Unite the study of AI in government—with a shared language and typology

1. Main text

With the arrival of generative Artificial Intelligence (AI), we have arguably entered a brave new world of opportunities and risks in technology adoption—and public institutions cannot afford to misstep. Although the emerging study of AI in government has advanced insights into important topics such as algorithmic aversion and bias, we believe several weaknesses undermine the field's validity and policy usefulness. As a result, it is high time scholars and policy advocates across social and technical fields unite so that we realize the tremendous potential AI holds to build resilience in government—and wider society. Here, we offer some thoughts on the issues in current research practice as we see them, and provide a few suggestions for how social scientists and AI scholars might overcome them by adopting a shared language and typology.

By now AI strategies and applications have been adopted by various states and public agencies across the world in the hope of making government services more effective, fair and responsive. In the context of the UK, for instance, a recent a survey of public sector professionals carried out by Bright and colleagues (2024) found that 22% of respondents are actively using a generative AI system in their work. Meanwhile, Egypt's recent Charter for Responsible AI sets out how the country seeks to responsibly develop and use AI systems. Yet, as the state of play continues to evolve, the integration of machine intelligence into state bureaucracy risks becoming riddled with ever more "AI tensions": sociotechnical challenges relating to the fairness, transparency, and explainability of AI systems, to name a few.

Whilst the adoption of AI in myriad contexts will likely continue to proliferate, we now know that for citizens to trust the government adoption of AI systems, these must overcome such tensions by satisfying a range of criteria—big on the list are accuracy, safety and interpretability. Consequently, calls for new metrics, technical standards and governance mechanisms to guide the use of trustworthy AI have now become commonplace. With the development of the AI Act, the EU has started to make the most progress on this by introducing the most ambitious attempt to regulate AI in a way that accounts for the various risks that AI systems pose. Yet, like most other research and policy efforts, the act falls short in accounting for all issues potentially relevant for the use of AI and likely underestimates the risks posed by large language models (LLMs). To overcome this, we need to arguably rethink how we structure research on institutional AI adoption to begin with.

How can researchers begin doing that? Currently, a major issue is that most research efforts still tend to prioritize only a handful of concepts; they do not fully connect all the different perspectives relevant to understand the implications of real-world use cases. At times this can even involve focusing on hypothetical future scenarios at the expense of highlighting the societal harms AI is causing right now. This omission stems in part from what may be called the 'relational problem' in socio-technical discourse: fundamental terminological issues have not yet been settled—including semantic ambiguity, a lack of clear relations between terms and differing standard glossaries. In short, the concept of 'fairness', for instance, can mean very different things depending on who you ask across the social sciences and the fast-moving fields of AI, ML, and robotics. Taken together, this contributes to the prevalence

of conceptual isolation in the fields that study government AI and the development of disparate metrics, standards and governance mechanisms.

While there will always be differences in how countries govern technology, the widespread risks and benefits of generative AI, especially LLMs, mean it is critical that these are embedded using standard operational procedures, clear epistemic criteria, and behave in alignment with the normative expectations of society. It is high time research on public sector AI unite to address these challenges. As starting points, we wish to promote the call made in Straub et al (2023a) to adopt a shared glossary and unified framework for describing AI across social and technical fields and linking new metrics to existing legal frameworks.

At the heart of our suggestion is the call for the development of new, multidimensional concepts that better capture the complexities of government AI systems, encourage cross-disciplinary dialogue, and foster a move to legal standardization. New examples discussed in a recent integrative review by the Alan Turing Institute include: operational fitness, epistemic alignment and normative divergence (see Straub et al., 2023b for definitions). Each of these seeks to possess depth, be bounded, that is, operationalizable, and offer theoretical utility beyond the usual suspects of single-construct concepts like transparency and accuracy. In other words, given the essentially contested nature of many existing terms used in current scholarship, we argue that the field needs to work collectively towards bridging past efforts to conceptualize AI in government by developing a shared language that is connected to emerging technical standards proposed by the likes of the EU.

Given the diverse conceptual strands of current scholarship, another theoretical innovation that can help unite the field of AI in government is the development of a novel framework to analyse and classify AI systems across disciplines. Specifically, we believe that AI in government is lacking an accepted typology, understood both as a method to classify observations in terms of their attributes and the development of theories about configurations of variables that constitute conceptual types. Whilst typologies take time to be developed, they are a tried and tested analytical strategy that can be “put to work” in forming concepts, refining measurement, exploring dimensionality, and organizing explanatory claims—all factors that can help unite the study of AI in government. Researchers have already proposed what such a typology may look like but a unified framework will ultimately require input from multiple disciplines (see references).

Overall, the advent of generative AI marks a pivotal juncture demanding concerted efforts to align technological advancements with societal values. Effectively addressing challenges in the deployment of AI in government clearly requires interdisciplinary collaboration. By advocating for a shared glossary, unified framework, and the development of multidimensional concepts, we believe it is possible to foster a comprehensive understanding and responsible integration of AI in the public sector. We submit these suggestions with the hope that they will engender interdisciplinary dialogue, facilitated by interdisciplinary forums like *AI & SOCIETY*, to drive advancements in the trustworthy use of trustworthy AI.

2. References

Bright J, Enock F E, Esnaashari S, Francis J, Hashem Y, & Morgan D (2024). Generative AI is already widespread in the public sector. https://doi.org/10.48550/arXiv.2401.01291.

Straub V J, Morgan D, Hashem Y, Francis J, Esnaashari S, & Bright J (2023a). A multidomain relational framework to guide institutional AI research and adoption. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*. Association for Computing Machinery, New York, NY, USA, 512–519. https://doi.org/10.1145/3600211.3604718.

Straub V J, Morgan D, Bright J, & Margetts H (2023b). Artificial intelligence in government: Concepts, standards, and a unified framework. *Government Information Quarterly*, *40*(4). https://doi.org/10.1016/j.giq.2023.101881.

3. Data availability statement

No data is associated with this manuscript.