AI Ethics: Seven Traps

MARCH 25, 2019 BY ANNETTE ZIMMERMANN

By **Annette Zimmermann** and **Bendert Zevenbergen**

The question of how to ensure that technological innovation in machine learning and artificial intelligence leads to ethically desirable—or, more minimally, ethically defensible—impacts on society has generated much public debate in recent years. Most of these discussions have been accompanied by a strong sense of urgency: as more and more studies about algorithmic bias have shown, the risk that emerging technologies will not only *reflect*, but also *exacerbate* structural injustice in society is significant.

So which ethical principles ought to govern machine learning systems in order to prevent morally and politically objectionable outcomes? In other words: what is AI Ethics? And indeed, "is ethical AI even possible?", as a **recent New York Times article** asks?

Of course, that depends. What does 'ethical AI' mean? One particularly demanding possible view would be the following: 'ethical AI' means that (a hypothetical, extremely sophisticated, fully autonomous) artificial intelligence *itself* makes decisions which are ethically justifiable, all things considered'. But, as philosopher Daniel Dennett argues in a **recent piece in *Wired***, "AI in its current manifestations is parasitic on human intelligence. It quite indiscriminately gorges on whatever has been produced by human creators and extracts the patterns to be found there—including some of our most pernicious habits. These machines do not (yet) have the goals or strategies or capacities for self-criticism and innovation to permit them to transcend their databases by reflectively thinking about their own thinking and their own goals". Of course, reflecting on the kinds of ethical principles that should underpin decisions by future, much more sophisticated artificial intelligence ('strong AI') is an important task for researchers and policy-makers. But it is also important to think about how ethical principles ought to constrain 'weak AI', such as algorithmic decision-making, here and now. Doing so is part of what AI ethics is: which values ought we to prioritise when we (partially) automate decisions in criminal justice, law enforcement, hiring, credit scoring, and other areas of contemporary life? Fairness? Equality? Transparency? Privacy? Efficiency?

As it turns out, the pursuit of AI Ethics—even in its 'weak' form—is subject to a range of possible pitfalls. Many of the current discussions on the ethical dimensions of AI systems do not actively include ethicists, nor do they include experts working in relevant adjacent disciplines, such as political and legal philosophers. Therefore, a number of inaccurate assumptions about the nature of ethics have permeated the public debate, which leads to several flawed assessments of why and how ethical reasoning is important for evaluating the larger social impact of AI.

In what follows, we outline seven 'AI ethics traps'. In doing so, we hope to provide a resource for readers who want to understand and navigate the public debate on the ethics of AI better, who want to contribute to ongoing discussions in an informed and nuanced way, and who want to think critically and constructively about ethical considerations in science and technology more broadly. Of course, not everybody who contributes to the current debate on AI Ethics is guilty of endorsing any or all of these traps: the traps articulate *extreme versions* of a range of possible misconceptions, formulated in a deliberately strong way to highlight the ways in which one might prematurely dismiss ethical reasoning about AI as futile.

1. The reductionism trap:

> **"Doing the morally right thing is essentially the same as acting in a *fair* way. *(or: transparent, or egalitarian, or <substitute any other value>)*. So ethics is the *same* as fairness *(or transparency, or equality, etc.)*. If we're being fair, then we're being ethical."**

Even though the problem of algorithmic bias and its unfair impact on decision outcomes is an urgent problem, it does not exhaust the ethical problem space. As important as algorithmic fairness is, it is crucial to avoid *reducing* ethics to a fairness problem alone. Instead, it is important to pay attention to how the ethically valuable goal of optimizing for a specific value like fairness *interacts* with other important ethical goals. Such goals could include—amongst many others—the goal of creating transparent and explainable systems which are open to democratic oversight and contestation, the goal of improving the predictive accuracy of machine learning systems, the goal of avoiding paternalistic infringements of autonomy rights, or the goal of protecting the privacy interests of data

Freedom to Tinker is hosted by Princeton's Center for Information Technology Policy, a research center that studies digital technologies in public life. Here you'll find comment and analysis from the digital frontier, written by the Center's faculty, students, and friends.

CENTER FOR INFORMATION TECHNOLOGY POLICY
PRINCETON UNIVERSITY

Search this website … Search

What We Discuss

AACS bitcoin CD Copy Protection censorship CITP Competition Copyright Cross-Border Issues cybersecurity policy DMCA DRM Education Events Facebook FCC Government Government transparency Grokster Case Humor Innovation Policy Law Managing the Internet Media Misleading Terms NSA Online Communities Patents Peer-to-Peer Predictions Princeton Privacy Publishing Recommended Reading Secrecy Security Spam Super-DMCA surveillance Tech/Law/Policy Blogs Technology and Freedom transparency Virtual Worlds Voting Wiretapping WPM

Contributors

Select Author... ▾

Archives by Month

- **2020:** J F M A M J J A S O N D
- **2019:** J F M A M J J A S O N D
- **2018:** J F M A M J J A S O N D
- **2017:** J F M A M J J A S O N D
- **2016:** J F M A M J J A S O N D
- **2015:** J F M A M J J A S O N D

subjects. Sometimes, these different values may conflict: we cannot always optimize for everything at once. This makes it all the more important to adopt a sufficiently rich, pluralistic view of the full range of relevant ethical values at stake—only then can one reflect critically on what kinds of ethical trade-offs one may have to confront.

2. The simplicity trap:

> **"In order to make ethics practical and action-guiding, we need to distill our moral framework into a user-friendly compliance checklist. After we've decided on a particular path of action, we'll go through that checklist to make sure that we're being ethical."**

Given the high visibility and urgency of ethical dilemmas arising in the context of AI, it is not surprising that there are more and more calls to develop actionable AI ethics *checklists*. For instance, a **2018 draft report** by the European Commission's High-Level Expert Group on Artificial Intelligence specifies a preliminary 'assessment list' for 'trustworthy AI'. While the report plausibly acknowledges that such an assessment list must be context-sensitive and that it is not exhaustive, it nevertheless identifies a list of ten fixed ethical goals, including privacy and transparency. But can and should ethical values be articulated in a checklist in the first place? It is worth examining this underlying assumption critically. After all, a checklist implies a one-off review process: on that view, developers or policy-makers could determine whether a particular system is ethically defensible at a specific moment in time, and then move on without confronting any further ethical concerns once the checklist criteria have been satisfied once. But ethical reasoning cannot be a static one-off assessment: it required an *ongoing process* of reflection, deliberation, and contestation. Simplicity is good—but the willingness to reconsider simple frameworks, when required, is better. Setting a fixed ethical agenda ahead of time risks obscuring new ethical problems that may arise at a later point in time, or ongoing ethical problems that become apparent to human decision-makers only later.

3. The relativism trap:

> **"We all disagree about what is morally valuable, so it's pointless to imagine that there is a *universal* baseline against which we can use in order to evaluate moral choices. Nothing is *objectively* morally good: things can only be morally good *relative* to each person's individual value framework."**

Public discourse on the ethics of AI frequently produces little more than an exchange of personal opinions or institutional positions. In light of pervasive moral disagreement, it is easy to conclude that ethical reasoning can never stand on firm ground: it always seems to be *relative* to a person's views and context. But this does not mean that ethical reasoning about AI and its social and political implications is futile: some ethical arguments about AI may ultimately be more persuasive than others. While it may not always be possible to determine 'the one right answer', it is often possible to identify at least some paths of action are clearly wrong, and some paths of action that are comparatively better (if not optimal all things considered). If that is the case, *comparing* the respective merits of ethical arguments can be action-guiding for developers and policy-makers, despite the presence of moral disagreement. Thus, it is possible and indeed constructive for AI ethics to welcome value pluralism, without collapsing into extreme value relativism.

4. The value alignment trap:

> **"If relativism is wrong (see #3), there must be *one* morally right answer. We need to find that right answer, and ensure that everyone in our organisation acts in alignment with that answer. If our ethical reasoning leads to moral disagreement, that means that we have failed."**

The flipside of the relativist position is the view that ethical reasoning necessarily means advocating for one morally correct answer, to which everyone must align their values. This view is as misguided as relativism itself, and it is particularly dangerous to (in our view, falsely) attribute this view to everyone engaged in the pursuit of AI ethics. A **recent Forbes article** (ominously titled "Does AI Ethics Have A Bad Name?") argues, *"[p]eople are going to disagree about the best way to obtain the benefits of AI and minimise or eliminate its harms. […] But if you think your field is about ethics rather than about what is most effective there is a danger that you start to see anyone who disagrees with you as not just mistaken, but actually morally bad. You are in danger of feeling righteous and unwilling or unable to listen to people who take a different view. You are likely to seek the company of like-minded people and to fear and despise the people who disagree with you. This is again ironic as AI ethicists are generally (and rightly) keen on diversity."* AI ethics skepticism on the grounds that AI ethics prohibits constructive disagreement means attacking a straw man. By contrast, any plausible approach to AI ethics will avoid the value alignment trap as much as it will avoid relativism.

5. The dichotomy trap:

> **"The goal of ethical reasoning is to 'be(come) ethical'.**

- **2014:** J F M A M J J A S O N D
- **2013:** J F M A M J J A S O N D
- **2012:** J F M A M J J A S O N D
- **2011:** J F M A M J J A S O N D
- **2010:** J F M A M J J A S O N D
- **2009:** J F M A M J J A S O N D
- **2008:** J F M A M J J A S O N D
- **2007:** J F M A M J J A S O N D
- **2006:** J F M A M J J A S O N D
- **2005:** J F M A M J J A S O N D
- **2004:** J F M A M J J A S O N D
- **2003:** J F M A M J J A S O N D
- **2002:** J F M A M J J A S O N D

author log in

Using 'ethical' as an adjective—such as when people speak of 'ethical AI'—risks suggesting that there there are *exactly two options*: AI is either 'ethical' or 'unethical'; or we (as policy-makers, technologists, or society as a whole) are 'ethical' or 'unethical'. We can see this kind of language in **recent contributions to the public debate**: "*we need to be ethical enough* to be trusted to make this technology on our own, and we owe it to the public to define our ethics clearly" or "building *ethical artificial intelligence* is an enormously complex task". But this, again, is too simplistic. Rather than thinking of ethics as an attribute of a person or a technology, we should think of it as an *activity*: a type of reasoning about what the right and wrong to do is, and about what the world ought to look like. AI ethics (and ethics more generally) is therefore best construed as something that people *think* about, and something that people *do*. It is not something that people, or technologies, can simply *be*—or not be.

6. The myopia trap:

> **"The ethical trade-offs that we identify *within one context* are going to be the *same* ethical trade-offs that we are going to face in other contexts and moments in time, both with respect to the *nature* of the trade-off and with respect to the *scope* of the trade-off."**

Empirical evidence and public discussion can present a clear picture of the value tradeoffs and consequences with regards to the introduction of an AI technology in a particular context that may inform governance decisions. However, artificial intelligence is an umbrella term for a wide range of technologies that can be used in many different contexts. The same ethical trade-offs and priorities do not therefore necessarily–and are indeed unlikely to–translate across contexts and technologies.

7. The rule of law trap:

> **"Ethics is essentially the same as the rule of law. When we lack appropriate legal categories for the governance of AI, ethics is a good substitute. And when we do have sufficient legal frameworks, we don't need to think about ethics."**

To illustrate this view, consider the following point from the aforementioned **NYT article**: "Some activists—and even some companies—are beginning to argue that the only way to ensure ethical practices is through government regulation." While it is true that realizing ethical principles in the real world usually requires people and institutions to advocate for their enforcement, it would be too quick to conclude that engaging in ethical reasoning is the *same* as establishing frameworks for legal compliance. It is misguided to frame ethics as a substitute for the rule of law, legislation, human rights, institutions, or democratically legitimate authorities. Claims to that extent, whether it is to encourage the use of ethics or to criticize the discipline in the governance of technology, should be rejected as it is a misrepresentation of ethics as a discipline. Ethical and legal reasoning pursue *related but distinct* questions. The issue of discriminatory outcomes in algorithmic decision-making provides a useful example. From an ethical perspective, we might ask: what makes (algorithmic) discrimination *morally* wrong? Is the problem that is wrongly *generalizes* from judgments about a set of people to another set of people, thus failing to respect them as individuals? Is it that it *violates individual rights*, or that it *exacerbates existing structures of inequality* in society? On the other hand, we might ask a set of legal questions: how should democratic states enforce principles of non-discrimination and due process when algorithms support our decision making processes? How should we interpret, apply, and expand our existing legal frameworks? Who is legally liable for disparate outcomes? Thus, ethics and the law is not an 'either—or' question: sometimes, laws might fall short of enforcing important ethical values, and some ethical arguments might simply not be codifiable in law.

Conclusion

This blog post responds critically to some recent trends in the public debate about AI Ethics. The seven traps which we have identified here are the following: (1) the reductionism trap, (2) the simplicity trap, (3) the relativism trap, (4) the value alignment trap, (5) the dichotomy trap, (6) the myopia trap, and (7) the rule of law trap. We will soon publish a white paper clarifying the role of ethics as a discipline in the assessment of AI system design and deployment in society, which addresses these points in more detail.

FILED UNDER: ARTIFICIAL INTELLIGENCE

Return to top of page