Dynamic Algorithmic Impact Assessment to Promote an Ethical Use of AI in Businesses

Shefeh Prisilia Mbuy

School of Computing and Mathematics
Keele University, Staffordshire, UK
s.p.mbuy@keele.ac.uk

Abstract

My PhD research focus is to produce a critical review of literature in Algorithmic Impact Assessment (AIA) and to develop an AIA tool that can be used to evaluate potential unintended impact of AI systems.

Introduction

The use of Artificial Intelligence (AI) is bringing about transformation to the way we live, work and interact with society in ways that were previously not envisaged.

While businesses are reaping the benefits of using AI, there is increasing concern about the ethical implications and unintended consequences of their specific implementations. In particular, when AI is used in automating high stakes decision making which affect people's lives such as using facial recognition algorithms in crime investigations. Automating such high stakes decisions raises concerns about fairness, bias, privacy, and transparency (Hagendorff 2020).

With policy and regulation lagging behind while the adoption of AI continues to accelerate, the gap to address the lack of ethical design in systems will continue to widen. This gap is at the core of what my PhD research is looking to address.

Research Aims

The aim of my PhD research is to produce a critical review of literature in Algorithmic Impact Assessment (AIA) and to develop an AIA tool that can be used to assess the impact of an AI system. The target of the tool will be Small and Medium Enterprises (SME) who build AI systems. Currently, system developers who want to incorporate ethics into the designs and implementations of AI systems lack the tools to translate abstract principles into practice (Morley et al. 2020; Vakkuri et al. 2019).

The dilemma for small businesses and start-ups in the private sector is that so many are facing the challenges of how to approach this in a structured and cost effective way due to the lack of established tools and methodologies. Therefore, the AIA tool will help SMEs to design and monitor AI systems in a way which mitigates unintended consequences.

Although the target is SMEs, the larger stakeholder group includes different classes of stakeholders all with conflicting requirements and goals: the SME (goal: profit), the public (goal: getting a service, maintaining privacy), the government (goal: producing and enforcing legislation).

Key Research Questions

Some key research questions include defining what impact is, how impact can be measured, what constitutes an impact assessment in the context of an AI system, defining what it means to adopt a dynamic approach to AIA and also the question on the effectiveness of AIAs for SMEs.

Defining what counts as impact is fundamental to my PhD research. A literature review in this area has highlighted that defining impact in algorithmic systems is complex. In their paper, Metcalf et al. point out that impact should be defined in a way which reflects real world harms. The authors argue that there is a risk that if impact is not properly defined, AIAs could be designed to measure evaluative metrics which do not reflect actual harms. The authors recommend adopting "the co-construction of impacts" as an approach to overcome this mismatch.

Measuring impact is also a key consideration for AIA. Several AIAs to date have adopted a quantitative approach using defined metrics. However not all harms can be mapped to a metric and therefore the challenge to consider is how to incorporate a quantitative approach to the AIA process.

The question of when to perform an AIA is also a key one. Typically, assessments would take place in a defined period. Valid arguments have been made on both sides of the ex-ante and ex-post debate. An ex-ante approach allows for potential unintended impact to be identified during system development and therefore creates the opportunity to take steps to address these prior to system deployment. On the other hand, the ex-post approach allows for real impact to be identified once the system is deployed in the environment of its intended use. It is argued that at this point, an impact assessment will reflect what the true impact is. While these are sound arguments, our view is that AIAs need to adopt a dynamic approach throughout the lifecycle of an AI system i.e. from design through to post deployment. This approach ensures that unintended impact resulting from future system maintenance and enhancements including changes to use case can be identified and addressed.

On the question of how effective an AIA would be for SMEs, most AIAs have so far focused on the public sector

in attempt to bring accountability to the use of AI in public service decision making. Although SMEs do not have to comply with certain types of regulations, some may choose to adopt a responsible approach to implementing AI. The question here is, are AIAs that are designed for public sector accountability suitable for SMEs? How can these toolsets be designed in a way which will address SME challenges such as pressures on time, funds, resources, and expertise?

These key questions will inform the approach that will be adopted for the AIA tool for my PhD research.

Progress to Date

For my PhD research, I am working as part of the Keele University Explainable and Responsible AI (XRAI) research group whose interests are to promote an ethical use of AI. We have completed the initial preliminary work on identifying the key factors to be considered when evaluating an automated decision-making system. These are based on accountability and responsibility, fairness, transparency and explainability.

- For the accountability and responsibility aspect we will consider the need to provide traceable explanation for the system's decision, the representation of societal norms as well as ascribe decision to specific parts of the algorithm.
- From a fairness perspective, the considerations we will be looking beyond traditional accuracy metrics, considering bias in data (for example skewed samples, tainted samples, limited features and sample size disparity). The fairness aspect will also include bias in algorithm and formal definitions such as demographic parity, counterfactual, etc.
- The transparency and explainability aspect will consider the "black box" effect. This will include the agnostic (algorithmic-independent) explanations, feature-based (i.e. trace impact of features on outcome), the algorithmic method i.e. the use of intrinsically explainable methods (such as Bayesian, tree-based, rule-based) as well as counterfactual advice that can help in understanding what change is needed to achieve a desired outcome in the future.

Notable Related Work

Over the past few years, support for the use of AIAs as an approach for identifying and mitigating unintended impact from algorithmic systems has gathered momentum. This has led to a growing area of research into how impact assessments which have been used in other domains for accountability purposes can be adapted for algorithmic systems.

To this end, Watkins et al. (2021) highlight the many challenges within AIAs related to the lack of universal approach and "ambiguity" on how AIAs are used to mitigate potential harmful impact. The authors offer six observations on AIAs as instruments for constructing accountability.

Metcalf et al. (2021) present the argument that, for AIAs to be effective, the process must map impacts to potential harms. The authors argue that current AIA proposals do not reveal potential harm but rather *proxies* for harm, and propose the "co-construction of impacts" as an approach to address this challenge.

Raji et al. (2020) propose an internal algorithmic audit framework (SMACTR) which has five distinct stages: Scoping, Mapping, Artifact Collection, Testing and Reflection. The authors argue that internal audits have a part to play in mitigating potential harmful impact from algorithmic systems. The point is made that unlike external audits which take place after a system has been deployed and which are "limited by lack of access" to the inner workings of these systems, internal audits can help to identify risks that might not have been identified as part of an external audit.

All these and many more proposed forms of AIAs have contributed to the research on how AIAs can be adopted as a form of accountability for algorithmic systems. My research will build on these with specific focus on areas such as the dynamic aspect of AIAs and the application of AIAs in the private sector.

Brief Timeline

I am currently writing a critical comprehensive review of AIA and I anticipate to have completed this work by February 2022. Subsequent phases of my research will include identifying requirements for the private sector and then a use case to conduct analyses. The findings and results will be published as a journal or conference paper (hopefully in a future AAAI conference).

References

Hagendorff, T. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30(1): 99–120.

Metcalf, J.; Moss, E.; Watkins, E. A.; Singh, R.; and Elish, M. C. 2021. Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 735–746.

Morley, J.; Floridi, L.; Kinsey, L.; and Elhalal, A. 2020. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and engineering ethics*, 26(4): 2141–2168.

Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 33–44.

Vakkuri, V.; Kemell, K.-K.; Kultanen, J.; Siponen, M.; and Abrahamsson, P. 2019. Ethically aligned design of autonomous systems: Industry viewpoint and an empirical study. *arXiv preprint arXiv:1906.07946*.

Watkins, E. A.; Moss, E.; Metcalf, J.; Singh, R.; and Elish, M. C. 2021. Governing Algorithmic Systems with Impact Assessments: Six Observations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21.