IRIE

International Review of Information Ethics Vol. 28 (06/2020)

Katrina Ingram

AI and Ethics: Shedding Light on the Black Box

Abstract:

Artificial Intelligence (AI) is playing an increasingly prevalent role in our lives. Whether its landing a job interview, getting a bank loan or accessing a government program, organizations are using automated systems informed by AI enabled technologies in ways that have significant consequences for people. At the same time, there is a lack of transparency around how AI technologies work and whether they are ethical, fair or accurate. This paper examines a body of literature related to the ethical considerations surrounding the use of artificial intelligence and the role of ethical codes. It identifies and explores core issues including bias, fairness and transparency and looks at who is setting the agenda for AI ethics in Canada and globally. Lastly, it offers some suggestions for next steps towards a more inclusive discussion.

Keywords:

Artificial intelligence, Data, Data Privacy, Ethical Codes, Ethics, Transparency

Outline:

1. Introduction 3

2. Why ethics? 3

3. Who is setting the agenda? 5

3.1. Professional Associations 5

3.2. Research Organizations 5

3.3. Industry 5

3.4. International Governments 6

3.5. Canadian Government 6

3.6. University of Montreal 6

3.7. Human Rights Groups 6

4. Core Issues 7

4.1. Bias and Discrimination 7

4.2. Inclusion and Diversity 8

4.3. Data Privacy 8

4.4. Transparency, Auditability and Accountability 8

5. Conclusion 9

6. References 10

Author:

Katrina Ingram

MA student: Master of Arts in Communications and Technology, University of Alberta

✉ reganing@ualberta.ca

1. Introduction

Artificial Intelligence is a game changing technology. It is already impacting our lives in ways we do not always see. As research advances and more AI technology is deployed, from autonomous vehicles, to the Internet of Things, to a wide range of AI enabled algorithms and robots, it will have an almost ubiquitous reach.

This raises a host of questions and issues around how AI is developed and deployed. There are concerns about bias, fairness and transparency, as well as safety. Issues around privacy and control of data also need to be examined. Data is a key input in training artificial intelligence systems. Reams of data are being gathered and used to train systems in a "wild-west" approach. There is a lack of transparency about what is being gathered, by whom and for what purposes. The people generating this information are not told how their information may be used and what datasets are being combined.

AI systems are being deployed that are making decisions which in some cases can impact lives. The people being impacted by these technologies often do not know or understand the implications. There are several documented cases where decisions made by AI technologies have adversely impacted people's lives without a clear indication or explanation of why and how these decisions are being made. It is not clear who is setting the agenda and what standards or rules, if any, are being applied. If AI is going to impact all of us, shouldn't we have the opportunity to weigh in on the discussion?

During a G7 summit in December 2018, Canada took the lead on hosting a multi-stakeholder conference on artificial intelligence where some of these issues were acknowledged as key challenges that need to be addressed.

> "Artificial Intelligence opens up wonderful opportunities, but democratic societies will need to address major challenges to ensure that the positive effects of algorithm development and use are shared equitably. In particular, we must ensure that AI growth does not amplify existing prejudice or lead to the exclusion of vulnerable populations" (AI for Society).

This paper outlines some of the core ethical issues to consider in ensuring we develop and deploy AI in ways that are beneficial to society. It takes a non-technical approach with the goal of being as inclusive as possible to provide some perspectives on these issues to a wide range of people.

Ethics itself is an enormous field and even ethics, as it applies to AI, is a broad topic. To further focus this paper, the discussion will centre on existing applications and how these technologies are impacting people today. Ethical issues around super-intelligence, sentient robots, the singularity and other future oriented technologies are fascinating, but beyond the scope of this paper. Furthermore, if we can successfully address today's issues, its likely we will be much better equipped to address those future topics.

Lastly, this paper is premised on the idea of moving forward with artificial intelligence. While its easy to be fearful of new technologies, in general historical terms, new technologies tend to advance society and human well-being overall. It is important to consider the many beneficial aspects of AI as we work to address ethical concerns. This paper advocates taking a balanced and responsible approach.

2. Why ethics?

If we agree with the premise that artificial intelligence will have a vast impact on society then we need to decide how we want to shape it to ensure it benefits humans and aligns with societal goals. Ethics is a starting point to determine what values we want to uphold in the development, design and deployment of artificial intelligence.

> "The extension, enhancement and replacement of human agency and reasoning in AI serve as the loci of many of the ethical issues that arise in its use, sometimes presenting us with vivid versions of old questions" (Boddington 29).

These age-old questions take us back to various philosophical schools of thought. "Aristotle emphasized virtues, Immanuel Kant emphasized duties and utilitarians emphasized the greatest happiness for the greatest number" (Tegmark 269). There are also deontological theories that emphasize "doing the right thing" and consequentialist theories that claim the best action is the one that drives the best consequences (Boddington).

While there is not a consensus on what ethical theory should prevail, there are general ethical principles that can be applied. These principles in general, relate to questions of fairness and justice (Boddington).

For the purposes of this paper, we will focus on the following areas of ethical concern: bias, fairness and transparency.

One example that illustrates these issues in practice is the case of Amazon's hiring algorithm that inadvertently created a bias against women. "The company's experimental hiring tool used artificial intelligence to give job candidates scores ranging from one to five stars - much like shoppers rate products on Amazon…They literally wanted it to be an engine where I'm going to give you 100 resumes, it will spit out the top five, and we'll hire those" (Dastin). However, the company recognized that the new system was not recommending female candidates for certain technical jobs, an error caused by a decade's worth of AI training data that reflected a male gender bias for these roles. Even after attempting to correct the situation, the system still found ways to discriminate. Its decision-making process was not even transparent to those who designed it. The project was eventually scrapped in 2017, three years after the project launched. The story was first reported by Reuters in 2018 after a tip from whistleblowers that were close to the project and came forward on the condition on anonymity (Dastin).

As we consider ethics from the perspective of protecting society in the design, development and deployment of artificial intelligence, it's also important to ask if ethics alone forms an adequate response to these concerns. The construction of ethical codes is a form of soft regulation or self policing. Its one way to get "ahead of government in an effort to shape the regulatory framework that could eventually govern the use of AI" (Serebrin). Thus, it can be an appealing starting point for industry and the AI research community. This is not to dismiss ethics as a consideration, but rather, to point out that the ethics discussion itself may lean towards these interests, and therefore, may have limitations in fully addressing societal concerns. This is noted in the *AI Now 2018 Report* which states that, "Ethical codes can only help close the AI accountability gap if they are truly built into the processes of AI development and are backed by enforceable mechanisms of responsibility that are accountable to the public interest" (Whittaker et al. 9).

It is also important to at least touch on the context around codes of ethics and why we have them at all. As Paula Boddington notes in *Towards a Code of Ethics for Artificial Intelligence*, often "codes of ethics (and laws and other regulations) have developed in response to catastrophes or scandals" (Boddington 49). However, she also says this worst-case scenario approach is far from ideal. It's a better scenario to develop codes that anticipate possible problems, and in the case of AI ethics, that's the approach being attempted (Boddington). While AI ethical codes will build on other existing codes as it pertains to specific industries (i.e. healthcare), there are some ethical issues being raised that are distinctive to AI. "Characteristic ethical questions regarding AI concern is typical enhancement or replacement of human agency; crucially, questions of agency are at the heart of how we see ethics" (Boddington 27).

3. Who is setting the agenda?

It is important to take a look at who is setting the agenda for AI ethics and defining what ethical guidelines are to be put into place. Given the far reaching, high stakes consequences, it seems necessary to be as inclusive as possible in framing this conversation. However, at the moment, the discussion is primarily taking place in a fairly limited and *ad hoc* fashion. In most cases it is being led by the AI research community with support from industry. Some government agencies are also starting to participant in this conversation. It is worth noting that the majority of the work has taken place in the last 12-24 months with most of it being less than a year old.

3.1. **Professional Associations**

The Association for the Advancement of Artificial Intelligence (AAAI) is the North American association for AI researchers. Searching the term "ethics" on the AAAI website yielded only a handful of technically focused papers and reports. There does not appear to be any development of an ethical framework, code or standards by AAAI.

The Institute of Electrical and Electronics Engineers (IEEE) is a global technical professional organization of 420,000 members dedicated to advancing technology for the benefit of humanity. The IEEE has published a book on ethical design principles specific to AI called *Ethically Aligned Design*. It also has a community of 2,000 people, largely AI designers, who are working to set standards as part of the Global Initiative on Autonomous and Intelligent Systems working group.

3.2. Research Organizations

The Machine Intelligence Research Institute (MIRI) seeks to "ensure that the creation of smarter-than-human intelligence has a positive impact" (MIRI). Two of its researchers, Eliezer Yudkowsky and Katja Grace, are doing work directly aligned with AI ethics.

Max Tegmark, an MIT Professor, formed the Future of Life Institute with funding from Elon Musk. This organization is behind the Asilomar Principles, an ethical code developed by an elite group of AI insiders in early 2017.

The Future of Humanity Institute led by philosopher Nick Bostrom is based in the UK. It's largely focused on the topic of existential risk and touches on AI as a potential existential risk.

The AI Now Institute is an interdisciplinary research institute based out of New York University. It is focused on taking more of a "watch dog" approach to the core issues. It releases an annual state of the union report which outlines key recommendations that address ethical considerations.

3.3. Industry

Some corporations that are conducting work using artificial intelligence have developed their own ethical codes of conduct or ethical guidelines. These include larger organizations such as Google and IBM as well as a long list of smaller players. How these codes are implemented in practice and whether we should trust industry to police itself are obvious questions to ask about the efficacy of corporate guidelines.

3.4. International Governments

The House of Lords in the UK has drafted a recommended approach to AI in April 2018 which includes a small reference to ethics. The European Union has also released a draft ethics guideline for trustworthy AI in December 2018 which aims to:

> "Ensure that AI is human-centric: AI should be developed, deployed and used with an 'ethical purpose,' grounded in, and reflective of, fundamental rights, societal values and the ethical principles of Beneficence (do good), Non-Maleficence (do no harm), Autonomy of humans, Justice, and Explicability" (European Commission 5).

3.5. Canadian Government

There are a few government initiatives in Canada taking place at the federal level. The G7 Multi-stakeholder Conference on Artificial Intelligence, which took place in Montreal in December of 2018, involved the discussion of a several working papers which touched on ethical issues. The Treasury Board has also created a directive on the use of Automated Decision Systems which came into effect on April 1, 2019 and applies to Automated Decision Systems procured after April 1, 2020.

> "The objective of this Directive is to ensure that Automated Decision Systems are deployed in a manner that reduces risks to Canadians and federal institutions, and leads to more efficient, accurate, consistent, and interpretable decisions made pursuant to Canadian law.
>
> "The expected results of this Directive are as follows:
>
> - Decisions made by federal government departments are data-driven, responsible, and complies with procedural fairness and due process requirements.
> - Impacts of algorithms on administrative decisions are assessed, and negative outcomes are reduced.
> - Data and information on the use of Automated Decision Systems in federal institutions are made available to the public, where appropriate" (Government of Canada).

3.6. University of Montreal

The Montreal declaration was driven by the University of Montreal with support from various partners in Quebec, including MILA, Quebec's artificial intelligence institute (Montreal Declaration). The process was more inclusive than most, with a series of citizen discussions led by industry experts as part of the methodology. The two co-chairs were notably both female and both from non-computer science backgrounds. While the process was inclusive within a Montreal/Quebec context, it was limited in its scope. The code which consists of seven principles, all fairly high level in nature, has been signed by over 1100 citizens and 28 organizations.

3.7. Human Rights Groups

A coalition of human rights groups including Access Now and Amnesty International came together in 2018 in Toronto to produce the Toronto Declaration. The declaration is an appeal to align artificial intelligence ethical guidelines with existing Human Rights laws and standards.

> "As the 'ethics' discourse gains ground, this Declaration aims to underline the centrality of the universal, binding and actionable body of human rights law and standards, which protect rights and provide a well-developed framework for remedies. They protect individuals against

> discrimination, promote inclusion, diversity and equity, and safeguards equality. Human rights are 'universal, indivisible and interdependent and interrelated" (Sterling).

It is still early days for discussions around ethics and AI. There are many players missing from the discussion which seems to be primarily driven at the global level by US and UK interests. China is an obvious player that is missing from the international discussion on ethics, though it appears to be having some national dialogue around the issue now. It has recently appointed top AI scientist, Chen Xiaoping, to lead an ethics committee on behalf of the state sponsored Chinese Artificial Intelligence Association (Zhang).

Also, the Montreal Declaration aside, the views of average people who are not technologists, but will be impacted by the technology, are not being factored into the discussion yet. There are some hyper-local initiatives taking place, such as meetup groups, which are forming to have discussions but there is still much work to be done to make this conversation truly inclusive.

4. Core Issues

There are many issues that could be examined in terms of AI ethics. This paper focuses on bias, discrimination, fairness, transparency, and privacy. These are issues that are impacting people right now and need to be addressed to correct current deployments of AI technologies as well as prevent future technologies from exhibiting dangerous design flaws or being used in ways that perpetuate inequality. Broadly speaking, these issues can be addressed through creating the conditions for diversity and inclusion, accountability, and trust in both the development and deployment of AI technologies.

4.1. Bias and Discrimination

A briefing document prepared by the technology research organization, Forrester, titled *The Ethics of AI: How to Avoid Harmful Bias and Discrimination states:* "By their very nature, machine learning algorithms can learn to discriminate based on gender, age, sexual orientation or any other perceived differences between groups of people (Purcell 2).

The document outlines how bias can be baked into machine learning models through various types of bad data which include:

- Incomplete datasets, such as a lack of sufficient data to accurately reflect the population. For example, facial recognition programs that do not have enough data for darker skinned individuals can create racist outcomes (Purcell).
- Data sets containing errors. Most data sets contain errors and have to be "cleaned." Cleaning can be a time-consuming and expensive process, and thus some companies would prefer to "look the other way."
- Historical bias that gets encoded when a model uses a proxy for race, age or another unethical discriminator. For example, the majority of single parents in the US are female, thus the resulting model can use single parent as a proxy for female, creating gender bias in the model (Purcell).

The report goes on describe a FAIR model, which stands for creating models that are Fundamentally sound, Assessable, Inclusive and Reversible (Purcell). It is one example of how companies developing AI might approach correcting their data sets if necessary and ensuring they can explain their models. There is self-interest in this approach for companies whose brand, reputation and revenue are ultimately at risk if they do not attempt to address these issues.

4.2. Inclusion and Diversity

Inclusion is part of a sustainable solution to address bias, and it was a key issue at the G7 Multi-stakeholder Conference on Artificial Intelligence in Montreal in December 2018. Some of the questions outlined in a discussion paper on this topic were:

> "How do we make sure gender, social and cultural diversity fuel AI design and development? What are the best practices examples of inclusion and diversity in AI and how can we replicate them? What can government, academia, and industry do to promote an inclusive use of AI technology that will benefit a diverse society?" (AI for Society).

One of the best ways to ensure bias does not enter into the design of AI technologies is to ensure the people designing the technologies represent a diverse perspective. "The tech industry in general suffers from an underrepresentation of female, black and Hispanic employees" (Purcell).

In it is 2018 report, AI Now calls for fairness, accountability and transparency of the "full stack" supply chain (Whittaker et al). This is an attempt to account for all aspects of the components that make up the AI system.

> "For meaningful accountability, we need to better understand and track the component parts of an AI system and the full supply chain on which it relies; that means accounting for the origins and use of training data, test data, models, application program interfaces (APIs), and other infrastructural components over a product life cycle" (Whittaker et al. 5).

4.3. Data Privacy

The topic of data privacy is vast and could easily be the focus of an entire paper. It needs to be mentioned as part of the ethical considerations pertaining to AI ethics because data is a key input for training AI systems. As it related to AI ethics, the issue is one of using data for training AI without the proper consent or visibility from those who's data is being used. In general, most countries do not have strong data protection laws and as such, companies are gathering data and using it with little to no oversight.

4.4. Transparency, Auditability and Accountability

Two core ethical question that need to be addressed are how a decision is made and who is responsible or accountable. These questions relate to issues of transparency and auditability – being able to see and understand how the process works, so that there can be accountability. This is an important concept not only from an ethical standpoint but form a legal standpoint.

> "Accountability is about a clear acknowledgement and assumption of responsibility and 'answerability' for actions, decisions, products and policies….building explainability into the AI systems…determining which individuals or groups are accountable for the impact of AI algorithms….and as a feature of the broader sociotechnical system that develops, procures, deploys and uses AI" (AI for Society).

Currently there is an accountability gap between those who develop and profit from AI and those most likely to suffer the consequences and that gap is growing larger (Whittaker et al.).

Finally, its not enough to apply any of these issues in a “one and done” fashion because AI systems learn and evolve. Thus, what might be onside at the time its deployed, may, with new datasets and new learning, move out of alignment. Therefore, its important to continually monitor these systems to ensure that they remain in alignment with the intended goals and objectives. This monitoring aspect is part of the ethical framework.

This is by no means an exhaustive list of ethical issues, but it does serve to highlight some of the major issues that need to be addressed.

5. Conclusion

There seems to be general alignment with the need for AI ethics across industry, academia, government and the AI research community. Efforts made to date, for the most part, have been largely individual and limited to a particular organization or government rather than a more holistic, coordinated approach. They’ve also largely been centered on the topic of designing and deploying AI as a discrete area rather than informing the topic of AI ethics within the context of the domain in which it will be applied (i.e. Healthcare, Education, Employment, etc.).

On March 26, 2019, Google announced that it was launching an AI ethics external advisory board to “consider some of Google's most complex challenges that arise under our AI Principles” (Walker). Yet, the exercise quickly turned into a PR crisis. The composition of the board itself raised ethical issues, fueling protests by Googlers and an online petition with calls to remove a controversial board member. This in turn led to the resignation of others on the board and ultimately the effort was disbanded a mere 10 days after it was announced (Levin). Any efforts by other corporations to attempt to self-govern on the issue of AI ethics is now set against that backdrop.

At the heart of the matter is trust and inclusivity. Members of the general public haven’t had an opportunity to weigh in on these issues. Public engagement and consultations need to take place to ensure a broad range of perspectives are considered. This needs to be done in tandem with educating the public about artificial intelligence, presenting not only the risks, but also the benefits of the technology, in a way that is balanced, objective and as agenda-free as possible. Who should take on this work? Government? The AI research community? A not for profit? Thus far, there is not a clear answer. For practical reasons, this work needs to take place at a local or national level which could feed into a global conversation. At a global, macro-level, AI researchers and product designers would benefit from having a governing body to help define appropriate standards that are based on ethical principles.

There is a tremendous amount of work to be done to address the issues that have been raised by the development and deployment of AI technologies. Its still early days but the challenges will only become more difficult to address the longer we wait.

6. References

AI for Society. *Inclusion in AI Development and Deployment*, 2018. Retrieved from - https://www.ic.gc.ca/eic/site/133.nsf/vwapj/1_Discussion_Paper_-_AI_for_Society_EN.pdf/$FILE/1_Discussion_Paper_-_AI_for_Society_EN.pdf

Amnesty International & Access Now. *The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems*, 2018. Retrieved from - https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf

Boddington, P. *Towards a Code of Ethics for Artificial Intelligence*. Springer. UK, 2017

Bostrom, N. Yudkowsky E. *Draft for Cambridge Handbook of Artificial Intelligence*, eds. William Ramsey and Keith Frankish (Cambridge University Press, 2011), 2011. Retrieved from - http://faculty.smcm.edu/acjamieson/s13/artificialintelligence.pdf

Dafoe, A. *AI Governance: A Research Agenda. Future of Humanity Institute*, University of Oxford, 2018. Retrieved from - https://www.fhi.ox.ac.uk/wp-content/uploads/GovAIAgenda.pdf

Dastin, J. *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters, 2018. Retrieved from - https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

European Commission. *Draft ethics guidelines for trustworthy AI*, 2018. Retrieved from - https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai

Government of Canada. *Directive on Automated Decision Making*, 2019. Retrieved from - http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

IEEE. *Ethics in Action*, 2019. Retrieved from - https://ethicsinaction.ieee.org/?utm_campaign=EAD1e&utm_medium=ieeeorgstdpg&utm_source=Web&utm_content=report#join

Levin, A. "Google scraps AI ethics council after backlash: 'Back to the drawing board'". *The Guardian*, 2019. Retrieved from - https://www.theguardian.com/technology/2019/apr/04/google-ai-ethics-council-backlash

Parliament.uk. *UK can lead the way on ethical AI, says Lords Committee*, 2018. Retrieved from - https://www.parliament.uk/business/committees/committees-a-z/lords-select/ai-committee/news-parliament-2017/ai-report-published/

Purcell, B. *The Ethics of AI: How to avoid harmful bias and discrimination*, 2018. Retrieved from - https://www.ibm.com/downloads/cas/6ZYRPXRJ

Serebrin, J. "E is for ethics in AI – and Montreal's playing a leading role". *Montreal Gazette*, 2019. Retrieved from - https://montrealgazette.com/news/local-news/can-montreal-become-a-centre-not-just-for-artificial-intelligence-but-ethical-ai

Sterling, B. "The Toronto Declaration on machine learning: Preamble". *Wired*, 2018. Retrieved from - https://www.wired.com/beyond-the-beyond/2018/05/toronto-declaration-machine-learning-preamble/

Tegmark, M. *Life 3.0: Being Human in the Age of Artificial Intelligence*. First Vintage Books. New York, 2017.

Walker, K. *An external advisory council to help advance the responsible development of AI*, 2019. Retrieved from - https://www.blog.google/technology/ai/external-advisory-council-help-advance-responsible-development-ai/

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S.M., Richardson, R., Schultz, J., Schwartz, O. "AI Now Report 2018". *AI Now Institute*, NYU, 2018. Retrieved from - https://ainowinstitute.org/AI_Now_2018_Report.pdf

Zhang, P. "China's top AI scientist drives development of ethical guidelines". *South China Morning Post*, 2019. Retrieved from - https://www.scmp.com/news/china/science/article/2181573/chinas-top-ai-scientist-drives-development-ethical-guidelines