Uses and Abuses of AI Ethics

Lily Eva Frank (0000-0001-8659-2390)
Eindhoven University of Technology

Michał Klincewicz (0000-0003-2354-197X)
Tilburg University / Jagiellonian University

Abstract (150 words)

In this chapter we take stock of some of the complexities of the sprawling field of AI ethics. We consider questions like "what is the proper scope of AI ethics?" And "who counts as an AI ethicist?" At the same time, we flag several potential uses and abuses of AI ethics. These include challenges for the AI ethicist, including what qualifications they should have; the proper place and extent of futuring and speculation in the field; and the dilemmas concerning how we use our public and academic resources.

Keywords: AI Ethics, AI regulation, Responsibility Gaps, Algorithms, Ethics Washing, Technology Ethics, Expertise, Value-sensitive Design

Introduction

Media and public discourse surrounding AI have reached an all-time fever pitch. As we write this chapter, in 2023 the United Nations proposed the creation of an international AI monitoring agency, akin to the International Atomic Agency, specifically citing threats of disinformation that could undermine democracy. UN Chief Antonio Gutierez said he has been "warned by AI developers themselves "[t]hese scientists and experts have called on the world to act, declaring AI an existential threat to humanity on a par with the risk of nuclear war. We must take those warnings seriously" (UN Audiovisual Library, June 2023). The claim that AI poses an existential threat or risk to humanity has several diverse sources, among them Jaan Tallin, co-founder of Skype and the Future of Life Institute, Oxford University philosopher Nick Bostrom, and X (Twitter) CEO Elon Musk, just to name a few.

Over the past couple of years, news coverage and academic debate over large language models (LLMs) has exploded in volume and in level of hyperbole. In the field of education, for example, LLMs are seen as a major challenge for assessing students' competencies in disciplines that involve writing and argumentative skills, as well as programming; creating new and virtually undetectable ways for students to have their work done by someone (or something) else, and undermining students' abilities to hone their own skills. The open letter "Pause Giant AI Experiments: An Open Letter," which has been

signed by approximately 34,000 people as of today, calls “on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.”[1]

Perhaps in response to some of this, interdisciplinary research centers and policy institutes dedicated to research and policy about responsible and ethical AI are proliferating and the work of the European Commission High level-expert group on AI issued its final report on trustworthy AI in 2020, partially informing the proposed legislation in the European Parliament—“The AI Act”[2]—2023. At the same time, mega-corporations like Microsoft and Google have eliminated their ethics boards; Microsoft disbanded their “Ethics and Society” team in March 2023 and Google eliminated their AI ethics board in April 2019. And these are just some of the highlights.

Amid all this activity, it is important to reflect on and find ways to distinguish things that are genuinely concerning about AI from a moral perspective from what may perhaps be best characterized as a firehose of fears, misunderstandings, and competing narratives. This is essential to combatting the backlash of "ethics bashing," which is "the trivialization of ethics and moral philosophy now understood as discrete tools or pre-formed social structures such as ethics boards, [corporate] self-governance schemes or stakeholder groups." (Bietti 2021), and cynicism about AI ethics in public policy and as a scholarly enterprise. The central aim of this chapter is to give people concerned with AI and ethics the means to distinguish the signal from the noise in some of the prominent public and academic debates. We will do this by first addressing the boundary conditions for discussions of AI ethics. This will be followed with a discussion of the way in which a lack of clarity about these boundaries has been used and abused. This section will also advance an argument for the definition of boundaries in AI ethics that was developed and defended in the first section. Throughout the chapter we will flag what we call "uses" and "abuses" of AI ethics, some of the abuses beset applied ethics or ethics of technology more generally and many of the methods, topics, and issues we discuss have the potential to function as uses or abuses.

Boundaries and Uses of AI Ethics

Under what conditions does it make sense or is it appropriate for AI itself to be the unit of ethical analysis? As a comparison, consider ethical debates surrounding the use of life-sustaining technologies in medicine, such as mechanical ventilation or artificial nutrition

and hydration. These technologies use electricity. However, to gain insight into the relevant and challenging ethical issues surrounding these technologies, when they should be used, or discontinued, who should have decision-making authority when it comes to their use, allocation of these technologies as a scarce resource, etc., their use of electricity is of very limited relevance. Perhaps the fact that these technologies use electricity could be of ethical relevance in specific cases, for example, if we consider electricity use in context of its production and the effect this has on the emission of greenhouse gases. Other than these special cases, life-sustaining medical technology is not an appropriate topic for "electricity ethics" should such a sub-discipline exist.

Artificial intelligence is a general term for a set of computing technologies, which include algorithms and programs. So, right out of the gate we have an important question to ask: in what sense are computer programs or algorithms the subject of ethical analysis? Is it the technological hardware itself, the algorithm that runs on it, the way the algorithm is used, or the socio-technical system that includes all these things the proper target of ethical analysis? Depending on the answer we give, we may or may not end up leaving the domain of AI ethics. If we target the algorithm exclusively, then a good case can be made that this is indeed a matter for computer or engineering ethics; if we focus on an algorithm or technology that falls into the relevant subset of AI-based programs or algorithms, then we have a case for AI ethics.

It is important to remember this broader context of computational technologies and not just those that use AI. All technologies, including algorithms (and the special subset of these that involve AI), are not value neutral (Winner 1980, 1986; van de Poel 2001; Illies & Meijers 2009; Van de Poel & Kroes 2013). Guns are designed for violence. Bridges are designed for ease and efficiency in travel. Valuing violence or ease and efficiency contributes to guns and bridges coming to exist and then being used in particular ways. In the same sense, algorithms or the systems that use them may not be morally neutral. They may embody values, which are reflected in their design and functioning.

For example, an algorithm may be better at identifying patterns of recidivism than other sorts of patterns or be designed specifically for that purpose. And if that is the case, then it itself becomes a proper target of ethical analysis. This is especially so if the algorithm involves technologies or processes that are themselves value-laden. For example, a machine learning model designed to identify potential criminal recidivists can inherit the values of

the society that produced data that was used to train it (McKay, 2020). In this sense, the algorithm, model, and the data *itself*, may be biased and can even be said to embody values.

If we focus on the way that data used to train machine learning (supervised or unsupervised) was collected, then there are other, more general domains of ethics that take precedence over AI ethics, such as research ethics. If we focus on the factors responsible for the patterns we may see in this data, then perhaps we should move to the ethics of policing, policy, law, or the criminal justice system. If it is the socio-technical system, then it may not be clear at all whether the target of our analysis is in ethics or some other normative field altogether, such as law or law enforcement. So, the involvement of AI algorithms and technology is not, on its own, sufficient to warrant being a proper target for AI ethics.

Further, AI ethics is a subfield within the broader fields of practical ethics, ethics of technology, and professional ethics, e.g., computer ethics or engineering ethics. And some of the debates in those broader domains naturally apply to examples of technologies that use AI. For example, in the example of the recidivism machine learning algorithm, ethics of technology may interrogate the use of any technology that can be used to identify potential criminals, independently of whether it uses AI at all. From that perspective, a magic eight ball that identifies recidivists is no less a target of ethical analysis than a machine learning algorithm that does the same. What may be morally problematic is the practice of predictive policing, not the technology used for that purpose. So, the target of moral analysis for that worry is not the algorithm as such, but its use for that purpose.

In addition, moral problems that occur in the fields of professional or engineering ethics may arise in cases where AI is used but are not necessarily best characterized as problems of AI ethics. For example, the problem of many hands can apply to AI technologies. The problem of many hands arises when (Thompson, 1980; Coeckelbergh, 2020) attributing moral responsibility to any one individual or group of individuals is complicated by the fact that many different people and groups contributed to different extents and at different times to the creation or allowance of the moral wrong. The problem of many hands affected the Ariane rocket disaster, in which a group of programmers re-used old computer code (Kroll, 2020). No AI was involved (to our knowledge) at all. Given all this, identifying the proper scope of analysis and target, is the first thing that AI ethicists

should do: they should make an argument supporting the claim that what they are analyzing is an ethical issue and of special concern for AI ethics.

These initial observations give us a working definition for a proper target of AI ethics analysis that we will simply assume throughout the chapter: *the AI technology, AI algorithm or AI-empowered system that can be used for some a specific purpose that involves values.* Of course, what values are important would be highly context-dependent and itself open to debate; but that is outside the scope of this chapter. Formulated in this fashion, AI ethics would have its own appropriate target(s) for analysis (for a full account of these, see: Brey, 2012; Umbrello et al, 2023).

Once they have a proper target, what do AI ethicists do, exactly? It is good to start with some examples. In general, values are embedded in technologies through design decisions, for example, the height of a bridge can be designed to make it difficult for people using buses as public transportation to access public parks (Winner, 1980). Similarly, values can end up being built into the architecture of AI-powered technologies. Take the example of the machine learning algorithms that rank profiles in a dating app based on some criteria of attractiveness (Klincewicz, Frank, Jane, 2022; Frank & Klincewicz, 2021) or social robots designed to aid in the care of the elderly, which modulate the tone of voice the robot uses so as to not be infantilizing. AI ethicists can presumably tell us whether the technologies embody particular values and whether there are morally salient consequences from their use.

The second, and perhaps more important use of AI Ethics, is to elucidate which values should or should not be embedded in the AI or AI-powered system, and why. Perhaps AI-powered robots designed for sexual gratification embody values that we, on reflection, do not endorse or, perhaps, should not endorse (Nyholm & Frank 2019). Bringing this to light may matter to policy, law, and how we lead our lives, once sex robots are technologically and economically viable. The deliverable here is a persuasive and philosophically informed argument that bears on these questions. In policy documents, this project is sometimes referred to as "value alignment."

Value alignment aims to answer the question "how to ensure that AI systems are properly aligned with human values and how to guarantee that AI technology remains properly amenable to human control" (Gabriel & Ghazavi, 2022). This has several complications that need to be addressed by those working in AI ethics: 1) diversity and

incompatible values (Risse, 2021; Wan, Hooker, Thomas Donaldson, 2021; Sutrop, 2020; Mohamed, Png, Isaac, 2020); 2) assuming an ought from an is (Wan, Hooker, and Donaldson, 2021); and 3) neglect of principles and theory (Wan, Hooker, Thomas, and Donaldson, 2021). Gabriel and Ghazavi (2022) argue that AI poses a special challenge for value alignment because of its unique properties and capacities compared with other complex technological systems. The unique features they suggest are 1) opacity of algorithmic decision making; 2) greater freedom and autonomy in decision making; and 3) that fact that "AI models have agential properties that manifest to a higher degree than in non-AI systems" (p. 7).

Another thing that an AI ethicist can do is to bring into relief so-called techno-responsibility gaps (Tigard, 2021). For example, Nyholm (2023) has argued that the use of self-driving cars presents responsibility gaps when they are involved in traffic accidents, because it is not obvious who is responsible or blameworthy. This topic has gotten significant academic and policy attention as of late, as AI technologies become more generally available and their ability to facilitate or altogether replace human decision-making has increased. Debate has focused on whether such gaps exist in specific cases, what is their nature and moral significance, and how or even if they should be closed, resolved, or plugged (Santoni de Sio, F., & Mecacci, G., 2021; Burton et al, 2020; Danaher 2022; Nyholm, 2023). Arguments in favor of the existence of techno-responsibility gaps tend to see them as presenting a set of moral problems that are unique to the AI and robotics context because of the level of autonomous action of these systems.

Fields like animal ethics, environmental ethics, and bioethics grapple with (potentially) ethically relevant questions about the ontological status of the entities they study. This bears on whether said entities deserve moral consideration, have rights and or obligations, in other words whether they can be moral patients or agents. AI ethicists can do the same. This is not limited to the properties these entities possess, (e.g., are fetuses sentient beings?) but also the social/political/community roles they play in relation to others (e.g., in what ways are social robots recognized by other members of the moral community?), as in Gunkel's "relational turn" (Gunkel, 2022) or Danaher's behaviorism (2021). Parts of what AI ethicist do here is try to provide reasons whether we *should* (assuming it is possible) create artificial entities that either possess the properties relevant

for moral patiency or agency or play the requisite relational roles (Bryson 2010; 2018; Gunkel 2018).

Since the difference between AI and non-AI computer algorithms and systems defines the disciplinary boundaries for AI ethics, it is particularly important for AI ethicists to know what AI is, in a technical sense. One rule of thumb here is that AI is defined externally by the boundaries of the academic sub-discipline of artificial intelligence within computer science. This means that *the boundaries of AI ethics at least to some extent depend on what the computer scientists say artificial intelligence is*. Let's call this the *disciplinary instrumentalism thesis* about AI ethics. If some machine learning algorithm ceases (for whatever reason) to be considered AI by them in the future, then AI ethics should have little or no interest in talking about such a system and the values that it may or may not promote or embody.

Instrumentalism is easy to put into practice but may be unattractive in principle. The operative definition of artificial intelligence in the academy may piggyback on un-argued-for or un-articulated answers to questions in the philosophy of mind or cognitive science, such as: is machine learning really a form of learning? Is reinforcement learning intelligent? Is information just a measure of entropy? Can machines have minds? These questions and many others unfortunately raise ("unfortunate" for the instrumentalist) fundamental ontological questions that are of concern to other disciplines. Defining the boundaries, then, becomes more complicated than asking a computer scientist about AI, especially since they may know little about what is going on in those other fields.

This may lead one to shrug and wonder why it is important to attempt to delineate the borders of AI ethics in the first place. To make our case we will go through an array of abuses of AI ethics, which are facilitated by a lack of boundary conditions.

Abuses of AI ethics and its boundaries

The first problem with mis-identifying the proper targets of AI ethics is the opportunity cost. Academics, governments, and the media all have limited resources to devote to topics of societal importance. Time that is devoted to "AI ethics" as such, is time not spent on other concerns, which may be the proper targets of analysis of other efforts in engineering ethics, political philosophy, or law. Given this, there is an obligation to prioritize

certain ethical concerns, moral outrage targets, and energy consumed in scholarship, activism, and public debates.

AI has taken airtime away from other issues that are arguably more important, such as global poverty and inequality, climate change, international conflicts, etc. In the 2023 Future of Life "Open Letter," for example, its authors ask: "Should we automate away all the jobs, including the fulfilling ones?" This question, in the context of demanding a pause on the training of advanced generative systems, suggests that the elimination of jobs, especially fulfilling jobs, is primarily an AI issue, rather than a result of globalization, insufficiently regulated capitalism, and automation much more generally. Why, one may wonder, do people like Elon Musk not sign open letters about the effects of globalized capitalism on income inequality or the erosion of democracy instead?

Independent of this, within the academy, incentives have been created to do research on AI and AI ethics or to reframe existing research programs as being about AI or AI ethics, while being about something else. These incentives come from funding agencies that emphasize AI in their calls for applications, conferences, workshops, courses, institutes, and job openings, including funded Ph.D.'s in the field. Working in AI ethics also opens opportunities for the academic ethicist to collaborate with industry, serve on advisory committees, and present keynotes at prestigious international events, raising one's profile and the profile of one's associated university. The creation of ethical guidelines is happening very fast and is itself "big business" (van Maanen, p. 4; Floridi & Cowls 2019). Without any kind of ill-intention research resources end up being poured into AI ethics that very well may have been better used elsewhere. This is also an opportunity cost of not being clear about proper targets of AI ethics.

Mark Coecklebergh's book *AI Ethics* (2020), contains a chapter playfully titled, "It's the Climate, Stupid! On Priorities, The Anthropocene, and Elon Musk's Car in Space." Research focused on the ethical threats of AI may not only preclude *research* on more important threats, such as climate change. It also risks the ethical dangers of techno-solutionism or techno-optimism, which lead us to mischaracterize the nature of the problems we are facing and to put unjustified hope in technological breakthroughs and solutions that do not require substantial political change or disruption of lifestyles. Techno-solutionist or optimistic beliefs may also give us reason to delay actions that could gradually produce solutions or mitigate problems because we are waiting for the technological

solution and furthermore to ignore the role that past technological developments (in particular political-economic contexts) have contributed to current challenges. Techno-solutionism encourages hope that complex social, political, economic, etc., problems can be remedied by implementing the right technological fix and transformed into “neatly defined problems with definite, computable solutions or as transparent and self-evident processes that can be easily optimized—if only the right algorithms are in place!” (Morozov 2013).

Despite the danger AI has to make some of humanity's critical preexisting problems worse, it also has the potential to contribute solutions. An important and under-researched area of AI ethics is the critical work of settling “whether a particular problem ought, morally, to be conceived of as a technological problem and consequently be addressed with the help of (AI-based) technology in the first place." (Heilinger 2022). In the ethics of social robotics and care robots as well as some applications of AI in medicine, there is a debate about the need for further technological solutions when underlying problems have to do with inadequate staffing and pay, doctors not having enough time to spend with patients, or inequitable distribution of health care resources and infrastructure. Debates about the ethics of self-driving cars are also subject to the critique that traffic accidents and high fuel consumption are the result of developing a car culture, as opposed to public transportation or other forms of travel, resulting in path dependencies. Instead of developing self-driving cars, critics argue, we should invest in trains or perhaps even engage in deeper reflection on our need for hyper mobility, commuter culture, and cheap travel.

Abuses of AI ethics expertise

Another abuse of AI ethics, which is often a consequence of misidentifying the proper target of AI ethics, is misidentifying the expertise that is necessary to do ethics. There are many ways in which this manifests itself, but the thing that binds these abuses together can be expressed in the following tri-partite general formula: Expert 1 argues that *p*, Expert 2 argues that *not p*, and what ultimately matters whether *p* or *not p* is put in practice, is a distinct special interest. What makes these abuses work is a general lack of clarity about what makes Expert 1 or Expert 2 experts in the first place and what matters to that evaluation. Luciano Floridi has helpfully collected these abuses under the phrase "translating principles into practices" (2019). All of them fit this formula, especially in the context of creating policy and legislative guidance documents on AI.

The most obvious and perhaps also most common version of this is *ethics shopping* (Wagner 2018), which is predicated on the idea that there is a "market of principles and values" from which special interests can choose, adopt, or revise through mixing and matching the opinions of experts to fit with antecedent goals, instead of changing these goals or the ways in which these goals are pursued. One recent and particularly striking example of this is from the government of the United Kingdom, which in October 2023[3] dismissed its government data ethics advisory board. A reporter interviewed anonymous board members who indicated that the board's work was not well received by the government, nor were its whitepapers (Klovig Skelton 2023).

In lieu of continuing the work of the board, the UK government's Department for Science, Innovation and Technology (DCIT) told the data ethics advisory board "they favored a strategy of a more flexible approach to consulting advisers, picking from a pool of external people, rather than having a formal advisory board" (Klovig Skelton 2023). Without access to the details of the decision, nothing definitive can be said about it, but the noted resistance to the committee's advice and the stated pick and choose strategy for experts signals that the DCIT is taking advantage of the growing market for AI ethics expertise and opinion to find guidance more consistent with their preexisting scientific, political, or economic goals. In effect, they are ethics shopping.

Another way that not being clear about the requirements of expertise for AI ethics can be abused is by ethics *bluewashing*, which is an umbrella term for a range of activities that implement superficial measures of ethics enforcement. *Bluewashing* is a cousin of greenwashing and other forms of disinformation used to confuse public opinion or regulatory bodies intentionally or unintentionally about the nature of one's activities. For example, a major polluter sponsoring "green" public events, advertising its commitments to sustainability, etc., while simultaneously violating extant environmental protection laws or corrupting officials to look the other way is greenwashing. *Bluewashing* could take the form of an internal AI ethics committee essentially playing the role of a public relations team by promoting, say, AI algorithm transparency, etc., while doing nothing to stop abuse of AI algorithms for nefarious means within the company. Pick a public relations expert, give them an internal position as an AI expert, and proceed to accumulate profit. What makes *bluewashing* possible is the idea that the work of AI ethicists can be done independently of

the sort of expertise and institutional embedding that is described above. Being an AI ethicist is something that is reduced to a matter of being called one.

Beyond *bluewashing* or *ethics shopping*, there are other kinds of abuses. For one, there is *ethics dumping*, or the strategy of moving AI research, development, or deployment to locations where what is considered legal and ethically acceptable is more tolerant or AI legislation and ethics are less developed, if at all. Another is *ethics shirking*, or the practice of shifting activities from those that require ethics oversight to those that do not. And finally there is just straightforward *AI ethics lobbying*, akin to any other outside institutional influence, but this time directed at creating a more favorable legal and ethical context for AI research and development that would otherwise be ethically suspect. With all three, making it clear what AI ethics is, who the AI experts are, and what the deliverables of their work look like, would limit the effectiveness of these strategies or eliminate them altogether. AI ethics lobbyists, for instance, would have to contend with research and scholarship from actual AI ethics experts; AI ethics shirkers would have a harder time finding ways to circumvent their obligations when faced with expertise that can properly analyze the consequences of activity; and AI ethics dumpers would likely have their offshoring strategy scrutinized by people that understand what they are doing and why.

These are not easy problems to tackle and finding suitable experts is not easy. First, identifying who qualifies as an expert on any subject is a complex epistemological question (Quast & Seidel, 2018). Second, gatekeeping in the field poses risks, especially the risks of excluding from the discussion scholars and stakeholders who have been historically marginalized or who have insights from their own domains of knowledge that have not traditionally been considered. These two problems notwithstanding, identifying the proper target of AI ethics can help to solve this problem. Those who are best equipped to carry on the work of AI ethics, are those who should be identified as AI ethicists.

And who might these people be? The work of AI ethics is best done by individuals with PhD-level training in moral philosophy, who have published research in respected venues for moral philosophy scholarship, and who are part of the professional community. Some sub-set of this group is best positioned to carry out the task of AI ethics. This becomes especially evident compared to people with no such training, e.g. those who do not publish scholarship in moral philosophy or who are not a part of the relevant professional

communities. This observation also extends to applied ethics and the sub-domains of moral philosophy, such as bioethics, business ethics, technology ethics, etc.

Ethicists working in these sub-disciplines can be identified as bioethicists, business ethicists, technology ethicists, or AI ethicists, when they bring their expertise to bear on issues within the boundaries of those particular domains. In the case of AI ethics, this would be, as we have describe above, when the target of their analysis is an AI algorithm or AI-powered technology, with what counts as those is defined instrumentally, by reference to what computer scientists think AI is. Importantly, ethicists working in sub-domains are typically also assumed to have domain-specific knowledge sufficient to carry out their work. For example, business ethicists should have sufficient knowledge about markets, business processes, and more specific things, such as tax policy or legislation. This is also true of AI ethicists, who should have domain-specific knowledge of artificial intelligence. It should be noted, however, that this is especially challenging, given the relatively high level of expertise one needs in computer science to become sufficiently knowledgeable here.

There are other tasks that AI ethicists engage in which, arguably, are not so connected to their training or expertise, or to their domain-specific understanding of AI. Some people doing AI ethics take it to be part of their responsibility to speculate about the possible future consequences of new technologies. This is a way of exploring possible moral complexities and conflicts in advance of the reality of these advances. This is perhaps unsurprising, since all branches of applied ethics confront predictions about possible future technological developments, their applications, and speculate about consequences. In the field of bioethics, for example, these debates often focus on the consequences of developing new reproductive technologies, cloning, and human bio-enhancement.

Science-fiction has been ahead of ethicists in this realm for some time and are rightfully acknowledged as a means of meaningful speculation. Isaac Asimov's laws of robotics are still regularly brought up in the context of the rights of AI and robots; *The Terminator* films have been used to discuss autonomous weapon systems; and the novels of William Gibson has been employed as a plausible model of a technological future where global capitalism and artificial intelligence run everything. Consider, for example, the 1995 animated movie *Ghost in the Shell* and this dialogue between the fully manufactured cyborg Major Motoko Kutsanagi and the "living doll" Batou, whose brain is the only naturally organic thing in an otherwise synthetic body:

> **Major Motoko Kusanagi**: Well, I guess cyborgs like myself have a tendency to be paranoid about our origins. Sometimes I suspect I am not who I think I am, like maybe I died a long time ago and somebody took my brain and stuck it in this body. Maybe there never was a real me in the first place, and I'm completely synthetic like that thing [referring to an AI-powered robot].
>
> **Batou**: You've got human brain cells in that titanium shell of yours. You're treated like other humans, so stop with the angst.
>
> **Major Motoko Kusanagi**: But that's just it, that's the only thing that makes me feel human. The way I'm treated. I mean, who knows what's inside our heads? Have you ever seen your own brain?

It is sometimes astonishing just how far ahead of the academic debate science fiction has been all this time. Batou and Major Kusanagi may be the first relationist theorists (Gunkel 2022) in AI ethics. That said, ethicists should be careful not to overstate the importance of these examples, especially since they are the product of fiction and could be seen as a distraction from the serious nature of the issues involving future uses of AI.

We do not have to rely on science fiction to do this work, either. Ethicists developed methods for doing this more rigorously using transdisciplinary design methods of "responsible futuring" (Zaga forthcoming); the creation of techno-moral scenarios (Swiestra et al. 2009); prospective evaluation (Grunwald 2009); and anticipatory technology ethics (Brey 2012). All of these methods must grapple with several challenges, including the Collingridge dilemma (1980). The Collingridge dilemma explains that early in the development of a technology, while there is a greater possibility of controlling it, we have very limited ability to foresee its impacts; once the technology is diffuse in a society and its consequences begin to appear, our ability to control the technology declines dramatically. Critics of more speculative or hypothetical analyses of new and emerging technologies also point out potential several pitfalls: epistemological challenges and uncertainty; lack of philosophical and argumentative rigour; promotion of public hype and or fear of emerging technology that is unwarranted; and distraction from more urgent ethical concerns of the present.

None of this should be read as a blanket prohibition on philosophers engaging in speculative, futuristic, or even absurd thought experiments, no matter how far-fetched. From moral twin earth (Horgan & Timmons 1991) to trolley problems (Foot 1967) to the sick violinist (Thomson 1976), thought experiments, despite their limitations play an important role of teasing out intuitions, particularly in applied ethics. A distinction needs to be made between radical speculation about future AI, and thought experiments that are meant to be predictive. It is just another thing that AI ethicist do, when they use their training appropriately.

AI ethics expertise and diversity

Staying within the lane of one's expertise is challenging in interdisciplinary domains, especially when the criteria of entry may be unclear. It was perhaps inevitable, then, to see expertise lane-switching into AI ethics. Computer scientists, businesspeople, and even celebrities have all self-proclaimed to be AI ethicists or, even, AI ethics experts. And there is not much anyone can do about it, since the field is new, involves academic philosophy and artificial intelligence, appears to be esoteric to the public, and concerns issues that have very high visibility (especially in the popular media). Although, one might make a plea for the virtue of intellectual humility when working outside of one's primary field of training or in overlapping domains (Snow 2022), it became desirable and fungible to be an AI ethicist. And, until recently, there were not that many people around that fulfilled the high-bar criteria for being an ethicist specialized in artificial intelligence. This has unfortunately exacerbated many other problems with AI ethics.

Another complicating factor is the fact that the different activities that have been assembled under the umbrella of AI ethics are situated in several distinct but overlapping contexts: academic research, policy and regulatory advice, corporate governance, and public/popular debate. As Gijs van Maanen has pointed out, there is substantial "interminglement" between these spheres and thus calls for AI ethicists to make an "explicit acknowledgement of the situatedness of one's ethical research" (van Maanen, 2021, p. 3). Thus, what is needed is transparency about the potential role that politics or power relations play in this work.

There is also a potential danger to emphasizing the role of expertise in moral philosophy, which is important to keep in mind if we follow the approach we recommend

here. First, one has to be vigilant that important expertise and research is never excluded, or that epistemic injustices are not facilitated, reinforced, or created. Requiring from people that contribute to the work of AI ethics that they have a PhD in philosophy is not just a high bar, it creates a danger of excluding important voices and perspectives from the discussion. Often, these can be from the very people who could be disproportionately affected by abuse of AI technology, as in the predictive policing. Including voices from affected communities should be part of the conversation and may be an effective way to raise awareness among other stakeholders of the nature and importance of these issues. It may even prevent unethical use of AI technology before it happens.

There is a well-documented phenomenon that if one searches an online database for images associated with human robot interaction, AI ethics, trustworthy AI, value alignment etc., most commonly one will see the hand of a white man reaching out for the shiny (or snow white) metallic (or plastic) robot hand. These images consistently surface in advertisements for conferences, book covers, in popular media articles, and on research institute promotional materials. Arguably, image results from search engines have an impact on "individual and collective perception of social reality" (Makhortykh, Urman & Ulloa 2021). This is a superficial, but useful way of capturing the reality of the whiteness and maleness of AI research (Cave & Dihal 2020). Although academic researchers in AI ethics have a better gender balance, people who achieve PhDs in philosophy in the United States for example are over 80% white (https://ncses.nsf.gov/surveys/earned-doctorates/2022#tabs-2). What kinds of AI ethics do these demographics contribute to? How can this ensure adequate diversity in approaches and decision-making?

It should be noted that there are active research programs aimed at addressing this. An emerging area of research in AI and robot ethics has to do with justice and representation of race and gender (Gebru 2020; Intahchomphoo & Gundersen 2020; D'ignazio & Klein 2023; Broussard 2023). Scholars in the fields of human-computer interaction and social robotics have also been particularly active in these discussions, as have scholars working on algorithmic fairness and data science (e.g. Cathy O'Neill's 2017 *Weapons of Math Destruction* or Safiya Umoja Noble's 2018 *Algorithms of Oppression: How Search Engines Reinforce Racism,* in the popular press). Some researchers have also increasingly been drawing attention to the other intersecting identities and loci of oppression that are relevant in the ethics of technology and AI ethics, including socio-

economic status, disability, age (Stypinska 2023); ethnicity, and sexual orientation. Future *uses* of AI ethics will likely include a focus on power relations, politics, and the material consequences of AI.

Conclusion

AI ethics is currently more visible than ever, attracting increasing academic, industry, policy, and public attention. This has both positive and negative effects on the quality of work being done. It also creates affordances for abuse, which can render some of the issues that AI ethics aims to address, worse. Here we outlined these abuses and put them against the backdrop of proper uses of AI ethics. We also offered those proper uses of AI ethics, such as identifying responsibility gaps, identification of values and interest, and futuring, are best exercised by professional ethicists with a technical understanding of computer science. Being clear and strict about this and keeping AI ethics expertise within academic philosophy has the potential to limit the impact of affordances for abuse, which are often exploited by interests looking to escape or limit ethics oversight. On the other hand, being clear and strict in that way introduces a risk of exclusion, especially of voices from individuals and communities that may be negatively impacted by AI-powered technologies. This is an important thing to keep in mind, which we highlighted as a new and developing area of AI ethics.

References

Bietti, E. (2021). From ethics Washing to ethics bashing: a moral philosophy view on tech ethics. *Journal of Social Computing*, *2*(3), 266-283.

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, **9**, 1–30.

Brey, Philip AE. "Anticipatory ethics for emerging technologies." NanoEthics 6.1 (2012): 1-13.

Broussard, M. (2023). *More Than a Glitch: Confronting Race, Gender, and Ability Bias in Tech*. MIT Press.

Bryson, J. J. (2010). Robots should be slaves. *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*, 8, 63-74.

Bryson, J. J. (2018). Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology*, 20(1), 15-26.

Burton, S., Habli, I., Lawton, T., McDermid, J., Morgan, P., & Porter, Z. (2020). Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. Artificial Intelligence, 279, 103201.

Cave, S., & Dihal, K. (2020). The whiteness of AI. *Philosophy & Technology*, *33*(4), 685-703.

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics*, 26(4), 2051-2068.

Collingridge David. 1980. *The Social Control of Technology*. New York: St. Martin's Press.

Danaher, John. "Tragic choices and the virtue of techno-Responsibility gaps." P*hilosophy & Technology* 35.2 (2022): 26.

Danaher, J. (2021). What matters for moral status: Behavioral or cognitive equivalence? *Cambridge Quarterly of Healthcare Ethics*, 30(3), 472-478.

D'ignazio, C., & Klein, L. F. (2023). *Data feminism*. MIT press.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and machines*, *14*, 349-379.

Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. *Machine learning and the city: Applications in architecture and urban design*, 535-545.

Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology*, 32(2), 185-193.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, 5, 5.

Gabriel, Iason, and Vafa Ghazavi, 'The Challenge of Value Alignment: From Fairer Algorithms to AI Safety', in Carissa Véliz (ed.), The Oxford Handbook of Digital Ethics (online edn, Oxford Academic, 10 Nov. 2021)

Gebru, T. (2020). Race and gender. *The Oxford handbook of Ethics of AI*, 251-269.

Grunwald Armin. (2009). "Technology Assessment: Concepts and Methods." In Philosophy of Technology and Engineering Sciences, edited by Meijers Anthonie, 1103–46. Amsterdam, the Netherlands: North Holland.

Gunkel, D. J. (2018). The other question: can and should robots have rights?. *Ethics and Information Technology*, *20*, 87-99.

Gunkel, D. J. (2022). The relational turn: Thinking robots otherwise. Social Robotics and the Good Life: The Normative Side of Forming Emotional Bonds With Robots. Edited by Janina Loh and Wulf Loh. Transcript Verlag. Forthcoming.

Heilinger, J. C. (2022). The ethics of AI ethics. A constructive critique. *Philosophy & Technology*, 35(3), 61.

Horgan, T., & Timmons, M. (1991). New wave moral realism meets moral twin earth. *Journal of Philosophical Research*, *16*, 447-465.

Intahchomphoo, C., & Gundersen, O. E. (2020). Artificial intelligence and race: A systematic review. *Legal Information Management*, *20*(2), 74-84.

Illies, C., & Meijers, A. (2009). Artefacts without agency. *The Monist*, *92*(3), 420-440.

Kim, T. W., Hooker, J., & Donaldson, T. (2021). Taking principles seriously: A hybrid approach to value alignment in artificial intelligence. Journal of Artificial Intelligence Research, 70, 871-890.

Klovig Skelton (26 September 2023). " UK government quietly disbands data ethics advisory board." Computer Weekly.com https://www.computerweekly.com/news/366553297/UK-government-quietly-disbands-data-ethics-advisory-board

Kroll, J. A. (2020). Accountability in computer systems. The Oxford handbook of ethics of AI, 181-196.

Makhortykh, M., Urman, A., & Ulloa, R. (2021, April). Detecting race and gender bias in visual representation of AI on web search engines. In *International Workshop on Algorithmic Bias in Search and Recommendation* (pp. 36-50). Cham: Springer International Publishing.

McKay, C. (2020). Predicting risk in criminal procedure: actuarial tools, algorithms, AI and judicial decision-making. Current Issues in Criminal Justice, 32(1), 22-39.

Mohamed, S., Png, M. T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. Philosophy & Technology, 33, 659-684.

Morozov, Evgeny. *To Save Everything, Click Here: The Folly of Technological Solutionism*., Public Affairs, 2013.

Noble, S. U. (2018). Algorithms of oppression. In *Algorithms of oppression*. New York university press.

Nyholm, S., & Frank, L. E. (2019). It loves me, it loves me not: is it morally problematic to design sex robots that appear to love their owners?. *Techne: Research in Philosophy & Technology*, *23*(3).

Nyholm, S. (2023). Responsibility gaps, value alignment, and meaningful human control over artificial intelligence. In Risk and Responsibility in Context (pp. 191-213). Routledge.

O'neil, C. (2017). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown., Chicago.

Quast, Christian, and Markus Seidel. "Introduction: The Philosophy of Expertise—What is Expertise?." Topoi 37 (2018): 1-2.

Risse, M. (2019). Human rights and artificial intelligence: An urgently needed agenda. Hum. Rts. Q., 41, 1.

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. Philosophy & Technology, 34, 1057-1084.

Snow, N. (2022). The value of open-mindedness and intellectual humility for interdisciplinary research. *Scientia et Fides*, *10*(2), 51-67.

Stypinska, J. (2023). AI ageism: a critical roadmap for studying age discrimination and exclusion in digitalized societies. *AI & society*, *38*(2), 665-677.

Sutrop, M. (2020). Challenges of aligning artificial intelligence with human values. Acta Baltica Historiae et Philosophiae Scientiarum, 8(2), 54-72.

Thomson, J.J. (1976). A Defense of Abortion. In: Humber, J.M., Almeder, R.F. (eds) Biomedical Ethics and the Law. Springer, Boston, MA. https://doi.org/10.1007/978-1-4684-2223-8_5

Thompson, D. F. (1980). Moral responsibility of public officials: The problem of many hands. American Political Science Review, 74(4), 905-916.

Tigard, D. W. (2021). (2021) There is no techno-responsibility gap. Philos. Technol., 34, 589–607. https://doi.org/10.1007/s13347-020-00414-7

Umbrello, Steven, et al. "From speculation to reality: Enhancing anticipatory ethics for emerging technologies (ATE) in practice." Technology in Society 74 (2023): 102325.

Van de Poel, I. (2001). Investigating ethical issues in engineering design. *Science and engineering ethics*, *7*(3), 429-446.

Van de Poel, I., & Kroes, P. (2013). Can technology embody values?. In *The moral status of technical artefacts* (pp. 103-124). Dordrecht: Springer Netherlands.

Wagner, B. (2018). Ethics As An Escape From Regulation. From "Ethics-Washing" To Ethics-Shopping?. In E. Bayamlioglu, I. Baraliuc, L. Janssens & M. Hildebrandt (Ed.), *BEING PROFILED: COGITAS ERGO SUM: 10 Years of Profiling the European Citizen* (pp. 84-89). Amsterdam: Amsterdam University Press. https://doi.org/10.1515/9789048550180-016

Winner, L. (1980). Do artifacts have politics?. Daedalus, 121-136.

Winner, L. (1986). Myth information: Romantic politics in the computer revolution. In *Philosophy and Technology II: Information Technology and Computers in Theory and Practice* (pp. 269-289). Dordrecht: Springer Netherlands.

Notes

[1] https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[2] https://artificialintelligenceact.eu/

[3] https://www.computerweekly.com/news/366553297/UK-government-quietly-disbands-data-ethics-advisory-board