Selected Papers of #AoIR2019:
The 20th Annual Conference of the
Association of Internet Researchers
Brisbane, Australia / 2-5 October 2019

ASSESSING ETHICAL AI-BASED DECISION-MAKING: TOWARDS AN APPLIED ANALYTICAL FRAMEWORK

Paul Henman
University of Queensland

Introduction

Globally, there is strong enthusiasm for using Artificial Intelligence (AI) in government decision making. In activities as diverse as assessing individuals' national security risk, allocating social housing services, deciding parole applications, and identifying tax fraud, AI is being promoted and adopted as an efficient, effective, unbiased and error-free approach to government decision making. Simultaneously, people are expressing words of caution that this algorithmic approach is not without significant downsides including bias, exacerbating discrimination and inequalities, and reducing government accountability and transparency (e.g. Eubanks 2017; Pasquale 2015).

Consequently, there has been a flurry of work striving to identify challenges, principles, policies, regulations and institutions for enacting ethical AI within government, and society more broadly (e.g. UK, 2018; Australia, 2018; Council of Europe, 2018; Campolo et al 2017). These ethical AI frameworks operate primarily at the abstract level of issues and dimensions.

An elusive dimension in these discussions and proposals is applied or practical mechanisms and methodologies by which specific AIs can be assessed as un/ethical. We therefore remain poorly equipped to assess the ethics of an particular AI based on its design and operation. One approach suggested by Australia's Department of Industry, Innovation and Science (2019) provides a "toolkit for ethical AI" that includes impact assessments, risk assessments, review, best practice guidelines, industry standards, collaboration, monitoring, improvement and recourse mechanisms, and consultation. The UK government recently published *A guide to using artificial intelligence in the public sector* (2019), which focuses on assessing, planning and managing AI, and using AI ethically and safely.

Suggested Citation (APA): Henman, P. (2019, October 2-5). *Assessing ethical AI-based decision-making*. Paper presented at AoIR 2019: The 20th Annual Conference of the Association of Internet Researchers. Brisbane, Australia: AoIR. Retrieved from http://spir.aoir.org.

The purpose of this paper is to complement and extend these works, by proposing an more practical analytical framework that gets closer to the actual characteristics of an AI, and which can help guide the design, building and assessment of AI-based decision making. The analytical tool is constituted by a series of key questions that in turn can highlight areas for caution or concern.

An applied analytical framework for assessing ethical AI

AI based decision making can be summarized as:

> AI decision making = (data * context) + code + (use * context)

An AI algorithm requires **code**, a set of instructions, which operates on input **data** to produce output data (and actions). Such input data and the writing of the code are outcomes of the **context** in which they originate. The effects of the algorithm are intimately tied to how the algorithmic output is **used** and its wider context.

Data

AI-based decisions are determined on the basis of data. Algorithms define the data structures on which they operate. Bias is often listed as an ethical challenge of AI decision making (e.g. Noble 2018; Campolo et al 2017). The nature of data used to design an AI, as input data for learning, and to provide decisions, is a source for bias. What is known or not known, and the structure of that knowledge or data ontology matters (Davis 2017, Iliadis, 2018; Züllighoven & Keil-Slawik 1992). As the nature of data and data categories provide the foundation for AI bias, identifying its data and categories are therefore central for assessing the ethicality of AI.

> *Does the AI incorporate social categories associated with disadvantage (e.g. sex, ethnicity/race, religion, sexuality)?*
>
> *Are proxies for these used (e.g. names, place of birth, address, nationality)?*

Data context

The context of data generation and data for learning deeply shapes the quality and potential for bias in data. For example, critics of using AI in the criminal justice system argue that because criminal justice statistics arise out of a racially biased system, an AI trained on that data is likely to make decisions that reproduce that racial bias (c.f. Noble 2018).

> *Could the training set data be 'biased' (e.g. not representative, of poor quality, reflecting structural inequalities)?*

Code

Algorithms execute operations based on a program or code. The code defines how input data is operated on in order to produce output data and/or actions. Understanding how input data is used to determine outcomes is thus central for assessing the ethicality of AI. Importantly, just because social categorical data is available, it does not necessarily follow that those categories have any effect on outcomes.

Does the AI differentiate along social categories associated with disadvantage (e.g. sex, ethnicity/race, religion, sexuality)?

If so, does it make scientific and/or ethical sense to treat individuals differently based on cohort characteristics?

Is the algorithm based on proxy data to make decisions (e.g. assuming neglect = abuse; student performance = teacher performance)?

In traditional algorithms the relationships between input and output variables is defined by programmers and can be traced, whereas in learning algorithms these relationships are 'learnt' by trial and error in matching input data with a learning set of output data. Consequently, in evaluating AI alternative approaches are needed, such as testing and reverse engineering (Watcher et al 2018).

Use

AI-based decision-making about individuals, necessarily involves differentiating between individuals. A key ethical overarching consideration is whether these decisions could increase (or ameliorate) social structures and inequalities, or increase (or reduce) overall societal harms (such as bombing in an urban war environment).

Does the AI decision involve (a redistribution of) an increase of 'harms' or disadvantages (particularly for disadvantaged groups)?

Use context

Considerations about the use context of AI based decision making include whether or not humans can intervene to stop an AI decision, reverse its impact (unlike Diallo (2018) who was fired by an algorithm), appeal an AI decision, and seek redress.

Is human intervention possible in halting, reversing and correcting an AI-based decision?

Do subjects have the right to an explanation for an AI based decision?

Can subjects affected by an AI based decision effectively appeal and overturn that decision and seek appropriate readdress?

Conclusion

Building ethical AI is a widely agreed objective, but there is little understanding of how this might be achieved. Developing applied analytical frameworks that can evaluate the design, development and operation of specific AI decision-making tools is an important and urgent task in achieving this objective. This paper analytical approach and series of questions can in turn be deployed as Lickert scale questions to score an AI's ethics.

References

Australia. Department of Industry, Innovation and Science (2019) *Artificial Intelligence: Australia's Ethics Framework – A discussion paper*, Canberra: DIIS, https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf

Australia. Human Rights Commissioner (2018) *New Technology and Human Rights,* Sydney: AHRC, tech.humanrights.gov.au/sites/default/files/2018-07/Human%20Rights%20and%20Technology%20Issues%20Paper%20FINAL.pdf.

Council of Europe. Parliamentary Assembly (2018) *Algorithms and Human Rights,* Stasbourg: CoE, rm.coe.int/algorithms-and-human-rights-study-on-the-human-rights-dimension-of-aut/1680796d10.

Campolo, A., Sanfilippo, M., Whittaker, M., & Crawford, K. (2017). *AI Now 2017 report*. New York University: AI Now Institute.

Davis, S. L. (2017). 'The uncounted: politics of data and visibility in global health'. *The International Journal of Human Rights*, 21(8), 1144-1163.

Diallo, I. (2018) 'The machine fired me: No human could do a thing about it!' idiallo.com/blog/when-a-machine-fired-me, accessed 25/02/19.

Eubanks, V. (2018). *Automating inequality*. New York: St. Martin's Press.

Iliadis, A. (2018). 'Algorithms, ontology, and social progress'. *Global Media Communication*, DOI: 1742766518776688.

Pasquale, F. (2015). *The black box society*. Cambridge, MA: Harvard University Press.

Noble, S. U. (2018). *Algorithms of oppression*. New York: NYU Press.

UK. House of Lords. Select Committee on Artificial Intelligence. (2018). *AI in the UK: ready, willing and able?,* London: publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

UK. (2019) 'A guide to using artificial intelligence in the public sector', https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector

Wachter, S., Mittelstadt, B. D. M., & Russell, C. (2018). 'Counterfactual explanations without opening the black box: automated decisions and the GDPR'. *Harvard Journal of Law and Technology*, 31(2).

Züllighoven, C. F. H., & Keil-Slawik, R. B. R. (1992). *Software Development and Reality Construction*. Berlin: Springer