Artificial Intelligence
Ethics in Practice

JESSICA CUSSINS NEWMAN AND RAJVARDHAN OAK

Jessica Cussins Newman is a Research Fellow at the UC Berkeley Center for Long-Term Cybersecurity where she leads the AI Security Initiative. She is also an AI Policy Specialist for the Future of Life Institute and a Research Advisor with The Future Society. Jessica was a 2016–17 International and Global Affairs Student Fellow at Harvard's Belfer Center, and has held research positions with Harvard's Program on Science, Technology, and Society and the Center for Genetics and Society. Jessica received her master's degree from the Harvard Kennedy School and her bachelor's from the University of California, Berkeley, with highest distinction honors. jessica.cussins@berkeley.edu.

Rajvardhan Oak is a graduate student at the UC Berkeley School of Information. His research interests are security, privacy, and their intersection with machine learning. He obtained his bachelor's degree in computer science from the University of Pune, India. Presently, he is a graduate researcher at the UC Berkeley Center for Long Term Cybersecurity, where he works at the Citizen Clinic and is involved in several public interest projects for low-resource organizations. rvoak@berkeley.edu

This article describes key ethical challenges associated with the design, process, use, and impacts of artificial intelligence. We go beyond naming the problems that have garnered significant attention in recent years, and additionally reference several ongoing efforts to mitigate and manage key ethical concerns. This article is part of a series about ethics intended to encourage ongoing discussion and debate in the research community about ethical considerations that may arise in the course of networking, security, and systems research. We hope that this article will result in researchers as well as industry practitioners being more mindful in their design and use of AI systems.

The role of ethics in AI is sometimes contested, particularly as companies are accused of "ethics washing" in an effort to gain consumer trust while avoiding regulation. Ethics is an important lens for consideration but should not be a substitute for fundamental rights, human rights, or requirements by national and international law. Though this article focuses on AI ethics, it references meaningful intersections with politics, justice, and rights.

Why AI?

From health care to education, from space science to genomic research, AI has revolutionized the way we make decisions. The rise of AI, coupled with the development of computing technology, has allowed us to quickly look at vast amounts of data, discern useful patterns, and use our findings to shape future directions in research or business. Initially intended to be a tool for data analysis and classification tasks, machine learning has now been used to write stories, synthesize images, and even compose music. The desire for AI is understandable; in many cases humans simply cannot match the speed, accuracy, pattern recognition, and large-number-crunching ability of these algorithms. By 2030, AI technologies are expected to contribute $15.7 trillion to the global economy.

What Counts?

The economic promise of AI has led some companies to exaggerate the abilities of their products and services. Companies are able to exploit confusion about what counts as AI because artificial intelligence is an umbrella term, encompassing large sub-fields, including machine learning and deep learning. A simple overarching definition of AI that accounts for the diversity of methodologies actively used is, "a collection of technologies that can enable a machine or system to sense, comprehend, act, and learn" [1]. AI is also considered to be an omni-use technology, meaning that AI technologies have many uses across countless domains, including for good and for ill.

Four Categories

The rise and integration of machine intelligence into the world around us raises numerous ethical challenges, which can be considered to fall within four categories: design, process, use, and impact. The design category includes decisions about what to build, how, and for whom. The process category includes decisions about how to support transparency and accountability through institutional design. The use category includes ways in which AI systems can be used and misused to cause harm to individuals or groups. Lastly, the impact category includes ways in which AI technologies result in broader social, political, psychological, and environmental impacts.

Design

AI systems emerge as a result of numerous human decisions. Many of these may seem innocuous, but they can have profound implications. Most AI systems work by training on large data sets and learning to associate features with outcomes. Machine learning aims to establish a relationship between a target variable and one or more feature variables. It optimizes the parameters of this relation so that the predicted value is as close as possible to the ground truth. However, these systems still make mistakes that a human would never make. Data sets are always imperfect representations of reality and can generate blind spots and biases. Ethical AI is not just judicious use of AI but also thinking carefully about what goes into making these systems.

For example, the tech giant Amazon had been using AI to identify talent and match candidates to jobs since 2014. In 2018, it was discovered that their algorithms systematically discriminated against female candidates. The AI taught itself that the company would prefer male candidates over female ones. For example, according to experts, the system would downgrade candidates from two prominent women's colleges. The algorithm does not know that it is discriminating against women; it simply notices that if it does not select candidates with a certain value (0 or 1) for gender, the results are closer to a defined goal. The problem, in this case, lies in the data that the AI is trained on. Amazon used data that included 10 years of resumes, but only a fraction of them came from women due to women's historical underrepresentation in the technology industry. The AI system, therefore, ranked male candidates over female candidates, since it had seen a greater number of them succeeding.

These failure modes are particularly disturbing when they impact people's livelihoods. In April 2019, over 40,000 residents of Michigan were falsely accused of unemployment fraud based solely on decisions by a machine learning-based computer program. They were forced to repay money, along with substantial penalties. Although the Supreme Court eventually ruled against the governor's office, the fines caused substantial financial burden and even forced some into bankruptcy.

Other design decisions include the composition of engineering teams, and decisions about what technology to build, and for whom. Fewer than 14% of AI researchers are women, and that percentage has decreased over the last 10 years [2]. Racial diversity in AI fares even worse; Google's workforce is only 2.5% Black and 3.6% Latinx, and the percentages at Microsoft and Facebook are similar [3]. The lack of diversity among the teams designing AI systems can also generate blind spots.

For example, AI researcher Joy Buolamwini was a graduate researcher at the MIT Media Lab and found that the facial recognition algorithms she was working with could not “see” her because of her dark skin. She realized that this was not a unique problem; most of the facial recognition community was using the same benchmark data sets for testing the accuracy of models, and the data sets contained extremely limited racial representation. Facial recognition systems were considered to be “accurate” when in fact they were primarily accurate for white men. Joy founded the Algorithmic Justice League to increase awareness about algorithmic bias and develop practices to promote accountability.

Data sets generally reflect historical realities about our world, including structural racism and sexism. When AI systems learn from these data sets, they can then automate and amplify those biases, all while under the veil of technological neutrality. As we rely on algorithmic decision-making in an increasing number of high-stakes environments, including decisions about credit, criminal justice, and jobs, the design and training of the systems should be an area of active consideration.

Process

Just as we need to consider the design of AI systems, we also need to assess the processes in place to support ethical AI. Processes include the implementation of standards and legal requirements, the recognition of principles and best practices, communication with users, and the monitoring of systems' efficacy and impacts. Processes of this kind are necessary to promote transparency and accountability as well as safety.

For example, the utility of massive amounts of data for data analytics and machine learning has contributed to significant privacy breaches. The European Union General Data Protection Regulation (GDPR), which went into force May 2018, is an example of an early regulatory response to help establish data rights and mitigate potential harms from the abuse of personal data. Other data privacy laws have come since, including the California Consumer Privacy Act (CCPA), which went into effect January 2020.

Moreover, it is not common for companies to be forthright about the weaknesses of their models, which can lead to the overestimation of a system's abilities. Unfortunately, we have learned that too much reliance on AI can be dangerous. For example, Uber has been testing their autonomous vehicle technology in Arizona since early 2017. In a shocking incident, a car running in the automatic mode ran a woman over which led to her death. In a similar incident in 2016, a Tesla car running in autopilot mode collided with a truck, leading to the driver's death. Both these cars had human drivers behind the wheel, human drivers who deferred to AI to make the right decision.

Another flaw in AI systems that requires mitigation and monitoring is the susceptibility to *adversarial attacks* [4]. Adversarial examples are those that have been crafted specifically to fool a classifier. Typically, these are constructed by adding a small perturbation to the input. This change is so small that humans cannot identify it; but an algorithm might produce a completely different result. The reason for this is that neural networks, which lie at the heart of most classifiers today, are highly complex and consist of a number of sum functions of logarithms and exponents. As a result, a small change in the input can result in unexpectedly large changes in the output. Research has shown that minor alterations to text, such as dropping a character or capitalizing a letter, can lead to hateful and obscene content being classified as safe. In another example, minute, pixel-level changes to images led a classifier to falsely classify them as facial images.

All machine learning models are capable of making mistakes and being tricked in these ways. And these flaws can be exploited to damaging effect in the real world. For example, researchers have shown how adversarial attacks can be used to confuse medical imaging software, leading to incorrect diagnoses. There are not well-established norms around the mitigation and communication of these risks, and better processes are needed.

Use

Another category of ethical dilemmas associated with AI stems from the technology's broad array of possible uses and misuses. For example, recent advances in AI systems capable of generating synthetic text, audio, and video have beneficial uses, but they can also be used to cause significant harm. Language models can write short stories and poetry, but they can also generate misleading news articles, impersonate others online, automate the production of abusive content, and automate phishing content. *Generative Adversarial Networks* (or GANs) can look at thousands of images of people, learn how faces are constructed, and generate new faces of people who do not exist.

Deepfakes can insert anyone's face into existing video footage, offering a powerful tool for disinformation and information warfare. Doctored videos can quickly spread to millions across social media platforms and can be difficult to detect. Even when quickly proven to be false, doctored videos can have lasting political impact. The rise of deepfakes demands people to be skeptical of what they see, which can breed widespread distrust and corrode democratic processes. For now, however, deepfakes are not widely being used for political destabilization. A study that analyzed thousands of deepfake videos found that the vast majority of deepfakes are being used to create pornographic material, all of which targeted women [5].

Another consequential use of AI is facial recognition technology. In April 2019, it was reported that the Chinese government is using a massive network of facial recognition technology to track and monitor the Uighurs, a largely Muslim minority. The technology has provided unprecedented ability to automate surveillance and repression. Use of the technology has been controversial in the United States as well, where several states have banned the use of facial recognition technology in police body cameras and by law enforcement. Companies have also joined the call for greater regulation of the technology, with Microsoft, for example, calling out the problems of discrimination, privacy abuses, and mass surveillance [6].

AI has also been used to execute military actions. Autonomous weapons are a controversial class of weapons that select and attack targets with limited or no human intervention. Frequently referred to as the third revolution in warfare, after gunpowder and nuclear arms, these weapons may help in reducing human casualties during wars. However, they may also cause terror and destabilization globally; they can be used to conduct assassinations, destabilize nations, and even execute terror attacks on a large scale. In addition to these misuses, these systems are also susceptible to adversarial attacks, biases, and mistakes. Biases in Amazon's systems caused discrimination against women; biases in autonomous weapons can lead to deaths of innocent people.

Impact

AI technologies have economic, political, social, psychological, and environmental impacts that extend well beyond their immediate uses. The long-term impacts to labor markets are one example. The Organization for Economic Cooperation and Development (OECD) estimates that AI and robotics in advanced economies will contribute to radical changes in 32 percent of jobs and fully automate 14 percent of jobs over the next 15–20 years, with disproportionate impacts on low-skilled people and youth [7]. Many countries are now exploring policies to help ease labor transitions for large numbers of people, including retraining programs and social welfare programs.

Another shift that may occur due to AI development is the worsening of economic inequality regionally and between nations. Due to reliance on data and computing infrastructure, AI companies experience network effects, meaning those at the forefront are likely to get increasingly further ahead over time. AI pioneer Kai-Fu Lee has warned that emerging economies are likely to face even greater hurdles as previous pathways to economic growth, for example in China and India, will no longer be available due to the automation of tasks involved in repetitive manual labor of factories and cognitive labor of call centers.

Countries are eager to ensure their economic future and are quickly adopting strategies to generate new talent and innovation. The so-called "race" for AI advancement risks other consequential impacts, however, including international instability and underinvestment in key safety and ethical challenges.

Additionally, AI systems can have long-lasting psychological impacts. For example, e-commerce websites use cookies and demographic data to recommend products to customers. People may feel objectified or unsafe because of the perception that their behavior is being predicted at every step. Most prominent technology platforms also optimize for time spent on their sites, which has led to disturbing advances in “attention hacking” and the facilitation of filter bubbles where people only encounter familiar or provocative content they are likely to engage with. As people communicate more frequently with AI, for example via chatbots, there are also likely to be impacts on human emotions and relationships.

AI also has implications for security infrastructure. Traditionally, security consisted only of the CIA triad; confidentiality, integrity, and authentication. Now, however, there are new loopholes introduced such as susceptibility to adversarial attacks and privacy concerns due to leakage of model parameters. These new vulnerabilities are especially significant for critical infrastructure such as nuclear plants, power grids, and election systems; we now have to ensure security across these additional axes as well.

Lastly, the design and use of AI systems has impacts for the environment. The carbon footprint of training a single AI model has been estimated to result in 284 tons of carbon dioxide—five times the number from an average car over its entire lifetime. Deep learning is particularly energy intensive, as it requires the use of significant computational power for processing vast amounts of data.

Ongoing Efforts

Many institutions are cognizant of the ethical challenges described here and have developed principles to guide their development and use of AI. Notable examples include the Asilomar AI Principles, developed in 2017 through a consultative multi-stakeholder process and signed by thousands of AI researchers and others; Google’s AI Principles, developed in 2018, which notably include categories of AI applications that the company will not pursue such as the development of weapons, illegal surveillance, or technologies that would violate international law and human rights; and the Defense Innovation Board’s recommendations for AI principles to guide the ethical use of AI by the Department of Defense, published in 2019. Also in 2019, the OECD released AI Principles, which have been endorsed by more than 40 countries as well as by the European Commission and the G20, creating the first intergovernmental standard for the responsible stewardship of AI.

More than two-dozen nations have also released national AI strategies, many of which include discussion of how to manage the ethical implications of AI [8]. For example, France and Singapore have developed policy mechanisms to address ethical issues, including impact assessments and an AI ethics advisory council. In the United States, DARPA has a program dedicated to improving the explainability of AI systems, and the NSF has a program to promote fairness in AI systems [9].

In March 2018, The ACM Future of Computing Academy was sufficiently concerned about the negative impacts of advances in computing that they proposed a change to the peer review process, recommending that peer reviewers require papers to consider both positive and negative impacts. Listing the erosion of privacy and threats to democracy among other concerns, they stated, “we can no longer simply assume that our research will have a net positive impact on the world.” The lack of attention to potential negative consequences was described as “a serious and embarrassing intellectual lapse.”

Others have proposed different mechanisms for minimizing misuse. For example, when AI company OpenAI developed a new language model capable of generating paragraphs of text based on any prompt, the company described its concerns about how the tool could have negative societal impacts, and announced that they would engage in a staged release plan [10]. OpenAI only released a small version of the model at the outset and then subsequently released larger models over the course of nine months alongside research papers identifying potential social implications and threats. This process was undertaken with the hopes of providing time for more in-depth research into the technology’s misuse potential.

Another important mechanism that has been proposed to promote transparency and accountability in AI is the idea of Model Cards. In a 2018 paper titled “Model Cards for Model Reporting,” AI researchers proposed that machine learning models should be accompanied by documentation that details their performance characteristics [11]. This is intended to provide benchmarks for evaluation, including whether the model performs consistently across diverse populations, and to clarify intended uses and ill-suited contexts. Model cards are designed to be accessible for both technical and non-technical audiences, and to provide further transparency about how models were trained. Google recently established a web resource to further promote the idea [12].

Algorithmic impact assessments are another tool being used to promote AI accountability [13]. They are intended to examine the use of AI systems; evaluate their impacts on fairness, justice, bias, and other concerns; and to track impacts over time. Additionally, human rights impact assessments, a tool more broadly used for managing the human rights impacts of businesses, projects, and products, are being proposed for use with AI [14]. Predictive policing, targeted surveillance, and disinformation, among other uses of AI, can threaten universal rights. Human rights impact assessments are a key part of the UN Guiding Principles on Business and Human Rights and have been used since 2011. For example, Oxfam America and the Farm Labor Organizing Committee conducted a human rights impact assessment to investigate the state of migrant labor in North Carolina's tobacco industry [15].

Conclusion

AI technologies are not neutral but are created with human goals in mind, taught by human data, and put to use to fulfill human needs; they necessarily have ethical implications. The question is how to increase awareness and establish practices to promote the ethical development of AI that is robust well into the future. Risks of ignoring AI ethics include losing trust from users and the public, as well as pushing away limited talent. The development of ethical AI is a necessary component of sustainable market competition and global leadership.

This article outlined key ethical challenges at stake with artificial intelligence, broken down into four categories of design, process, use, and impact. The article also referenced several ongoing efforts to achieve the goals of ethical AI including principles, strategies, publishing norms, and mechanisms for accountability.

In real-world decision-making scenarios, actors are likely to face tradeoffs between these different considerations. Few ethical guidelines address questions of prioritization, but most organizations will experience the need to decide how to weigh competing values in a given situation. For example, in some cases, there may be a tradeoff between fairness or explainability and accuracy in a machine learning model. Given limited resources, there is also a tradeoff in terms of where to focus.

The need for robust ethical assessment is likely to vary depending on the degree of risk and impact of a given system. However, ethics should not be thought of as an add-on to be considered at the end of production but as a key part of the design process from the outset. Similar to the concept of privacy by design, we need to inculcate the culture of ethics by design. The research community is already at the forefront of many of these debates and is well positioned to play a key role in shaping a positive AI future.

References

[1] "What Is AI Exactly?" *Accenture*, September 21, 2018: https://www.accenture.com/us-en/insights/artificial-intelligence/what-ai-exactly.

[2] "Gender Diversity Crisis in AI: Less Than 14% of AI Researchers Are Women with Numbers Decreasing over the Last 10 Years," Nesta, July 17, 2019: https://www.nesta.org.uk/news/gender-diversity-crisis-ai-less-14-ai-researchers-are-women-numbers-decreasing-over-last-10-years/.

[3] S. M. West, M. Whittaker, K. Crawford, "Discriminating Systems: Gender, Race, and Power in AI," *AI Now*, April 2019: https://ainowinstitute.org/discriminatingsystems.pdf.

[4] L. Huang, A. Joseph, B. Nelson, B. Rubinstein, J. D. Tygar, "Adversarial Machine Learning," Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence, ACM, 2011.

[5] G. Patrini, "Mapping the Deepfake Landscape," *DeepTrace*, July 10, 2019: https://deeptracelabs.com/mapping-the-deepfake-landscape/.

[6] B. Smith, "Facial Recognition: It's Time for Action," Microsoft Blog, December 6, 2018: https://blogs.microsoft.com/on-the-issues/2018/12/06/facial-recognition-its-time-for-action/.

[7] "The Future of Work: OECD Employment Outlook 2019," OECD, 2019: https://www.oecd.org/employment/Employment-Outlook-2019-Highlight-EN.pdf.

[8] J. C. Newman, "Toward AI Security: Global Aspirations for a More Resilient Future," Center for Long-Term Cybersecurity, February 2019: https://cltc.berkeley.edu/wp-content/uploads/2019/02/Toward_AI_Security.pdf.

[9] "Artificial Intelligence for the American People," United States White House, 2019: https://www.whitehouse.gov/ai.

[10] "Better Language Models and Their Implications," *OpenAI*, February 14, 2019: https://openai.com/blog/better-language-models/.

[11] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, T. Gebru, "Model Cards for Model Reporting," *arXiv*, January 14, 2019: https://arxiv.org/abs/1810.03993.

[12] Google, "Model Cards," 2019: https://modelcards.withgoogle.com/about.

[13] D. Reisman, J. Schultz, M. Whittaker, K. Crawford, "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability," *AI Now*, April 2018: https://ainowinstitute.org/aiareport2018.pdf.

[14] "Closing the Human Rights Gap in AI Governance," *ElementAI*, November 2019: http://mediaethics.ca/wp-content/uploads/2019/11/closing-the-human-rights-gap-in-ai-governance_whitepaper.pdf.

[15] Oxfam America and Farm Labor Organizing Committee, "A State of Fear: Human Rights Abuses in North Carolina's Tobacco Industry," 2011: http://hria.equalit.ie/pdf/en/22/A%20State%20of%20Fear.pdf.