neither aim at reducing political influence on AI development nor should their values be questioned when companies attempt to reduce this influence in the name of "ethics".

5.1.2 Uselessness

A related criticism is that AI ethics is not successful at shaping AI in the desired way, even when it is done sincerely. This criticism is present in [7], calling AI ethics *toothless*, *meaningless*, and *isolated*, and in the arguments against a principled approach to AI ethics [13] (see Sect. 2). It has fueled comments such as AI ethics being "useless" [7] or "failing" [14, p. 113], which question the general value of AI ethics.

Using our distinctions, we can again clarify that this criticism is directed only at a specific instance of AI ethics being used for a certain purpose, and is better captured as follows: AI ethics *guidelines* [7, 14], as an instance of AI ethics, often fail the *application purpose*. When specified in this way, it becomes clear that this criticism does not provide a reason to question the value of AI ethics as a whole. When it turns out that one instance of AI ethics is not useful for the purpose we are looking to fulfill, we should instead aim to find a more effective instance of AI ethics.

5.2 The danger of discrediting AI ethics

We believe that the clarity we aim to provide with our distinctions has not only theoretical benefits but also practical relevance. In any criticism of AI ethics, it is essential to be clear about which instance is criticized regarding which purpose. Otherwise, if people read about the "uselessness of AI ethics," there is a risk of discouraging people from doing any AI ethics and providing justification to those who would prefer to ignore ethical issues of AI altogether. Bietti calls this practice "ethics bashing" [3].

We see two potential effects of such a development away from AI ethics. Either companies might let "the market reign" as suggested by [8], meaning that ethical issues with AI will only be addressed when and where they coincide with economic gains. Alternatively—and in line with the criticism that the use of AI ethics discourages proper legislation—some might focus on pushing for better AI legislation while reducing funding for ethics research, ethics boards, ethical education, or other non-legal approaches.

Both outcomes would be problematic. There is little hope that market pressures alone will lead to the ethical development of AI [13, 75]. Though the public's ethical outcries might constitute one form of market pressure, other incentives clearly run counter to developing AI ethically, e.g., the financial costs of optimizing AI for fairness [68, p. 1576]. Focusing only on legislation would also have its downsides. First, such a position ignores the fact that legislation does not spring from thin air. It relies on the ethical debates that happen within the different instances of AI ethics, in ethics research, education, ethics boards, or between ethical advisors and practitioners. Second, legislation can never react as flexibly to new technological developments as, for example, an internal ethics board can, because drafting and passing legislation takes time. Thus, even if instances of AI ethics do not guarantee satisfactory application of specified norms, they remain valuable as long as they have *some* influence.[11]

5.3 Reconciling defenses of AI ethics

In Sect. 2, we have shown that the existing debate on AI ethics presents a muddled picture of its nature and value, as different authors use the term "AI ethics" inconsistently. We have now argued (see Sect. 5.1.2) that our distinctions allow us to specify a prominent criticism (i.e., that AI ethics is "useless") as the claim that the instance of AI ethics guidelines is not apt for the application purpose. Now we will show that our distinctions can be used to systematize and reconcile the defenses against this claim by clarifying the instance and purpose to which each refers.

Lundgren's defense sticks to the *same instance* and *same purpose* of AI ethics as the critics (i.e., guidelines and application) and negates their claims [9]. He points out the efforts undertaken to ensure that guidelines do affect application.

Other defenses discuss *other instances* and the *same purpose* as critics, i.e., they suggest that instances of AI ethics beyond guidelines can be useful for application. Morley et al. can be taken to aim at developing better tools within the *practice* of AI ethics to translate the principles stated in guidelines for application. Their account [52] reads as the suggestion to turn to *governance instruments* like ethical auditing and external ethics advisors. Hagendorff's call for virtue ethics can be understood as a call to improve the ethics *education* of AI practitioners, thereby leading them to develop better AI [18]. Munn's own position—though he claims that we should move away from AI ethics towards what he calls "AI justice"—can ultimately be interpreted as a call for more *AI ethics research* and better *AI ethics practice* to fulfill the application purpose [7].

Finally, some defenses refer to both *other instances* and *other purposes*, e.g., by denying both that AI ethics should be equated with guidelines and that application is the primary purpose of AI ethics. We can understand Rességuier and Rodrigues in this way: when they call AI ethics a kind of attention that enables us to see new problems arise, they seem to focus on AI ethics *research* and see its usefulness for the *epistemic* purpose [5]. Similarly, when Lundgren points out that he takes "AI ethics" to refer to the study of ethical issues about AI and claims that, for this reason, Munn's criticism is misguided [9, p. 1014], he seems to equate AI ethics primarily with research.

Based on our distinctions, it becomes clear that these different defenses are not in contradiction with one another: instead, they show that the value of different instances of AI ethics can lie in how they contribute to different purposes. It is both helpful to point out that, for instance, AI ethics research provides important insights and is usually not directly aimed at application, and to search for the best ways to bring these insights into application. Regarding the application purpose, it can be helpful to explore different avenues to determine which instance of AI ethics is most effective, both overall and in specific contexts (see, e.g., [80] for research on how to combine different governance instruments). Our arguments show that ethics guidelines (as well as other instances of AI ethics) can be valuable even if they (alone) cannot guide application as desired. They might still serve the epistemic or the communication purpose, e.g., when ethical disagreements need to be navigated before focusing on application. Overall, the value of any

[11] Evidence for the influence of ethical writings can be seen in the examples of John Locke's influence on the Declaration of Independence, of Immanuel Kant on the German "Grundgesetz" or of Peter Singer on the animal rights movement [76, 77]. Evidence for the impact of ethical education is gathered in [78] and [79].

instance of AI ethics relies on its ability to contribute to one of the purposes. Especially when we are primarily interested in fulfilling the application purpose, however, we need to keep in mind how important it is to also find ways to fulfill the epistemic and communication purposes.

6 Limitations and objections

Before we close, let us consider some limitations and objections to our arguments.

Objection 1: "Your conception of AI ethics is overly inclusive."

A first possible objection is that we have included instances of AI ethics that some people might not consider AI ethics. For example, [5] argue that guidelines are not part of ethics but rather one possible end of ethics. Similarly, some might argue that teaching about AI ethics is not itself AI ethics, or that legislation about AI should be seen as distinct from AI ethics.

While we do not want to deny that important distinctions can be drawn between our instances of AI ethics, we see several advantages in relying on a more inclusive conception. First, our conception allows us to unite the different defenses of AI ethics instead of entering tedious discussions on who is using the term correctly. Our inclusive conception, combined with the distinction, can even be useful to avoid confusion that could arise when people disagree about the "correct" use of the term "AI ethics". Second, we see some authors who have, as a result of the criticism of AI ethics, given their preferred approaches names that do not include the term "ethics" but still describe a way of dealing with moral norms regarding AI (e.g., [7]). We argue that even if someone does not describe themselves as doing ethics, when they are dealing with moral norms (as opposed to, e.g., functional requirements) they are still doing some form of ethics. Even the legal system needs people to think about the norms by which we should be required or forbidden to act. These are ethical questions and we thus cannot replace ethics by legislation, because we still need ways to answer the ethical questions.

However, we do not claim that ours is the only "correct" conception of AI ethics, that it perfectly captures everyone's conceptual intuitions (which is likely impossible), or is the most useful conception in every context. We believe that someone who does not agree with our understanding of the term can still draw relevant insights from our arguments, such as the importance of specifying the instance and purpose that is discussed in any given context.

Objection 2: "Your instances and purposes are incomplete."

Another potential objection to our arguments and our conception of AI ethics is that we leave out some instances that could reasonably be called "AI ethics" or some purposes these instances might serve. Especially within the governance instruments, it is conceivable that new approaches will be developed that do not neatly fit the categories we describe. Recently, for example, some developments can be observed to use common industry processes for functional requirements, like audits or industry standards, and include ethical considerations (e.g., on risk management [43, 44] or the role of ethics advisors [65]). These developments can be described as efforts to professionalize the existing instances, but also as a distinct governance instrument. However, the possible existence of further approaches or instances does not reduce the importance of the distinctions we describe.

Another thing that might come to mind here is "Ethical AI" which Siau and Wang [81, p. 74] define as "an AI that performs and behaves ethically." We intentionally leave out Ethical AI because we do not consider it an instance but rather an artifact resulting from the successful use of AI ethics. Regarding the purposes, our aim was to describe some important distinctions on a rather general level. Certainly, several more fine-grained purposes of AI ethics can be described. We hypothesize that most of these fine-grained purposes can be subsumed under our general level purposes, however, showing that would be beyond the scope of this paper.

Generally, as we have not conducted a systematic literature review, we cannot exclude the possibility that we have overlooked some relevant instances or purposes. However, there are several reasons why we did not perform such a review. First, as the literature on AI ethics is vast, distinctions like ours are a prerequisite even to begin structuring it. Second, none of the main insights we derive requires a complete overview of the debate. A literature review could only complement our insights and would not contradict them. Third, a literature review can only ever provide a snapshot of the discussion at a specific point in time. Thus, it also cannot ensure that we do not overlook a *possible* use of the concept or a *possible* purpose. We therefore chose a conceptual approach which we believe allows us to cover most, if not all, relevant ways of dealing with moral norms within the AI ethics debate and the most important purposes. We nevertheless encourage further research on the prevalence of the different instances and purposes.

Objection 3: "Your distinctions do not provide practical guidance."

A limitation of our account is that, on its own, it is not sufficient to provide clear answers on how the different purposes can be best fulfilled. As a consequence, one might wonder whether the conceptual distinctions by themselves provide any additional benefit over the existing literature

that we address. However, we see several benefits in our distinctions that only arise from taking the bird's-eye perspective we provide: The discussion in Sect. 5 illustrates that distinguishing different instances and purposes of AI ethics helps our assessment of claims about the value of AI ethics. Instead of evaluating AI ethics as a whole, it allows us to assess the value of different instances and different approaches within those instances towards the different purposes. This is useful for different stakeholders involved with AI.

For *researchers*, our distinctions advance the theoretical understanding of AI ethics by making visible which instance(s) and purpose(s) a given claim concerns. A criticism of ethics guidelines as ineffective tools for application is not the same as a criticism of AI ethics research as an epistemic enterprise, nor is it the same as a criticism of AI ethics education as communication. Researchers can use the distinctions to reconstruct debates more precisely and to avoid treating claims about one instance-purpose relation as claims about AI ethics as a whole. In this way, the distinctions increase precision for formulating theoretical claims about AI ethics.

For *governance actors*, our arguments show how different governance instances can contribute to different purposes. Regulation, ethical auditing, certification, advisory boards, and non-binding guidelines should not be assessed by a single standard alone. Some instances may be particularly useful for communicating norms, others for creating incentives for application, and still others for generating knowledge about emerging risks and appropriate responses. This helps governance actors avoid two opposite mistakes: dismissing non-legal forms of AI ethics because they do not directly enforce application, and overestimating such soft instruments because they express desirable norms. Our argumentation instead encourage us to ask which combination of governance instances is needed to move from knowledge to application.

The main benefit of our conceptual distinctions is that they turn a vague question ("Is AI ethics useful?") into a set of tractable questions. Instead of asking whether AI ethics is useful or useless, different stakeholders can ask which instance of AI ethics is supposed to serve which purpose, how likely it is to be successful, and which other instances are needed to support it. This does not settle all questions about (the usefulness of) AI ethics. It does, however, provide a structure within which such questions can be asked more precisely and answered more systematically.

7 Conclusion and outlook

In this paper, we have provided an overview of different ways to engage with ethical norms regarding AI, discussed the purposes these instances of AI ethics can serve, and examined the complex relationships between the instances and purposes of AI ethics (see Fig. 1). The distinctions we introduce highlight that to meaningfully assess the value or usefulness of AI ethics, we need to be clear about *which instance* of AI ethics is being considered and with respect to *what purpose*. Importantly, this allows us to show that even if a specific instance of AI ethics proves little useful for one purpose, it might be valuable for others, and that fulfilling any purpose can make an instance valuable. Our arguments also show that existing defenses of AI ethics are not in contradiction; they merely emphasize different instances and purposes. In summary, our insights provide clarity on the issues discussed in debates over the value of AI ethics and offer helpful guidance to anyone seeking a promising instance of AI ethics for a specific purpose.

Future work should examine in detail how promising a given instance is for a particular purpose. Hagendorff laudably offers such an assessment for guidelines and the application purpose [14], but many of the other relations in Fig. 1 warrant similar efforts. Especially for the application purpose, it remains an open question which instances are most effective in achieving ethical AI. It is an empirical question whether, for instance, ethics education, ethics advisors, ethics certificates, or regulation have the strongest effect on the products developed by AI companies. Partly, the effectiveness of these measures will depend on their effectiveness in communicating about norms, which is an independent aspect that can be empirically researched. Regarding the epistemic purpose, a crucial open question is which structures are best suited to determine how to operationalize abstract moral norms for specific application contexts, which requires familiarity with the development process, but also ethical expertise and a neutral position that allows to weigh the interests of developers against those of other stakeholders. The distinctions we discussed provide a useful starting point to engage with these issues, seeing them as related, but distinct purposes that need to be fulfilled.

Acknowledgements Work on this paper was funded by the Volkswagen Foundation grants AZ 9B830, 98510, and 98514 "Explainable Intelligent Systems" (EIS), by the DFG under grant No. 389792660 as part of TRR 248, as well as by DFG grant No. 547583482 and by the Project "For the Greater Good? Deepfakes in Law Enforcement (FoGG)" funded by the Bavarian Institute for Digital Transformation (bidt), an institute of the Bavarian Academy of Sciences and Humanities, under the code KON-024-008. The Volkswagen Foundation, the DFG and the bidt had no role in preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication. We thank Katsunori Miyahara for valuable input in the conceptualization phase, and two anonymous reviewers and the participants in the

EIS colloquium for their valuable feedback on earlier versions of the paper.

Author contributions Conceptualization: Everyone; Writing—original draft preparation: AS, TS, LD, TH, HZ; Writing—review and editing: Everyone; Project Administration: AS, TH; Visualization: LK, AS; Funding Acquisition: LK

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability No datasets were generated or analysed during the current study.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Brey, P., Lundgren, B., Macnish, K., Ryan, M., Andreou, A., Brooks, L., Jiya, T., Klar, R., Lanzareth, D., Maas, J., Oluoch, I., Stahl, B.: Guidelines for the Development and Use of SIS (2019). https://doi.org/10.21253/DMU.11316833.v3
2. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**(9), 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2
3. Bietti, E.: From ethics washing to ethics bashing: a moral philosophy view on tech ethics. J. Soc. Comput. **2**(3), 266–283 (2021). https://doi.org/10.23919/JSC.2021.0031
4. Freitas Netto, S.A.V., Sobral, M.F.F., Ribeiro, A.R.B., Soares, L., G.R.: Concepts and forms of greenwashing: a systematic review. Environ. Sci. Eur. **32** (2020). https://doi.org/10.1186/s12302-020-0300-3
5. Rességuier, A., Rodrigues, R.: AI ethics should not remain toothless! A call to bring back the teeth of ethics. Big Data Soc. **7**(2) (2020). https://doi.org/10.1177/2053951720942541
6. Wagner, B.: Ethics as an escape from regulation. from e thics-washing to ethics-shopping? In: Bayamlioglu, E., Baraliuc, I., Janssens, L.A.W., Hildebrandt, M. (eds.) Being Profiled: Cogitas Ergo Sum. 10 Years of Profiling the European Citizen, pp. 84–88. Amsterdam University Press, Amsterdam (2018). https://doi.org/10.1515/9789048550180-016
7. Munn, L.: The uselessness of AI ethics. AI Ethics **3**(3), 869–877 (2023). https://doi.org/10.1007/S43681-022-00209-W
8. Andreessen, M.: The Techno-Optimist Manifesto (2023). https://a16z.com/the-techno-optimist-manifesto/ Accessed 30 March 2026
9. Lundgren, B.: In defense of ethical guidelines. AI Ethics **3**(3), 1013–1020 (2023). https://doi.org/10.1007/S43681-022-00244-7
10. Coeckelbergh, M.: AI Ethics. MIT Press, Cambridge (2020) https://doi.org/10.7551/mitpress/12549.001.0001
11. Lambrecht, F., Moreno, M.: What is AI ethics? Am. Philos. Q. **61**(4), 387–401 (2024) https://doi.org/10.5406/21521123.61.4.07
12. Heilinger, J.-C.: The ethics of AI ethics. A constructive critique. Philos. Technol. **35**(3), 61 (2022) https://doi.org/10.1007/s13347-022-00557-9
13. Mittelstadt, B.D.: Principles alone cannot guarantee ethical AI. Nat. Mach. Intell. **1**(11), 501–507 (2019). https://doi.org/10.1038/S42256-019-0114-4
14. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. Mind. Mach. **30**(1), 99–120 (2020). https://doi.org/10.1007/s11023-020-09517-8
15. Whittlestone, J., Nyrup, R., Alexandrova, A., Cave, S.: The role and limits of principles in AI ethics: towards a focus on tensions. In: Conitzer, V., Hadfield, G.K., Vallor, S. (eds.) Proceedings of the 2nd AAAI/ACM Conference on AI, Ethics, and Society (AIES), pp. 195–200. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3306618.3314289
16. Speith, T., Xu, J.: Explainability and transparency in practice: a comparison between corporate and national AI ethics guidelines in Germany and China. In: Calvaresi, D., Najjar, A., Omicini, A., Aydogan, R., Carli, R., Ciatto, G., Hulstijn, J., Främling, K. (eds.) Proceedings of the 6th International Workshop on Explainable and Transparent AI and Multi-Agent Systems (EXTRAAMAS). Lecture Notes in Computer Science, vol. 14847, pp. 205–223. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-70074-3_12
17. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci. Eng. Ethics **26**(4), 2141–2168 (2020). https://doi.org/10.1007/s11948-019-00165-5
18. Hagendorff, T.: A virtue-based framework to support putting AI ethics into practice. Philos. Technol. **35**(3) (2022). https://doi.org/10.1007/s13347-022-00553-z
19. Gornet, M., Delarue, S., Boritchev, M., Viard, T.: Mapping AI ethics: a meso-scale analysis of its charters and manifestos. In: Baeza-Yates, R., Steibel, F., Young, M. (eds.) Proceedings of the 7th ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 127–140. Association for Computing Machinery, New York (2024). https://doi.org/10.1145/3630106.3658545
20. Müller, V.C.: Ethics of artificial intelligence and robotics. In: Zalta, E.N., Nodelman, U. (eds.) The Stanford Encyclopedia of Philosophy, Fall 2023 edn. Metaphysics Research Lab, Stanford University, Stanford, CA, USA (2023). https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/
21. Seymour, W., Zhan, X., Coté, M., Such, J.M.: A systematic review of ethical concerns with voice assistants. In: Rossi, F., Das, S., Davis, J., Firth-Butterfield, K., John, A. (eds.) Proceedings of the 6th AAAI/ACM Conference on AI, Ethics, and Society (AIES), pp. 131–145. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3600211.3604679
22. Hagendorff, T.: Blind spots in AI ethics. AI Ethics **2**(4), 851–867 (2022) https://doi.org/10.1007/s43681-021-00122-8
23. Birhane, A., Ruane, E., Laurent, T., S. Brown, M., Flowers, J., Ventresque, A., Dancy, C.L.: The forgotten margins of AI ethics. In: Isbell, C., Lazar, S., Oh, A., Xiang, A. (eds.) Proceedings of the 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 948–958. Association for Computing Machinery, New York (2022). https://doi.org/10.1145/3531146.3533157
24. Bentham, J.: An Introduction to the Principles of Morals and Legislation. Dover Publications, Mineola (2007). First Published in 1789
25. Binns, R.: Fairness in machine learning: lessons from political philosophy. In: Friedler, S.A., Wilson, C. (eds.) Proceedings of the 1st Conference on Fairness, Accountability, and Transparency

(FAT*). Proceedings of Machine Learning Research, vol. 81, pp. 149–159. Proceedings of Machine Learning Research Press, Cambridge (2018). http://proceedings.mlr.press/v81/binns18a.html
26. Fleisher, W.: What's fair about individual fairness? In: Fourcade, M., Kuipers, B., Lazar, S., Mulligan, D.K. (eds.) Proceedings of the 4th AAAI/ACM Conference on AI, Ethics, and Society (AIES), pp. 480–490. Association for Computing Machinery, New York (2021). https://doi.org/10.1145/3461702.3462621
27. Hall, J.S.: Ethics for self-improving machines. In: Anderson, M., Anderson, S.L. (eds.) Machine Ethics, pp. 512–523. Cambridge University Press, New York (2011). https://doi.org/10.1017/CBO9780511978036.035
28. Bostrom, N.: Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Oxford (2014)
29. Mulligan, D.K., Kroll, J.A., Kohli, N., Wong, R.Y.: This thing called fairness: disciplinary confusion realizing a value in technology. In: Proceedings of the ACM on Human-Computer Interaction 3 (CSCW), (2019). https://doi.org/10.1145/3359221
30. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., Rahwan, I.: The moral machine experiment. Nature **563**(7729), 59–64 (2018). https://doi.org/10.1038/s41586-018-0637-6
31. Galinkin, E.: Towards a Responsible AI Development Lifecycle: Lessons From Information Security (2022) https://arxiv.org/abs/2203.02958
32. Arif Khan, F., Manis, E., Stoyanovich, J.: Towards substantive conceptions of algorithmic fairness: normative guidance from equal opportunity doctrines. In: Falcettoni, E., Machuve, D., Wilder, B., Zhou, A. (eds.) Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO). Association for Computing Machinery, New York (2022). https://doi.org/10.1145/3551624.3555303
33. Ng, M.Y., Kapur, S., Blizinsky, K.D., Hernandez-Boussard, T.: The AI life cycle: a holistic approach to creating ethical AI for health decisions. Nat. Med. **28**(11), 2247–2249 (2022). https://doi.org/10.1038/s41591-022-01993-y
34. Deck, L., Schomäcker, A., Speith, T., Schöffer, J., Kästner, L., Kühl, N.: Mapping the potential of explainable artificial intelligence (XAI) for fairness along the AI lifecycle. In: Cerrato, M., Coronel, A.V., Ahrweiler, P., Loi, M., Pechenizkiy, M., Tamò-Larrieux, A. (eds.) Proceedings of the 3rd European Workshop on Algorithmic Fairness (EWAF). CEUR Workshop Proceedings, vol. 3908, pp. 1–25. CEUR-WS, Aachen (2024). https://ceur-ws.org/Vol-3908/paper_38.pdf
35. Fiesler, C., Garrett, N., Beard, N.: What do we teach when we teach tech ethics? a syllabi analysis. In: Zhang, J., Sherriff, M., Heckman, S., Cutter, P.A., Monge, A.E. (eds.) Proceedings of the 51st ACM Technical Symposium on Computer Science Education (SIGCSE), pp. 289–295. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3328778.3366825
36. Bargh, J.A., Schul, Y.: On the cognitive benefits of teaching. J. Educ. Psychol. **72**(5), 593–604 (1980). https://doi.org/10.1037/0022-0663.72.5.593
37. Fiorella, L., Mayer, R.E.: The relative benefits of learning by teaching and teaching expectancy. Contemp. Educ. Psychol. **38**(4), 281–288 (2013). https://doi.org/10.1016/j.cedpsych.2013.06.001
38. Nielsen, N.R.: Social responsibility and computer education. In: Pollack, S.V., Williams, L.H., Aiken, R.M. (eds.) Proceedings of the 2nd SIGCSE Technical Symposium on Computer Science Education, pp. 90–96. Association for Computing Machinery, New York (1972). https://doi.org/10.1145/800155.805011
39. Borenstein, J., Howard, A.: Emerging challenges in AI and the need for AI ethics education. AI Ethics **1**(1), 61–65 (2021). https://doi.org/10.1007/S43681-020-00002-7
40. Baum, K., Sterz, S.: Ethics for nerds. Int. Rev. Inf. Ethics **31**(1), 1–9 (2022). https://doi.org/10.29173/irie484
41. EU High-Level Expert Group on Artificial Intelligence: Ethics Guidelines for Trustworthy AI (2019). https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
42. Köhl, M.A., Baum, K., Langer, M., Oster, D., Speith, T., Bohlender, D.: Explainability as a non-functional requirement. In: Damian, D.E., Perini, A., Lee, S. (eds.) Proceedings of the 27th IEEE International Requirements Engineering Conference (RE), pp. 363–368. IEEE, Piscataway (2019). https://doi.org/10.1109/RE.2019.00046
43. ISO/IEC 23894:2023: Information Technology—Artificial Intelligence—Guidance on Risk Management. International Organization for Standardization and International Electrotechnical Commission. ISO/IEC 23894:2023
44. Tabassi, E.: Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology (U.S.), Gaithersburg, MD (January 2023). 10.6028/NIST.AI.100-1 . http://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf Accessed 31 July 2026
45. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S.: The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. Commun. ACM **64**(4), 136–143 (2021). https://doi.org/10.1145/3433949
46. Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I.G., Cosentini, A.C.: A clarification of the nuances in the fairness metric landscape. Sci. Rep. **12**(1) (2022). https://doi.org/10.1038/s41598-022-07939-1
47. Deck, L., Schöffer, J., De-Arteaga, M., Kühl, N.: A critical survey on fairness benefits of explainable AI. In: Baeza-Yates, R., Steibel, F., Young, M. (eds.) Proceedings of the 7th ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 1579–1595. Association for Computing Machinery, New York (2024). https://doi.org/10.1145/3630106.3658990
48. Weerts, H.J.P., Dudík, M., Edgar, R., Jalali, A., Lutz, R., Madaio, M.: Fairlearn: assessing and improving fairness of AI systems. J. Mach. Learn. Res. **24** (2023) https://arxiv.org/abs/2303.16626
49. Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J.T., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y.: AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. IBM J. Res. Dev. **63**(4/5) (2019). https://doi.org/10.1147/JRD.2019.2942287
50. Hardt, M., Chen, X., Cheng, X., Donini, M., Gelman, J., Gollaprolu, S., He, J., Larroy, P., Liu, X., McCarthy, N., Rathi, A., Rees, S., Siva, A.A., Tsai, E., Vasist, K., Yilmaz, P., Zafar, M.B., Das, S., Haas, K., Hill, T., Kenthapadi, K.: Amazon SageMaker clarify: machine learning bias detection and explainability in the cloud. In: Zhu, F., Ooi, B.C., Miao, C. (eds.) Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp. 2974–2983. Association for Computing Machinery, New York (2021). https://doi.org/10.1145/3447548.3467177
51. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., Vayena, E.: AI4People's ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind. Mach. **26**, 689–707. (2018). https://doi.org/10.1007/s11023-018-9482-5
52. Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., Floridi, L.: Ethics as a service: a pragmatic operationalisation of AI ethics. Mind. Mach. **31**(2), 239–256 (2021). https://doi.org/10.1007/S11023-021-09563-W

53. ISO/IEC 42001:2023: Information Technology—Artificial Intelligence—Management System. International Organization for Standardization and International Electrotechnical Commission
54. Buolamwini, J., Gebru, T.: Gender shades: intersectional accuracy disparities in commercial gender classification. In: Friedler, S.A., Wilson, C. (eds.) Proceedings of the 1st Conference on Fairness, Accountability, and Transparency (FAT*). Proceedings of Machine Learning Research, vol. 81, pp. 77–91 (2018) https://proceedings.mlr.press/v81/buolamwini18a.html
55. Kearns, M.J., Neel, S., Roth, A., Wu, Z.S.: Preventing fairness gerrymandering: auditing and learning for subgroup fairness. In: Dy, J.G., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 80, pp. 2569–2577 (2018) https://arxiv.org/abs/1711.05144
56. Green, B.: Escaping the impossibility of fairness: from formal to substantive algorithmic fairness. Philos. Technol. 35(4) (2022). https://doi.org/10.1007/s13347-022-00584-6
57. Blodgett, S.L., Barocas, S., Daumé, I., Wallach, H., H.M.: Language (technology) is power: a critical survey of bias in NLP. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 5454–5476. Association for Computational Linguistics, Dundee, Scotland, UK (2020). https://doi.org/10.18653/V1/2020.ACL-MAIN.485
58. Katzman, J., Wang, A., Scheuerman, M.K., Blodgett, S.L., Laird, K., Wallach, H.M., Barocas, S.: Taxonomizing and measuring representational harms: a look at image tagging. In: Williams, B., Chen, Y., Neville, J. (eds.) Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI), 35th Conference on Innovative Applications of Artificial Intelligence (IAAI), 13th Symposium on Educational Advances in Artificial Intelligence (EAAI), pp. 14277–14285. AAAI Press, Palo Alto (2023). https://doi.org/10.1609/AAAI.V37I12.26670
59. Loi, M., Ferrario, A., Viganò, E.: Transparency as design publicity: explaining and justifying inscrutable algorithms. Ethics Inf. Technol. **23**(3), 253–263 (2021). https://doi.org/10.1007/S10676-020-09564-W
60. European Parliament, Council of the EU: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons With Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (2016)
61. European Parliament, Council of the EU: Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act) (2022). https://eur-lex.europa.eu/eli/reg/2022/2065/oj
62. European Parliament, Council of the EU: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts (Artificial Intelligence Act) (2024). https://eur-lex.europa.eu/eli/reg/2024/1689/oj. Accessed 19 July 2024
63. Hay, D.F.: The roots and branches of human altruism. Br. J. Psychol. **100**(3), 473–479 (2009). https://doi.org/10.1348/000712609X442096
64. Blackman, R.: Why you need an AI ethics committee: expert oversight will help you safeguard your data and your brand. Harv. Bus. Rev. **100**(7–8), 118–125 (2022)
65. prEN 18274:2025: Competence Requirements for Professional AI Ethicists. European Committee for Standardization and European Committee for Electrotechnical Standardization. Draft European Standard
66. Metzinger, T.: EU Guidelines: Ethics Washing Made in Europe. Der Tagespiegel (2019). https://www.tagesspiegel.de/politik/ethics-washing-made-in-europe-5937028.html. Accessed 14 May 2024
67. Scharowski, N., Benk, M., Kühne, S.J., Wettstein, L., Brühlmann, F.: Certification labels for trustworthy AI: insights from an empirical mixed-method study. In: Harrington, C., Huq, A., Fox, S., Tan, C. (eds.) Proceedings of the 6th ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 248–260. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3593013.3593994
68. Costanza-Chock, S., Raji, I.D., Buolamwini, J.: Who audits the auditors? recommendations from a field scan of the algorithmic auditing ecosystem. In: Isbell, C., Lazar, S., Oh, A., Xiang, A. (eds.) Proceedings of the 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 1571–1583. Association for Computing Machinery, New York (2022). https://doi.org/10.1145/3531146.3533213
69. Cane, P.: Taking law seriously: starting points of the Hart/Devlin debate. J. Ethics **10**(1/2), 21–51 (2006) https://doi.org/10.1007/s10892-005-4590-x
70. Tudorach, D.: Press Conference Following the Artificial Intelligence Act Trilogue—Part 3 (2023). https://newsroom.consilium.europa.eu/events/20231206-artificial-intelligence-act-trilogue/142864-3-press-conference-part-3-20231209-1 Accessed 14 May 2024
71. Maanen, G.: AI ethics, ethics washing, and the need to politicize data ethics. Digital Soc **1**(2) (2022). https://doi.org/10.1007/s44206-022-00013-3
72. Wright, J.: The development of AI ethics in Japan: ethics-washing society 5.0? East Asian Science. Technol Soc Int J **18**(2), 117–134 (2024). https://doi.org/10.1080/18752160.2023.2275987
73. Cole, M., Cant, C., Ustek Spilda, F., Graham, M.: Politics by automatic means? A critique of artificial intelligence ethics at work. Front Artif Intell **5**, 869114 (2022). https://doi.org/10.3389/frai.2022.869114
74. Rawls, J.: A Theory of Justice: Original Edition. Harvard University Press, Cambridge (2005). https://doi.org/10.2307/j.ctvjf9z6v
75. Laat, P.B.: Companies committed to responsible AI: from principles towards implementation and regulation? Philos. Technol. **34**(4), 1135–1193 (2021). https://doi.org/10.1007/s13347-021-00474-3
76. Stern, K.D.: John locke and the declaration of independence. Cleveland-Marshall Law Rev. **15**(1), 186–195 (1966)
77. Hong, M.: Immanuel Kants 300. Geburtstag und das Grundgesetz. VerfBlog, 2024/4/22 (2024). https://doi.org/10.59704/275738a964c1c3cb. https://verfassungsblog.de/immanuel-kants-300-geburtstag-und-das-grundgesetz/
78. Knight, J.: Evaluating the impacts of a research ethics training course on university researchers. Soc. Sci. **12**(3) (2023). https://doi.org/10.3390/socsci12030182
79. Schuett, J., Reuel, A.-K., Carlier, A.: How to design an AI ethics board. AI Ethics (2024). https://doi.org/10.1007/s43681-023-00409-y
80. Lucaj, L., Van Der Smagt, P., Benbouzid, D.: AI regulation is (not) all you need. In: Harrington, C., Huq, A., Fox, S., Tan, C. (eds.) Proceedings of the 6th ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 1267–1279. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3593013.3593997
81. Siau, K., Wang, W.: Artificial intelligence (AI) ethics: ethics of AI and ethical AI. J. Database Manag. **31**(2), 74–87 (2020). https://doi.org/10.4018/JDM.2020040105

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