Schools of AI in the Public Sector: Fairness and Accountability Concerns

Marc T.J Elliott
School of Electronics, Electrical Engineering and Computer Science
Queen's University Belfast
Belfast, United Kingdom
melliott22@qub.ac.uk

Abstract

As decision-making algorithms become more prevalent in society, the importance efficiency and problem-solving abilities come into question when predictions impact individuals' lives. High-risk applications require trusted AI systems to be designed with fairness and accountability; such trust and consideration is essential for public acceptance and successful deployment. Despite growing advocacy for ethical and trustworthy AI, along with the emergence of guidelines like the EU AI Act, controversies surrounding AI persist in the media. Public sector AI systems are being implemented haphazardly, whether in judicial decision-making, healthcare diagnostics, or social welfare distribution. These high-risk applications directly affect citizens' quality of life, highlighting the need for a critical assessment of how AI are being designed and deployed in the public sector. My thesis explores the integration of fairness, accountability, and uncertainty in public sector AI to assess whether these systems are appropriately designed, effectively adapted, and capable of enhancing societal well-being. The research aims to provide actionable insights for designing AI systems that align with public sector needs and maximize societal benefits.

Motivation

As decision-making algorithms become more prevalent in society, the importance efficiency and problem-solving abilities come into question when predictions impact individuals' lives. High-risk applications require trusted AI systems to be designed with fairness and accountability; such trust and consideration is essential for public acceptance and successful deployment.

Despite growing advocacy for ethical and trustworthy AI (Kaur et al. 2022), along with the emergence of guidelines like the EU AI Act, controversies surrounding AI persist in the media. Public sector AI systems are being implemented haphazardly, whether in judicial decision-making, healthcare diagnostics, or social welfare distribution. These high-risk applications directly affect citizens' quality of life, highlighting the need for a critical assessment of how AI are being designed and deployed in the public sector.

My thesis explores the integration of fairness, accountability, and uncertainty in public sector AI to assess whether these systems are appropriately designed, effectively adapted, and capable of enhancing societal well-being. The research aims to provide actionable insights for designing AI systems that align with public sector needs and maximize societal benefits.

Accomplished Work

To this end, two chapters on assessing fairness and accountability of AI practices within the public sector have been completed, with additional research underway. My first publication, presented at the 2023 AI Safety Engineering workshop, explored leveraging group-level learning to produce implicitly fairer decisions (Elliott and P 2023). The second work advances our understanding of the relationship between AI schools and their capacities for meaningful accountability. This led to a subsequent publication in the ACM Digital Government journal (Elliott, P, and Maccarthaigh 2024). The following expands on the importance of the research accomplished in each of these areas:

Fairness with Group-Level Learning

Prior research on fair AI has focused on embedding mathematical notions of fairness into training objectives, balancing performance with fairness metrics via loss functions (Agarwal et al. 2018). While effective, these approaches often require access to sensitive attributes during prediction, or a detailed understanding of the task context to select appropriate fairness notions. These requirements complicate real-world deployment due to privacy laws restricting sensitive attribute access and the complexity of justifying fairness metrics, which often conflict when multiple notions are applied (Kleinberg, Mullainathan, and Raghavan 2016).

Our research diverged from this standard approach by developing a novel, implicit in-processing method to achieve fairer outcomes through group-level training (Elliott and P 2023). By adapting a logistic regressions training phase, we train independently on data from different sensitive groups to produce coefficient weights that align with each group. These weights are then aggregated after each training iteration to update the model parameters, ensuring a more equal representation of each group. This technique addresses real-world group imbalances, providing underrepresented groups with more equitable influence on model parameters, thereby

reducing unfairness without requiring additional data handling techniques.

Following this, subsequent experiments explored different aggregation methods, revealing that even with implicit fairness, the choice of aggregation method can influence the underlying fairness agenda, directing the nature of the decisions made.

AI Schools and Public Sector Accountability

My second work focuses on the relationship between different AI techniques and the types of accountability within the public sector. While the literature addresses AI accountability, it has significantly overlooked the multiple forms of public accountability and the challenges they present [2, 10]. To address this gap, three interconnected studies exploring how different AI paradigms affect public sector accountability were conducted.

Domingos (2015) identified five schools of AI, each with unique theoretical and technical concepts defining their approach to knowledge formation and the categorisation of AI algorithms. Our first study analyses the two dominant AI schools in the public sector against the multiple forms of accountability. Findings indicate that a single AI paradigm may not seamlessly align with every phase of meaningful accountability across different forums. This study highlights the need to assess the suitability of technological environments to meet specific accountability requirements before public deployment, thereby preventing hindrances in the accountability process. It sheds light on an overlooked aspect of AI accountability literature, emphasising how algorithmic design influences accountability relationships even before implementation. Marking an initial effort to align technical AI schools with specific accountability types, encouraging further discussions on the impact of evolving technologies on multiple accountability forums and designing systems for greater public benefit.

Our second study, presented at the 2024 ASPA Conference, further explores the relationship between AI schools and their capacity to support accountability. Using a database of real-world AI incidents and controversies (Pownall 2021), we identified trends in accountability issues through the lens of Peters (2018) analysis on policy coordination challenges. The study reveals correlations between the type of AI school deployed and the most likely accountability problems, providing insights into how different AI paradigms influence accountability outcomes.

Lastly, our third study develops a framework to address changing accountability dynamics introduced by generative AI (Elliott, P, and Maccarthaigh 2024). It examines how these systems transform government-citizen relationships, introducing dual-phase relationships involving the actor, forum, and AI. We provide actionable recommendations for public servants and AI designers to ensure accountability while maximising the benefits of advanced AI technologies.

Future Work

Building on completed work, I plan to explore another critical characteristic of public sector AI usage: uncertainty. Drawing on Gigerenzer (2024) 'Adapt-to-AI Principle', which suggests that adapting physical environments or human behaviors can improve AI performance in uncertain situations (Gigerenzer 2022). This research will contextualise the principle within public sector systems, an area inherently filled with uncertainty and unforeseeable factors that affect AI performance. This raises several key questions: How must societal norms evolve for the public to adapt to AI? When are such changes necessary, and when should task automation be avoided to preserve what makes society inherently human?

We will analyse case studies of public sector AI with varying risks and complexities to determine the environmental adaptations needed to maximise the AI potential. With the goal to develop a framework for identifying and implementing these adaptations, thus pushing the boundaries of AI uncertainty research for public sector applications. This work acknowledges the potential advantages of integrating psychological insights with AI systems, enhancing our understanding of managing uncertainties, and offering valuable guidance for public practitioners.

Our goal is to assist practitioners in adapting and designing environments that maximise AI potential without compromising citizen safety. Additionally, we aim to provide insights into how uncertainty affects various AI algorithms and how these can be better designed to handle varying degrees of uncertainty in unstable environments.

Conclusion

At the core of my thesis is the question: How compatible are current public sector AI systems with social normative standards and the goal of social good? Our current results indicate that while AI systems are significantly transforming society, they require more careful deliberation and consideration in the design phase before deployment into real-world applications.

Our fairness studies reveal that seemingly minor decisions on weight aggregations can lead to vastly different interpretations of fairness and subsequent algorithmic outputs. Meanwhile, our accountability studies show that the choice of algorithm affects the meaningfulness of accountability and determines who benefits most from AI systems. Both studies highlight a common theme: the choice and design of the algorithms driving these systems is often an overlooked factor in the public sector. By addressing this oversight, our research offers insights that could shape public policy, ensuring AI development aligns with the societal goals and enhances government-citizen relationships.

Our proposed future study on uncertainty will examine whether the public environment itself can be feasibly adapted to better integrate AI systems from the ground up, rather than attaching these systems to pre-existing processes haphazardly. By identifying the necessary adaptations to integrate AI systems more effectively, we aim to ensure these technologies truly benefit society and enhance the quality of public sector decision-making.

References

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. In *International conference on machine learning*, 60–69. PMLR.

Domingos, P. 2015. *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Basic Books.

Elliott, M.; and P, D. 2023. A Group-Level Learning Approach Using Logistic Regression for Fairer Decisions. In *International Conference on Computer Safety, Reliability, and Security*, 301–313. Springer.

Elliott, M. T.; P, D.; and Maccarthaigh, M. 2024. Evolving Generative AI: Entangling the Accountability Relationship. *Digital Government: Research and Practice*.

Gigerenzer, G. 2022. *How to stay smart in a smart world: Why human intelligence still beats algorithms*. MIT Press.

Gigerenzer, G. 2024. Psychological AI: Designing algorithms informed by human psychology. *Perspectives on Psychological Science*, 19(5): 839–848.

Kaur, D.; Uslu, S.; Rittichier, K. J.; and Durresi, A. 2022. Trustworthy artificial intelligence: a review. *ACM computing surveys (CSUR)*, 55(2): 1–38.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Peters, B. G. 2018. The challenge of policy coordination. *Policy design and practice*, 1(1): 1–11.

Pownall, C. 2021. AI, Algorithmic and Automation Incident and Controversy Repository (AIAAIC).