AI and Ethics (2021) 1:61–65
https://doi.org/10.1007/s43681-020-00002-7

OPINION PAPER

Emerging challenges in AI and the need for AI ethics education

Jason Borenstein[1] · **Ayanna Howard**[2]

Received: 15 July 2020 / Accepted: 23 July 2020 / Published online: 6 October 2020

Abstract
Artificial Intelligence (AI) is reshaping the world in profound ways; some of its impacts are certainly beneficial but widespread and lasting harms can result from the technology as well. The integration of AI into various aspects of human life is underway, and the complex ethical concerns emerging from the design, deployment, and use of the technology serves as a reminder that it is time to revisit what future developers and designers, along with professionals, are learning when it comes to AI. It is of paramount importance to train future members of the AI community, and other stakeholders as well, to reflect on the ways in which AI might impact people's lives and to embrace their responsibilities to enhance its benefits while mitigating its potential harms. This could occur in part through the fuller and more systematic inclusion of AI ethics into the curriculum. In this paper, we briefly describe different approaches to AI ethics and offer a set of recommendations related to AI ethics pedagogy.

Keywords AI ethics · Artificial intelligence · Design ethics · Ethics education · Professional responsibility

1 Introduction

Artificial Intelligence (AI) is becoming pervasive. The technology is reaching into so many facets of our lives that we have no choice but to confront its impacts. The creation and deployment of AI is changing our lives and communities in countless ways. These changes are often difficult to understand and anticipate, and are only accelerating due to the ongoing COVID-19 pandemic. Although AI provides observable benefits, the collection, use, and abuse of data used to train and feed into AI, as well as the algorithm itself, may expose people to risks that they were not even aware existed. Employers can monitor workplace performance and behavior in covert and unexpected ways. And a potential employee might be turned down for a job because of the information an automated tool collects while scraping the person's social media profile. A local government might use facial recognition to identify each and every individual that passes through a public area. It was not that long ago that such scenarios would seem farfetched. But now we see a rise in the use of these tools by industry, government, and even academic institutions as they deploy AI algorithms to make decisions that alter our lives in direct, and potentially detrimental, ways. The frequently voiced justification for the use of such AI tools is that they are "better" than a human decision-maker. Should not an algorithm be fair and free of human biases? After all, it shouldn't be burdened with the biases derived from our lived experiences, right? Then again, an algorithm is made by humans, and humans make mistakes, including during the designing, programming, calibrating, and evaluating of the algorithm's performance. Therein lies a key problem: how can fallible humans design AI that effectively lives up to its promised benefits while ensuring its outcomes aren't biased or otherwise harmful? Complicating matters is how do imperfect humans even go about defining "fairness"? It is a messy task, especially considering that the concept has numerous candidate definitions and what counts as "fair" can fundamentally shift over time.

At times, AI is intensifying societal ills, and it would be misleading to imply that a single, simple solution is on the horizon. Fix the bias in the data. Fix the bias within the algorithms. Fix the bias in the outputs. All of these practices may get us closer to mitigating part of the problem; but these

✉ Jason Borenstein
borenstein@gatech.edu

[1] Center for Ethics and Technology, School of Public Policy and Office of Graduate Studies, Georgia Institute of Technology, Atlanta, GA, USA

[2] Linda J. and Mark C. Smith Professor and Chair, School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, USA

are only necessary but not sufficient conditions for fixing them. Addressing these and other ethical concerns requires starting with the root of the problem (i.e., people). Tackling the problem head-on requires educating ourselves at the beginning stages of our interaction with AI–irrespective of whether we are developers, first learning about AI, or users, just starting to interact with AI. The opportunity to learn about how data are used to train AI, about the applications that the AI can enable, etc., should be available to any person that interacts at any stage with AI. If we focus just on those designing AI technology, there is tremendous potential to shape what developers are learning and encourage them to embrace the crucial message that ethics is intertwined with the entire design process (before, during, and after). Moreover, ethics should not be a slapped-on component after-the-fact, a standalone lesson, or a second thought. It is integral at every stage when learning about AI. When we teach the mathematical derivations of a linear regression function for supervised learning in AI, we can also mention the use of disparate impact as a metric to evaluate fairness of the output in the hopes that we move closer to a result that is "correct" and "fairer".

The underlying hypothesis we propose is thus based on the belief that a key measure for potentially adjusting to a world that is rapidly evolving due, in part to an increasing reliance on AI, is to revisit the instruction that future generations of developers (e.g., students) are receiving on AI-related topics. This is especially important given that they may have a direct role in reshaping society as developers-in-training or future adopters of AI technology. In this thought piece, our focus is on the rising need to prepare emerging developers, and working professionals as well, with the skills needed to grapple with the complex and multifaceted ethical challenges emerging from the growing infusion of AI in our day-to-day activities. The educational community, broadly defined, needs to renew its emphasis on nurturing the ability to recognize and engage with ethical issues emerging in relation to AI. Many essential topics are in this space, including the ethical design of AI algorithms, mitigating the risks of AI outcomes, and improving data acquisition and other research practices. Vague concepts of fairness and bias, separated from context or without understanding that people are more than just data or inputs, are not helpful. Within this realm, we highlight recent approaches to AI ethics education, especially as they pertain to current societal concerns.

2 Emerging ethical challenges in AI

AI technology is filtering into our personal and professional lives in countless ways, and not all of its impacts are positive. For instance, AI holds a lot of promise in terms of how it could alter the healthcare landscape. Some claim that AI algorithms could potentially read medical images more quickly than a radiologist could (e.g., [27]). Yet algorithmic bias and other ethical challenges must be overcome to prevent harm to patients. It has already been found, for example, that an AI system used for recommending follow-on healthcare services failed black patients by referring them at a lower rate than their white counterparts even when both groups had a similar diagnosis [26].

Over the past few years, governments and other entities have had a surge of interest in facial recognition. Yet the technology is drawing much scrutiny in part because it is far less reliable when used to identify people who are not white males. In addition, the increasing loss of privacy due to facial recognition is a real worry. During protests sparked by the death of George Floyd, the US government allegedly used facial recognition to identify protesters [19]. Recently, the use of facial recognition software in Detroit resulted in a Black man being falsely arrested for a crime he did not commit [1]. Even though the specific manner in which it might be used is difficult to discern, AI, including facial recognition, might come to play a key role in China's social credit scoring system [4], a system which many find to be ethically problematic. Responses to the use of facial recognition technology include calls from civil liberty groups to regulate this AI tool, along with recent announcements by a number of tech companies that they will purportedly no longer offer their technology to police departments [10]. Yet many thorny ethical issues still need to be resolved.

The contribution of AI to privacy erosion is also intensifying with the advent of tools such as Clearview AI, which can in principle search Internet sources for all of a person's online photos [7]. And given that much of our information is freely available for anyone to scrub when we post it online, without the typical safeguards found in physical infrastructures, it is profoundly difficult to even discern who is using such tools and for which purposes.

Trust in AI technology is another crucial and timely ethical issue. Going back to the aforementioned medical imaging example, if an AI algorithm proves itself "trustworthy", not only could it complement human judgment, it could become an eventual replacement for that judgment. This could perhaps even extend to cover cases the algorithm was not designed to handle (somewhat akin to the practice of "off-label" use of medical products). Or, in a different scenario, during a conversation with a therapy chatbot, the person may begin to trust it and think the technology can provide guidance for circumstances that go beyond the bounds of its programming. Another important facet of (over)trust is that users might believe AI can mitigate harm when it does not have the capacity to do so. For instance, a person wearing a robotic exoskeleton

might assume the device will provide warnings in dangerous circumstances when it does not actually possess that feature [2].

These examples are only a small fraction of the types of ethical issues and challenges circulating around AI [13, 14]. Yet its usage continues to expand. Thus, passively waiting, in the belief that ethical problems will somehow disappear or magically resolve themselves, is not a viable option. We must, instead, be deliberative and proactive in creating not just good AI applications, but ethically sound practices and policies surrounding these applications.

3 Attempts to address AI's ethical challenges

Many initiatives have arisen to address the ethical challenges emerging in relation to AI technology. This includes the drafting of AI ethics documents by a variety of stakeholders, including academic institutions, government agencies, NGOs, and industry. The Montreal Declaration, for example, is largely an initiative from an academic institution and focuses on the responsible development of AI [8]. The professional organization IEEE [16] has drafted a report on the ethics of intelligent systems and is in the process of developing a series of technical standards for such systems. Many companies are highlighting, through press releases or other documents, which ethical issues, such as fairness and transparency, they deem to be important (e.g., Google [9], Deloitte [6]). A sizeable collection of AI ethics is being produced around the globe, which has even led to topical analyses of such documents (e.g., [12, 17]). Whether these documents are generating tangible change, including in terms of new regulations or industry practices, is unclear.

The emergence of organizations such as the Partnership on AI and AI now, and conferences such as ACM FAccT with a mission tied to AI ethics-related issues is a relatively recent occurrence. Funding agencies, such as the National Science Foundation (NSF [25]), are supporting efforts to examine Fairness, Ethics, Accountability, and Transparency (FEAT) in computing fields. This can be taken as evidence of the seriousness that AI ethics should warrant.

Even if these attempts are trying to move the needle in terms of addressing AI's ethical challenges, the fundamental root of the problem remains: that of human fallibility and other related human shortcomings, and how they shape the design and use of technology. While it is doubtful that most people are intentionally designing AI to be malicious in nature or want their systems to deliver biased outputs, the fact remains that they are not consistently being asked to look in the mirror to identify their own biases and the values that they are building into the technology.

4 Fostering a professional mindset

As is illustrated by the above discussion, many individuals and organizations are proposing remedies to the ethical challenges resulting from AI, but solutions (technical or otherwise) are hard to identify and implement. Yet a key piece of the puzzle is enabling developers to understand that the technology they are building is intertwined with ethical dimensions, and that, as developers, they have a vital role and responsibility to engage with ethical considerations. The first aim in establishing an authentic professional mindset is related to cultivating moral sensitivity; in other words, they need the ability to recognize that professional, including "technical", decision-making is intertwined with ethical considerations. The view that technology is "value neutral" hides and obscures the reality that ethical issues are fundamentally embedded in the selection, design, deployment, and use of technology. For example, building a dating app that only offers a binary option for a user's gender is a value-laden choice by the app's creator. When you then integrate AI to identify the user's best match, you are thus constructing a system that has bias woven throughout its design.

A second related point is how those in the AI community view their professional responsibilities. Oftentimes, developers believe (a view sometimes reinforced through the STEM curriculum and in other ways) that ethics is someone else's problem. They may think something like "We deal with the technology; let the lawyers or ethicists resolve the ethical concerns." However, when making choices during the design process, those choices not only have ethical ramifications but they reflect the designer's ethical values (e.g., whether to err on the side of a false positive or a false negative with medical imaging or evaluating recidivism). Such choices not only shape the technology, but they end up shaping individual lives and society more generally.

Taking the example of medicine, physicians may promise to uphold the Hippocratic Oath. While a professional oath is not a panacea, it can serve as a statement of and a commitment to a social contract between a profession and the public. Even if physicians do not literally voice the pledge, the Hippocratic Oath is a reminder of their ethical obligation to improve the health of the public. When AI provides similar benefits, and harms, to the public, what should we expect in terms of the ethical responsibilities of those who develop the technology? Should their responsibilities be anything less? A key step is enabling AI developers and the broader computing community to more fully understand what those responsibilities are.

5 AI ethics instruction

Imparting lessons regarding what is means to be a professional and what one's associated ethical responsibilities are can ideally be achieved through both formal and informal education. Yet at the present time, AI ethics education has not fully taken root within the computing curriculum (e.g., [22]). According to Brundage et al. [3], "Educational efforts might be beneficial in highlighting the risks of malicious applications to AI researchers". In this regard, education can foster a professional mindset for the next generations of AI developers. Of course, if ethics is already taught within the engineering or computing curriculum, this requires evaluating and potentially rethinking how it's done—because either it is not working or not pervasive enough to impact a change in mindset.

Some attempts to incorporate ethics into the curriculum involve a focus on increasing students' familiarity with professional codes of ethics. While an important step, it is not sufficient. Skeptics may point out that just because one is aware of a code, it does not necessarily mean that it will influence behavior; for example, according to a study by McNamara et al. [23], introducing students to the ACM code of ethics did not seem to have a tangible impact on decision-making.

While AI is not the specific target, the Mozilla Foundation [24] is supporting the development of ethics pedagogy in order to try to reshape the computer science curriculum. Funding from the Mozilla program has enabled a team at Georgia Tech, including one of the authors of this editorial, to create an autonomous vehicle role-playing scenario for undergraduate CS courses. The scenario places students in different roles, including ones that are "technical" such as a computer scientist and ones that are "non-technical" such as an active transportation advocate. The students, representing different stakeholders, are supposed to work as a committee to advise a hypothetical city on whether to permit a fictional company to test a self-driving bus fleet in the city's downtown area. We hope that this kind of approach can foster students' moral sensitivity and enable them to appreciate a broader range of perspectives. Yet it is clear that educational efforts must continue to move beyond drop-in modules or single ethics courses [11].

Guided by the aim of nurturing a professional mindset in those who are part of the AI community, we propose three elements that could help familiarize students with the emerging ethical challenges of AI:

1. Teaching the ethical design of AI algorithms; this should include but not be limited to "FEAT" considerations. Learning about the importance of participatory design could also be an important lesson. For example, the new AI ethics course in the online Master of Science in Computer Science program at Georgia Tech, taught by one of the authors of this editorial, has the potential to train a huge generation of AI developers to think through the ethical design of their algorithms (Howard [15]).
2. Incorporating fundamental concepts of data science and the ethics of data acquisition; using real-world data sets that requires students to address privacy, fairness, and legal issues while developing AI solutions.
3. Offering ethics-related lessons in multiple ways and at multiple times; "ethics across the curriculum" is a model for putting this into practice (e.g., [20]), but the general notion is regularly reinforcing the significance of ethics, including in "technical" courses.

A related point is the importance of having interdisciplinary teams who create AI ethics content and potentially teach it. The challenges emerging in relation to AI cross over disciplinary lines and are too complex for any single type of expertise to handle. Insights from lawyers, sociologists, policy scholars, philosophers, and others along with scientists and engineers can be especially valuable when determining how to educate students about AI ethics. This hopefully will attune students to AI's ethical challenges and encourage them to have the willingness to engage with those challenges seriously. Another key facet of AI ethics education is cultivating critical thinking and ethical reasoning skills in students that are transferable across different professional contexts. While there are debates about the value of including ethical theory in professional ethics courses (an issue we will not seek to resolve here), such courses should nurture reason and reflection; they are vital components of the professional mindset.

6 The future of AI ethics

AI is changing our lives in ways that are difficult to anticipate and understand. If the technology is going to be directed in a more socially responsible way, it is time to dedicate time and attention to AI ethics education. Not only is it important for the computing community to more resolutely embrace ethics as a part of its core identity, but from a practical perspective, jobs are starting to emerge in the realm of AI ethics (e.g., [5]). Lewis [21] suggests that some companies may consider having a chief artificial intelligence ethics officer. One hopes that this is part of a sincere effort toward taking ethics more seriously rather than an exercise in "ethics washing" [18]. A pathway towards increasing that likelihood is making sure that ethics has a central place in AI educational efforts.

References

1. Allyn, B.: 'The Computer Got It Wrong': how facial recognition led to false arrest of black man. NPR. https://www.npr.org/2020/06/24/882683463/the-computer-got-it-wrong-how-facial-recognition-led-to-a-false-arrest-in-michig (2020). Accessed 24 June 2020
2. Borenstein, J., Mahajan, H.P., Wagner, A.R., Howard, A.: Trust and pediatric exoskeletons: a comparative study of clinician and parental perspectives. IEEE Trans. Technol. Soc. **1**(2), 83–88 (2020). https://doi.org/10.1109/TTS.2020.2974990
3. Brundage, M., Shahar, A., Jack, C., Helen, T., Peter, E., Ben, G., Allan, D. et al.: The malicious use of artificial intelligence: forecasting, prevention, and mitigation. Future of Humanity Institute, Centre for the Study of Existential Risk, Center for a New American Security, Electronic Frontier Foundation, OpenAI. https://arxiv.org/abs/1802.07228 (2018)
4. Chowdhury, D, Neha, M.: Reports of 'big brother' China social credit system untrue: AI expert Xue Lan. Reuters. https://www.reuters.com/article/us-davos-meeting-lan/reports-of-big-brother-china-social-credit-system-untrue-ai-expert-xue-lan-idUSKBN1ZL2P9 (2020). Accessed 22 Jan 2020
5. Davenport, T.: "What does an AI ethicist do?" World Economic Forum. https://www.weforum.org/agenda/2019/08/what-does-an-ai-ethicist-do/ (2019)
6. Deloitte.: AI ethics: a business imperative for boards and C-suites. https://www2.deloitte.com/us/en/pages/regulatory/articles/ai-ethics-responsible-ai-governance.html (no date)
7. Duffy, C.: The ACLU sues Clearview AI, calling the tool an 'unprecedented violation' of privacy rights. Cnn.com. https://www.cnn.com/2020/05/28/tech/clearview-ai-aclu-lawsuit/index.html (2020). Accessed 29 May 2020
8. Gibert, M., Christophe, M., Guillaume, C.: Montréal declaration of responsible AI: 2018 overview of international recommendations for AI Ethics. University of Montréal. https://www.montrealdeclaration-responsibleai.com/reports-of-montreal-declaration (2018)
9. Google: Artificial intelligence at Google: our principles. https://ai.google/principles/ (no date)
10. Greene, J.: Microsoft won't sell police its facial-recognition technology, following similar moves by Amazon and IBM. The Washington Post, https://www.washingtonpost.com/technology/2020/06/11/microsoft-facial-recognition/ (2020)
11. Greengard, S.: A new class of AI ethics. Commun. ACM (2020). https://cacm.acm.org/news/245121-a-new-class-of-ai-ethics/fulltext
12. Hagendorff, T.: The ethics of AI Ethics: an evaluation of guidelines. Minds Mach. (2020). https://doi.org/10.1007/s11023-020-09517-8
13. Howard, A., Borenstein, J.: The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. Sci. Eng. Ethics J. **24**(5), 1521–1536 (2018)
14. Howard, A., Borenstein, J.: AI, robots, and ethics in the age of COVID-19. MIT sloan management review. https://sloanreview.mit.edu/article/ai-robots-and-ethics-in-the-age-of-covid-19/ (2020).
15. Howard, A.: CS 8803 O10: AI, ethics, and society. Georgia Institute of Technology. https://omscs.gatech.edu/cs-8803-o10-ai-ethics-and-society (no date)
16. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.: Ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems, 1st edition. Piscataway, New Jersey. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html (2019)
17. 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
18. Johnson, K.: How AI companies can avoid ethics washing. VentureBeat. https://venturebeat.com/2019/07/17/how-ai-companies-can-avoid-ethics-washing/ (2019). Accessed 17 July 2020
19. Kanno-Youngs, Z.: U.S. watched George Floyd protests in 15 cities Uusing aerial surveillance. The New York Times, New York (2020)
20. Karoff, P.: Embedding ethics in computer science curriculum. The Harvard Gazette. https://news.harvard.edu/gazette/story/2019/01/harvard-works-to-embed-ethics-in-computer-science-curriculum/ (2019)
21. Lewis, N.: Why having a chief AI Officer should matter to HR. SHRM. https://www.shrm.org/ResourcesAndTools/hr-topics/technology/Pages/Why-Having-a-Chief-AI-Officer-Should-Matter-to-HR.aspx (2020)
22. Macaulay, T.: Study: only 18% of data science students are learning about AI ethics. TNW. https://thenextweb.com/neural/2020/07/03/study-only-18-of-data-scientists-are-learning-about-ai-ethics/ (2020)
23. McNamara, A., Justin, S., Emerson, M.-H.: Does ACM's code of ethics change ethical decision making in software development?" ESEC/FSE 2018. Proceedings of the 2018 26th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering (2018)
24. Mozilla Foundation.: Announcing a competition for ethics in computer science, with up to $3.5 Million in Prizes. The Mozilla Blog. https://blog.mozilla.org/blog/2018/10/10/announcing-a-competition-for-ethics-in-computer-science-with-up-to-3-5-million-in-prizes (2018). Accessed 31 July 2020
25. National Science Foundation.: Artificial intelligence (AI) at NSF. https://www.nsf.gov/cise/ai.jsp (2019)
26. Obermeyer, Z., Brian, P., Christine, V., Sendhil, M.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019). https://doi.org/10.1126/science.aax2342
27. Rajpurkar, P., Irvin, J., Ball, R.L., Zhu, K., Yang, B., Mehta, H., et al.: Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med. **15**(11), e1002686 (2018). https://doi.org/10.1371/journal.pmed.1002686