Ethics of Artificial Intelligence in the North-American Workplace

for
Dr. Teslenko
MECH226 Instructor
University of British Columbia

by

December 7, 2017

Contents

List of Figures

List of Tables

Abbreviations

AI artificial intelligence

MLAI machine learning and artificial intelligence

NN neural network

NSTC National Science and Technology Council

R&D research and development

SA sentiment analysis

Abstract

Artificial intelligence (AI) is becoming increasingly prominent in the workplace, with systems appearing everywhere from smartphones to cars. With this increase of AI systems, more ethical issues regarding their usage are becoming apparent. AI systems are are incredibly complicated and often considered black boxes due to the lack of understanding and trust society has with regards to their inner workings. The use of neural networks and data testing using known inputs and outputs, pose the challenge of verifying whether AI is suitable for a task. Using AI in daily life demands significant human computer interaction, which means that any biases inherent to AI will propagate to human interaction. These biases, usually the result of impressionable programming, data acquisition, and machine learning, are detrimental to workplace accessibility and equality. When AI makes mistakes, the question of liability and accountability arises. Two options exist for holding accountability: the AI system, and creator. Due to these ethical issues, guidelines must be created and observed by both users and creators. Organizations have already begun recommending AI guidelines following humanitarian law for a more ethical use of AI. Unless these issues are acknowledged, they risk multiplying as AI manages more responsibility and riskier tasks in the future.

1 Introduction

The notion of artificial intelligence has long been associated with progress and prosperity. Yet its recent emergence in modern life may not prove as much. The influx of artificial intelligence uses in the North-American workplace has introduced various ethical dilemmas. Specifically, it has reshaped communication to necessitate less and less human interaction by augmenting human-computer interaction. Despite the various advantages to automation, the ethical concerns it raises call into question its appropriate extent in the workplace, and beyond. Based on information collected from reputable sources made available through UBCs library databases, the main ethical concerns have been identified. The transparency, bias, accountability,and regulation of AI are being examined.

2 Allocation of Accountability

2.1 Assessing Responsibility

With the noticeable increase of AI use in the workplace and potential institutional dependency on it, several concerns arise. While AI functions by analyzing its surroundings, generating alternatives and finding the best outcome it deems possible, AI ethical decision making remains questionable by humans. As such, it is crucial to address the question: who is held accountable for risks arising from AI decision making?

AI accountability can be assessed through two approaches. The first approach, the classic approach, views machines as slaves or mechanical instruments controlled and owned by humans, thus bearing no responsibility. The second approach, the pragmatic approach, on the other hand, holds AI accountable for their actions and decisions under the coat of 'artificial morality' (Alaieri and Vellino, 2016). This, however, is dependent on how the AI was programmed initially. If the AI was programmed using the the 'top-down' method, where ethical codes are embedded by the programmer, then the coder is fully responsible for the actions for his or her AI. There may be inherent problems with this approach, Liu notes "guiding ethical frameworks overlook compound or aggregated effects which may arise, and which can lead to subtle forms of structural discrimination(Liu, 2017, p. 1). But if the AI was coded using the 'bottom up' method, where AI is able to learn for its surrounding and learn from experience, then the AI is fully responsible of its actions. The AI, its manufacturers as well as its users can all be held accountable for any negative outcomes resulting from the AI.

There are many use cases for AI, particularly in the case of user error or unscrupulous behaviour by users, AI outcomes can be undesirable. In such cases, of course, the user would be responsible for any outcomes. For example, the Microsoft's Twitter chatbot was noticed to be tweeting sexist and racist responses. This is due to the fact that the intelligence improved through experiential learning, Twitter users having learnt this, intentionally made sexist and racist comments in an effort to impact its decisions (Alaieri and Vellino, 2016).

2.2 Applications and Practice

AI is arguably better at 'rational' decision making than humans but the lack of human trust in AI decision making sustains a fear of negative outcomes. AI will give the output without any further explanation, even if the decision making process is opaque. This means that in some cases AI makes the 'rational' response but human perception may not reach the same conclusion and hence think the decision made by the AI was unethical (Alaieri and Vellino, 2016).

Thus for a more practical way of allocating accountability, it is suggested that a certifying agency be created under the purview of the Artificial Intelligence development Act (AIDA) that evaluates the safety of AI. If manufacturers pass the certificate, then they would hold limited liability of whatever the outcomes of their AI produces, however, if an AI product isn't 'certified' then the producer/programmer is to be hold responsible for the negative or unethical decisions made by their AI (Scherer, 2016). The accountability of AI thus, depends on multiple factors and differs when looked at from multiple perspectives. This indicates that more work should be put onto the regulations of the use of AIto resolve the ethical concerns of its decision making in the workplace.

3 Unintentional Bias of AI

The notion that AI can be objective is an attractive prospect. Using AI in lieu of a human could serve in eliminating any preconceptions or judgements that are unrelated to the task at hand. This symbolizes an advantageous elimination of unethical treatment in the workplace, and beyond. However, much like the people that make them, AI can equally be riddled with biases. The plausibility of a truly bias free AI is under question.

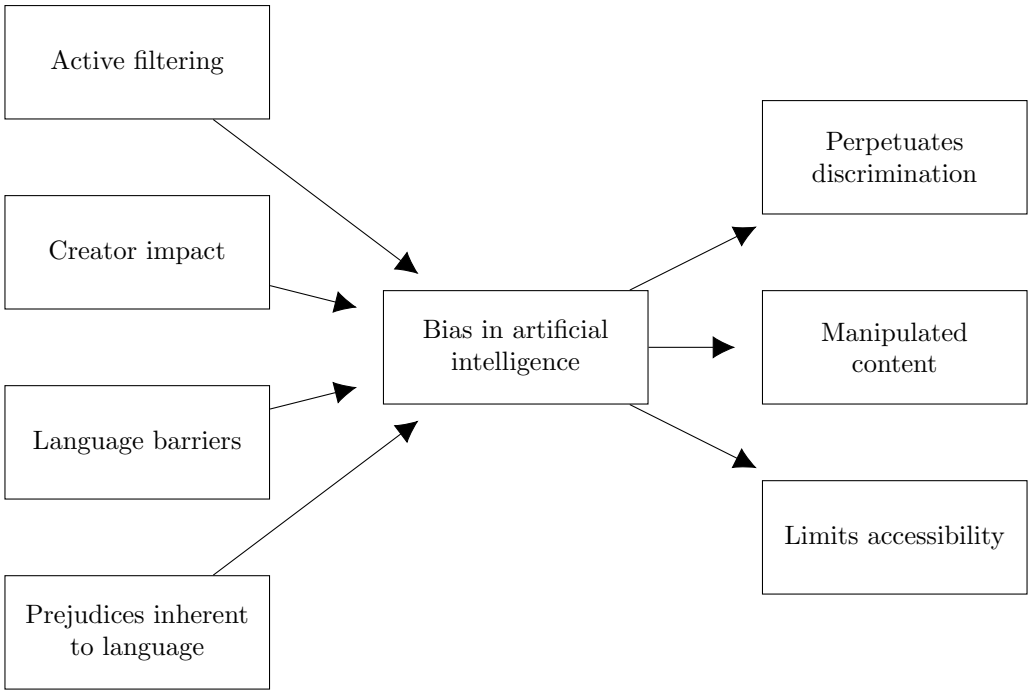

Figure 1: The Constituents and Potential Effects of AI Biases. Image created by [redacted].

3.1 Prejudice and Accessibility

With the ever-increasing presence of AI in the workplace, interactions with this technology become more significant. It is a powerful tool, but it can equally be detrimental. AI parameters may decrease accessibility and perpetuate stereotypes by the means illustrated in Figure 1.

A popular use of AI consists of virtual personal assistants, such as Apple's Siri or Microsoft's Cortana. Yet their use is not available to all. Voice automated technologies often fall short when it comes to recognizing various dialects or non-standardized speech (Hirayama et al., 2015). This bars people with such manners of speech from using a helpful technology that facilitates work.

Human interaction with AI involves a heavy amount of filtering. It is often used to better suit a user's needs. Yet it can also serve as a method of advertisement and withhold useful information. The bias in the filtering process is a product of personalized algorithms, the influence of the creator, and intentional content selection by the technology's owner (Bozdag, 2013). This can limit AI's flexibility and range in informational output, providing the user partial results.

The characterization of biases within AI is inexact and the bestowment of the title is often contested. Nonetheless, there exists biases in AI that are undeniable biases. Sexism, racism, and the likes are ethical dilemmas to which AI are susceptible. The documented instances of racial prejudice in AI (Danks and London, 2017) risk perpetuating harmful stereotypes and racism.

She	**He**
Homemaker	Maestro
Nurse	Protégé
Receptionist	Philosopher
Librarian	Captain
Socialite	Architect
Hairdresser	Financer
Nanny	Warrior

Table 1: The Strongest Word Vector Associations with "She" and "He" Found in Word2vec[1]. Created with data from Bolukbasi et al. (2016)

Methods of machine learning, such as semantics achieved through word embedding[2], are liable to precisely reflect the biases seen in humans (Caliskan et al., 2017). This compromises any hope of objectivity and holds ethical ramifications.

For example, Bryson, Caliskan, and Narayanan noted how the word she was associated with words pertaining to domestic roles, examples of these associations are available in Table 1. This is counteractive to the sociopolitical progress made by women. It compromises womens and minorities importance and personal development within the workplace.

3.2 The Search for Objectivity

Some argue that unbiased AI is likely unachievable (Bozdag, 2013). Any product of humans is prone to echo their same partiality. Nevertheless, the mitigation of some detrimental biases is possible. Speech recognition in AI has been improved upon to respond to a wider array of dialects (Hirayama et al., 2015), increasing accessibility. Biases are arguably an inherent part of the algorithms on which AI is based (Danks and London, 2017). The solution may lie in bias mitigation at the source: humans.

[1]Word2Vec is a complex embedding created on a corpus of Google News with 3 million words(Bolukbasi et al., 2016).
[2]Word-embedding is the mathematical representation of conceptual associations within language

4 Transparency in AI Decision Making

Being still an emerging technology, truly autonomous AI is still far from being a reality. There are still many problems that must be solved before AI can be used safely and ethically.

4.1 Overview of Neural Network Based Systems

A neural network (NN) can simply be thought of as a function that takes inputs and produces some output. For example, a NN that classifies numbers might take pictures of handwritten digits as inputs.

Figure 2: Images of handwritten digits and their labels from the MNIST database (Lecun and Cortes, 1999).

The input layer would have one neuron per pixel in the image, with the activation being the pixel's brightness. The output layer would one neuron for each digit. In between these layers are what are called "hidden layers". Each neuron in a layer are connected to every neuron from the previous layer. Each of these connections will have some unique weight that represents how strongly correlated they are.

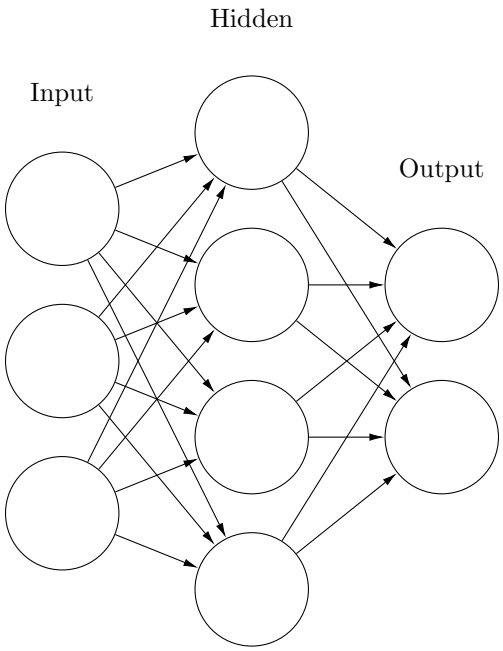

Figure 3: A visual representation of a small NN with 3 inputs, 2 outputs, and one hidden layer.

The activation of a given neuron is then calculated as follows:

$$a_{(i,j)} = \sigma\left(\left(\sum_{n=1}^{N} = a_{(n,j-1)} w_{(n,j-1)|(i,j)}\right) + b_{(i,j)}\right) \quad (1)$$

The weights are calculated for every neuron until they reach the output layer, where they hopefully give meaningful output. Here, a good output would be the neuron representing the correct digit having high activation, while the other neurons have low activation.

4.2 Using Training Data in the Real World

When a NN is created, the weights are unset and the network produces random noise for any given input. To make the network produce meaningful data it must be trained. Training is done by providing input, comparing its output to the expected output, and adjusting the weights accordingly[3].

One of the problems that arises from this is that the NN becomes optimized for the training data. Because of this, it is not guaranteed to perform well on a different set of data(Amodei et al., 2016). While this problem is not specific to AI[4], NNs are unable to recognize that they are untrained. This could create ethical issues when creating AI that must communicate to a user. By being trained on data from one cultural region, the AI may lack cultural awareness of others.

4.3 Current and Future Ethical Implications of Using AI in the Workplace

One of the applications of AI is sentiment analysis (SA). SA is used to identify someone's attitude towards some topic. Predictably, this technology has become highly sought after by businesses (Balahur et al., 2014). SA can be used to read the publics' reaction to something, or provide information for optimizing marketing campaigns. However, it can also be used to monitor an employee's communications. This could have serious ethical implications since it would be impossible for an employee to know what messages may get flagged. An employee could be fired without ever saying something against company policy.

As AI systems become more general, they will replace humans in more and more areas. It is imperative that these AI systems act in an ethical manner. One of the most important things to consider is how to define an AI's goals. For example, a general AI may be self-protective so that it may continue to achieve its goal (Omohundro, 2014), leading to unwanted behavior. Because of the opaque nature of AI, it will be hard to verify that a given system completely follows its intended purpose.

Equation (1) describes the summing of the activation of all of the previous neurons plus some bias, and then normalizing it using the sigmoid function.

[3]This is done with back-propagation, which essentially means adjusting the weights between the final layers, then adjusting the ones before that, until all the layers have been adjusted.

[4]i.e. an untrained human will also be poor at the same task

5 Guidelines for AI Usage

To address the issues with AI, guidelines must be created and observed by producers and users. Creating guidelines for AI is difficult, but crucial to address the downfalls of AI.

5.1 Creating Guidelines for The Future

Many organizations have already begun looking at the effects of AI on the future society. The One-Hundred-Year Study On AI from Stanford University outlines some aspects required to create effective guidelines: Place experts who understand AI interaction in government in order to property evaluate AI impact and recommend a path of action; Fund interdisciplinary studies to look at the social impact of AI; and, remove impediments to allow research on fairness and security of AI which is critical for examining accountability for AI systems (Stone et al., 2016). Creating guidelines will take time and a lot work is still needed, yet the first steps were already taken by the white house last year as two reports outlining a strategy for AI research and development (R&D) were published. The National Science and Technology Council (NSTC) machine learning and artificial intelligence (MLAI) community outlined over twenty recommendations, some major ones are presented in Table 2 below (National Science and Technology Council (U.S.) and Office of Science and Technology Policy and United States, 2016).

NSTC Number	Recommendation
1	Institutions should examine whether and how they can responsibly use AI and machine learning
2	Federal agencies should prioritize open data training and open data standards in AI
4	The NSTC MLAI subcommittee should develop a group for AI practitioners across government
11	The Government should monitor the milestones of AI development in other countries
13	The Federal government should prioritize short and long-term AI R&D
18	Schools should include ethics, and discussions about security, privacy, and safety, as part of a curricula on AI, and machine learning
20	The government should develop a strategy on international engagement, and a list of AI areas that need international engagement and monitoring
23	The government should finish developing of a single policy, consistent with humanitarian law, on autonomous and semi-autonomous weapons

Table 2: Key NSTC Recommendations

As a budding topic, AI is becoming more prominent and concerns are being raised on how it will change the society. A concern for many people is losing their jobs to automation. Moreover, the development of AI will cause inequality and greater bias in the labour market as low-skilled jobs disappear creating way for more high-skilled jobs (Leenes et al., 2017). Therefore, this issue, like many other, must be addressed in the future guidelines, but including all ethical issues is difficult.

5.2 The Challenge of Creating Guidelines

The major challenge of creating regulation for AI is keeping up with the technological advances. Since AI technology is new and evolving, there will be gaps in existing regulation and new laws to be made to adapt to the technology. However, since the technology is advancing rapidly more conflicts will arise with existing regulations. This dilemma for controlling evolving technology is called technology-neutrality versus legal certainty. One method is controlling the effects of AI using abstract regulations that can apply to many cases, yet this may not provide enough legal constraints and certainty. On the other hand, having strong legal certainty and premature laws may obstruct the scientific advancement and stop innovation. Legal regulations cant be adapted to include new advancing technology or use reclassification to include evolving technology in an existing distinction, new neutral and legal constrained need to be created (Leenes et al., 2017).

Regulating AI comes with the challenge of following ethical standards, and a strong value system of overarching principles. The issue of accountability must be addressed in the legal regulations, transparent research must be conducted to allow clear future guidelines for AI use, and biases must be considered in order to decide whether AI would strengthen existing bias or create new biases. The ethical principles should be key part in the creation of new guidelines but also impose the big challenge on the creation of proper guidelines.

6 Conclusion

It is undeniable that AI has potential. However, it has yet to be decided whether it has potential for good, or for harm.

It is in these formative stages, when the technology is quickly emerging, that the ethics of AI must be analyzed, discussed, and addressed. This can be done by looking closely at its enigmatic conception, its predispositions, its accountability, and the regulations that aim to govern them.

AI is often described as a “black box”. Its operation is based on inputs and outputs, and the link between the two remains cryptic to most. The formation of these NNs rely on inputting a set of data, analyzing the result and reconfiguring accordingly. This can lead to a network constrained to specific data or one that blunders unknowingly. The use of these networks in SA and marketing threatens privacy and freedom of speech. We even risk being outsmarted and outmaneuvered by our own technology.

Void of human sentiment and susceptibility to corruption, AI has the potential for objectivity. Yet the current biases in AI challenge the realism of this prospect. Biases in AI architectures may propagate social division through selective voice automated technology, partial filtering, and human based machine learning. In turn, this perpetuates racial prejudice, sexism and other harmful stereotypes. These effectively counteract social progress and cause imbalance in the workplace. However, improvement is possible by working case by case.

When complications in AI arise, the bestowal of blame can be complicated. AI accountability depends on its production and use. A programmer may be held responsible if an ethical framework is said to have been established. Yet if AI learns from experience, unethical consequences could be attributed to the AI, the entire body of individuals governing its behaviour, or the users themselves. To resolve this, a certifying agency ought to ascertain the safety of AI. Manufacturers would be cleared of liability if the AI is passable, otherwise the producer is liable.

Enlisting the help of experts, running studies, and establishing guidelines may mitigate the unfavorable ethical impacts of AI. Published reports recommend targeting government involvement, informing people about AI, and monitoring the international effect of AI. However, Establishing adequate regulation has proved difficult. Keeping pace with advancements without constraining growth is a fine balance. Moving forward, society should continue to innovate but must remain vigilant and recognize the ethical risks of AI.

References

Alaieri, F. and Vellino, A. (2016). Ethical decision making in robots: Autonomy, trust and responsibility. In *International Conference on Social Robotics*, pages 159–168. Springer.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565*.

Balahur, A., Mihalcea, R., and Montoyo, A. (2014). Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications. *Computer Speech & Language*, 28(1):1–6.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and information technology*, 15(3):209–227.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Danks, D. and London, A. (2017). Algorithmic bias in autonomous systems. *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017)*, pages 4691–4697.

Hirayama, N., Yoshino, K., Itoyama, K., Mori, S., and Okuno, H. G. (2015). Automatic speech recognition for mixed dialect utterances by mixing dialect language models. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(2):373–382.

Lecun, Y. and Cortes, C. (1999). The MNIST database of handwritten digits. Retrieved from `http://yann.lecun.com/exdb/mnist/`.

Leenes, R., Palmerini, E., Koops, B.-J., Bertolini, A., Salvini, P., and Lucivero, F. (2017). Regulatory challenges of robotics: some guidelines for addressing legal and ethical issues. *Law, Innovation and Technology*, 9(1):1–44.

Liu, H.-Y. (2017). Irresponsibilities, inequalities and injustice for autonomous vehicles. *Ethics and Information Technology*, 19(3):193–207.

National Science and Technology Council (U.S.) and Office of Science and Technology Policy and United States (2016). Preparing for the future of artificial intelligence. Technical report, Executive Office of the President.

Omohundro, S. (2014). Autonomous technology and the greater human good. *Journal of Experimental & Theoretical Artificial Intelligence*, 26(3):303–315.

Scherer, M. U. (2016). Regulating artificial intelligence systems: risks, challenges, competencies, and strategies. *Harvard Journal of Law & Technology*, 29(2):353.

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S., et al. (2016). Artificial intelligence and life in 2030. *One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel*.