A pluralist approach to human-centred government AI with a focus on gender-based equity

International School for Government, King’s College London

Amanda Dahl

April 2024

Introduction

The rise of artificial intelligence (AI) technology has led to both highly impactful advancements and also significant challenges across civil society. As governments increasingly rely on AI systems to automate decision-making processes, it becomes imperative to critically examine their impact on social equity, particularly in relation to gender-based issues. This essay seeks to explore the complex intersection of AI technology and gender equity when deployed by the UK government for decision-making, with a focus on the design of a policy intervention informed by a pluralist approach to human-centred AI.

The purpose of this essay is twofold: first, to apply complexity theory to the dynamics of AI systems within the context of government decision-making, and second, to propose a policy intervention that leverages concepts and tools acquired from the module. By adopting a pluralist perspective, which recognises diverse voices and perspectives, this intervention aims to foster a government AI ecosystem that prioritises the needs and rights of all genders. Through this exploration, I would like to contribute to the ongoing conversation on the ethical and social implications of AI technology while advocating for tangible solutions grounded in the theory and practice of complexity theory, systems thinking and design thinking.

Public issue

When adopted by governments for automated decision-making, AI has the potential to improve outcomes for women and those of marginalised gender identification, especially those affected by economic inequality. Equally, there is risk of further disadvantage to these populations should the automated decision-making be poorly implemented. The intersection of government AI and gender-based inequality is a public issue because it touches upon fundamental principles of democracy, human rights, social cohesion, economic development, public health, and ethics. Addressing this issue requires proactive efforts to design and implement AI systems that promote gender equality, especially as "women are often underrepresented in available data sets" used to train AI (Browne, Drage, et al, 2024).

For instance, there are gendered impacts of automating social safety net benefits, which can overlook and undervalue care as a legitimate economic practice (Barford, 2022). When technology aimed at benefits eligibility is introduced, "women's work becomes degraded and polarised from men's work" (Frennert, 2021). In this case, if historical data used to train the AI shows that women are more likely to be the primary caregivers in households and may have intermittent employment due to caregiving responsibilities, the algorithm may unintentionally disadvantage women when assessing eligibility for benefits. As was seen in a social safety net benefits system in Austria, the AI system might not adequately account for the economic impact of caregiving responsibilities on women's employment status, overestimating the employability of women with caring responsibilities, and leading to fewer women receiving the support they need (Human Rights Watch, 2021).

Complexity theory and government service provision

Provision of government services to citizens constitutes a “complex adaptive system” characterised by multiple interconnected components. It includes government departments, policymakers, service providers, the public, and the broader socio-economic and political context, all interacting to shape the delivery of services. Adaptation and learning are inherent to this system (Sterman, J. 2001), as departments and policymakers continuously adjust their strategies and practices in response to evolving needs, ministerial priorities, feedback from the press/social media, and changes in external factors such as technology and demographics. These interactions often show nonlinear dynamics, where small changes in one aspect of the system can lead to significant, sometimes unforeseen effects elsewhere (Sterman, J. 2001), illustrating the complexity and interconnectedness of government service provision.

The system also displays self-organising tendencies, with patterns and structures emerging from the interactions between various actors without centralised control. Feedback loops underscore the dynamic nature of service delivery processes (Sterman, J. 2001). Recognising government service provision as a complex adaptive system helps policymakers and practitioners to create more resilient, responsive, and human-centric approaches that effectively address the diverse needs of populations.

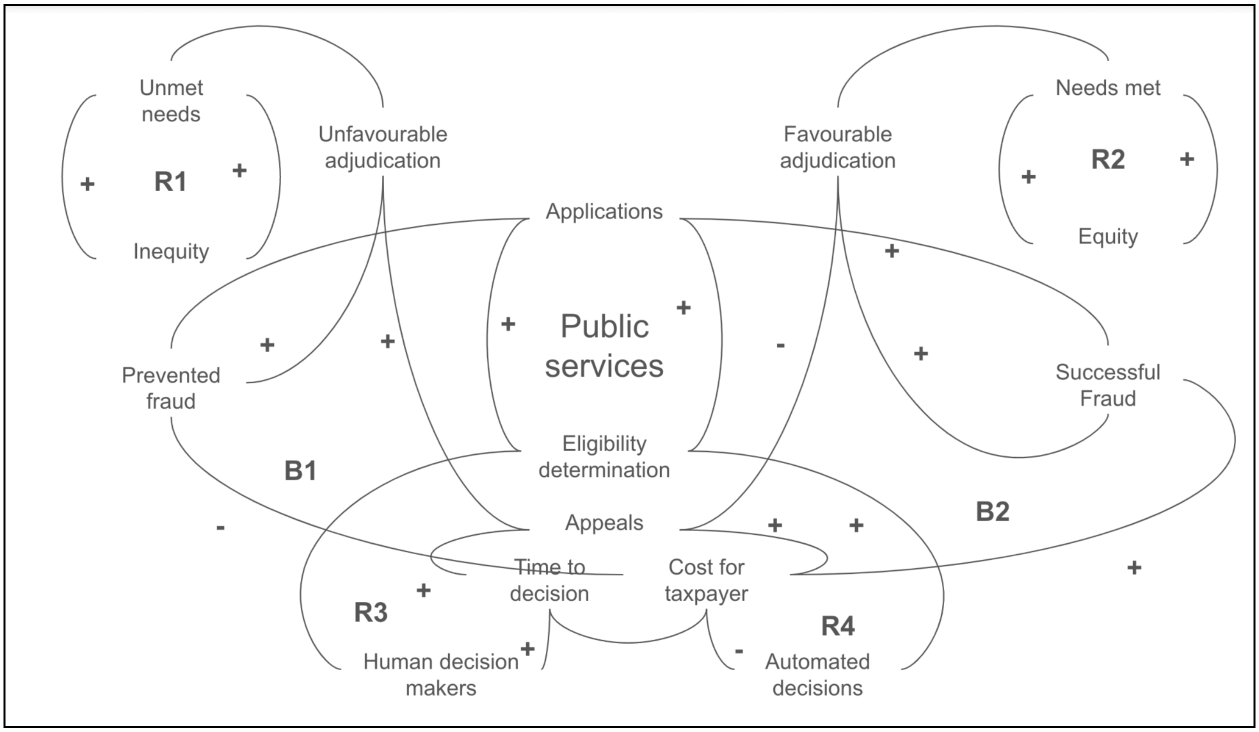


Figure 1: Causal loop diagram showing the reinforcing and balancing loops involved in public services attempting to meet the needs of citizens whilst avoiding fraud.

Pluralism and gender representation in AI development

Pluralism emphasises the inclusion of multiple voices and interests in decision-making processes. Connolly argues that “in a culture of multidimensional pluralism, support for the reduction of inequality requires the mobilisation of a majority assemblage” (Connolly, 2005). Yet women, and particularly those from marginalised communities, are often underrepresented in the development and implementation of AI technologies. In fact, “the number of male researchers working in STEM is much larger than that of women, and this is a constant in most countries all over the world” (Vallis, Gilbert, 2022) .

This limited representation perpetuates gender biases and reinforces existing inequalities, as AI systems may fail to account for the diverse needs and realities of women. “Without incorporating diverse expertise, deliberative results risk being stupid. Without being responsive to what matters, deliberative results are irrelevant or solve the wrong problems. And without commitments to action, deliberative results risk being just tal.” (Forester, 2018).

UK government policy on AI regulation is not without recognition that “unaccounted for bias in an AI-enabled automated decision making process could result in discriminatory outcomes against specific demographic characteristics” (DSIT, 2023). However this policy is a small win against bias and not targeted specifically at gender-based policy issues. As Ansell and Gash put it, “small wins may not be an appropriate strategy for trust building where stakeholders have more ambitious goals that cannot easily be parsed into intermediate outcomes” (Ansell, Gash 2008).

The government “is working closely with the Equality and Human Rights Commission (EHRC) and [the Information Commissioner's Office] ICO to develop new solutions to address bias and discrimination in AI systems” (DSIT, 2023). It has sponsored an innovation challenge aimed at developing novel solutions to prevent AI and algorithmic bias. The winning proposals focus on the higher education, healthcare, financial and recruitment sectors. Gender equity is a common thread running through these sectors but is not specifically addressed. Based on this, while there is evidence in current UK government policy of an attempt to discourage algorithmic bias in AI systems, the absence of comprehensive pluralism could easily undermine the spirit of the policy by creating unintended consequences in the resulting decision-making (Baert, 1991).

Counter-finality for gender-based equity in AI decision-making

Baert's typology of unintended consequences helps us to understand the complexities and uncertainties involved in social phenomena and policy interventions, emphasising the importance of considering unintended consequences in decision-making processes. Particularly relevant to the development of new AI technologies is the concept of "counter-finality" (Baert, 1991). This concept refers to the idea that actions or interventions designed to achieve specific goals can sometimes produce outcomes that are contrary to those goals.

Baert uses this concept to highlight the complexities and uncertainties inherent in social processes and policymaking. He suggests that even well-intentioned efforts to address social problems or achieve certain objectives can result in unexpected and undesirable outcomes.

The potential pitfall in the development of AI for decision-making in government which lacks a pluralistic approach to the data sets, rules and requirements of the technology could easily result in development of a counter-finality where an intentional effort to alleviate bias of one type creates an unintended bias for another cohort or segment of the population.

Emergence as a property of AI as a complex adaptive system

In complexity theory, the concept of emergence refers to when complex patterns result from the combination of simpler components within a system, often in ways that are not predictable when examining those components alone. Like the behaviour of a flock of birds, emergence is when “nonlinear interactivity leads to novel outcomes that are not sufficiently understood as a sum of their parts” (Goldstein, 1999).

Essentially, emergence describes how “the whole is different from, not greater than, the sum of the parts” (Jervis, 1997) — meaning that the whole system exhibits properties that are not reducible to or predictable from the properties of its individual parts.

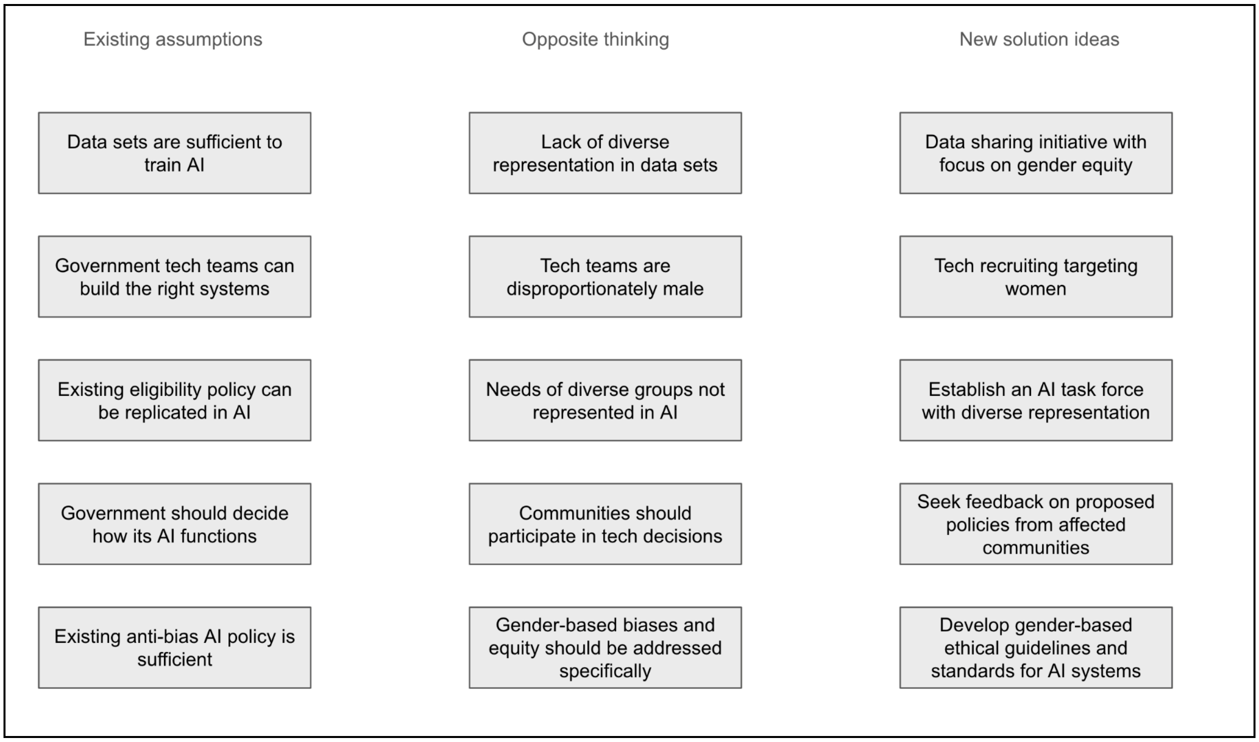
The use of artificial intelligence or machine learning for automated decision-making for public services is an example of how emergence can be applied to technological systems. This is especially true because many AI systems use opaque algorithms to make decisions without transparency to humans (Grimmelikhuijsen, 2023).

Human-centred AI policy intervention to reduce gender-based inequity

Policy intervention is essential to address the complex challenges surrounding equitable AI development and ensure that AI systems serve the interests of all citizens. However, with something as complex as AI decision-making, it can be challenging to arrive at a policy intervention that is sufficiently specific to impact the software development in a meaningful way. Therefore, the use of the “assumption reversal technique” (Vernon, Hocking, et. al., 2016) is helpful to identify areas where assumptions might be outdated or unquestioned.

In this case, I conducted an assumption reversal exercise with the goal of challenging current assumptions about the development of AI for complex decision-making within the UK government. It became apparent that while existing assumptions suggest confidence in the capabilities of government tech teams and the sufficiency of current policies, the opposing ideas highlight critical gaps that justify policy intervention for equitable AI development in government decision-making. Strikingly, the lack of diverse representation in both data sets and tech teams raises concerns about the inclusivity and fairness of AI systems.

Policy intervention is crucial to address these challenges and ensure that AI developed for government decision-making is created with diverse inputs. As Forester argues, “integrating inclusive participation and effective negotiation takes skill and preparation, thoughtfulness and a sense of humour, commitments to fairness and joint gains” (Forester, 2009). To this end, policymakers can mandate initiatives to actively recruit diverse talent and ensure that data collection processes prioritise representation from various demographic groups.



Existing assumptions	Opposite thinking	New solution ideas
Data sets are sufficient to train AI	Lack of diverse representation in data sets	Data sharing initiative with focus on gender equity
Government tech teams can build the right systems	Tech teams are disproportionately male	Tech recruiting targeting women
Existing eligibility policy can be replicated in AI	Needs of diverse groups not represented in AI	Establish an AI task force with diverse representation
Government should decide how its AI functions	Communities should participate in tech decisions	Seek feedback on proposed policies from affected communities
Existing anti-bias AI policy is sufficient	Gender-based biases and equity should be addressed specifically	Develop gender-based ethical guidelines and standards for AI systems

Figure 2: An assumption reversal exercise to uncover new solutions to the use of AI for decision making in UK government public services.

Policies should be implemented to facilitate community engagement in tech decisions, allowing affected communities to provide input and feedback on AI development processes. Specific measures targeting gender-based biases should be integrated into existing anti-bias AI policies to ensure that AI systems promote gender equity and fairness in government decision-making.

Based on the outcomes of the above exercise, one example of a policy intervention that addresses gender-based inequity in government AI decision-making through a pluralistic approach, while adhering to a human-centred AI approach, could involve the establishment of an interdisciplinary task force or advisory board.

Objective of the policy: Form a human-centred AI task force

The aim of the proposed policy would be for the UK government to establish an interdisciplinary task force consisting of experts from diverse fields, including gender studies, AI ethics, law, sociology, and relevant technical domains, to address gender-based inequity in government AI decision-making processes.

The task force would develop and implement policies aimed at promoting gender equity and ensuring a human-centred approach to AI decision-making. This intervention would be evaluated based on the reduction of gender disparities in AI outcomes, increased transparency and accountability in AI decision-making processes, and the representation of marginalised voices in policy development.

By adopting a pluralistic approach that integrates diverse perspectives and stakeholders, this policy intervention seeks to minimise the risk of counter-finality by identifying and addressing potential unintended consequences of AI decision-making while advancing gender equity and promoting human-centred AI principles.

The human-centred AI task force would operate in several areas, including stakeholder engagement and analysis, analysis of AI outcomes and their underlying data sets, and contributing to guidelines and standards for the development and deployment of government AI to ensure the reduction of gender-bias in decision making.

Stakeholder Analysis

The task force would be responsible for engagement with a variety of stakeholders across community organisations, interest groups, gender-equity advocates and the general public. The approach to stakeholder engagement should include interviews and user research following trauma-informed practices, in order to gain qualitative insight into how better to train AI systems to take into account real human needs.

Equity of data sets

In addition to qualitative analysis and insights, the task force should examine the underlying data sets and historical data being used to develop and train machine learning algorithms and automated decision making. The group should study the impact of AI systems on different demographic groups, with a focus on gender disparities, testing these systems against numerous use cases to determine where bias might exist.

Guidelines and standards

Based on their findings, the task force would develop ethical guidelines and standards for the design, implementation, and use of AI systems in government decision-making. These guidelines would prioritise fairness, transparency, accountability, and inclusivity, with a specific emphasis on gender-based biases and equity.

Policy implementation: bring human experience to software development practice

An ideal jumping-off point for the implementation of this policy intervention would be to make the evaluation of government AI for gender bias into an integral part of the “Service Assessment” process which all government services must adhere to as part of the “Service Standard” (Central Digital and Data Office, 2024). The human-centred AI task force could sit centrally in the Cabinet Office as part of either the Central Digital and Data Office (CDDO) or the Government Digital Service (GDS).

Spend control for new digital and data technology projects is delegated from HM Treasury to CDDO, and therefore any new instances of AI for decision making would need to pass through the task force both before receiving initial funding and also right before going live.

This central monitoring, evaluation for bias and control of deployment could help reduce the risk of undesirable emergence from unfettered use of AI within departments, and it could also help to provide a more joined-up user experience for the public when accessing government services.

By embracing pluralism and actively involving a range of stakeholders—including women's advocacy groups, gender scholars, and marginalised communities—in the design phase of AI systems, governments can address the root causes of gender inequality and create more inclusive and equitable decision-making processes.

Conclusion

This essay has examined the complex relationship between artificial intelligence (AI) technology and gender-based equity within the context of government decision-making processes. It has highlighted the importance of addressing gender biases in AI systems to ensure fair and inclusive outcomes for all citizens. By adopting a pluralistic approach that embraces diverse perspectives and stakeholders, we can design policy interventions that promote human-centred AI principles and advance gender equity.

Key points discussed include the recognition of AI as a complex adaptive system within government service provision, the need for pluralism and gender representation in AI development, and the potential unintended consequences of AI decision-making without proper oversight. Drawing upon complexity theory, I have explored how policy interventions can mitigate gender-based inequities in government AI through inclusive decision-making processes.

The proposed policy intervention, centred around establishing an interdisciplinary task force or advisory board, offers a tangible solution to address gender bias in government AI decision-making. By mandating diverse representation in both data sets and tech teams, facilitating community engagement, and integrating specific measures targeting gender-based biases into existing policies, we can reduce bias and promote equity.

If the UK government were to prioritise the implementation of such a policy intervention, it would be possible to create a more equitable and inclusive government AI ecosystem that serves the needs and rights of all genders. It would require effort from policymakers, politicians, technologists, advocates, and the broader public to collaboratively design and implement policies that uphold the principles of fairness, transparency, and accountability in AI decision-making. Only then can we truly harness the potential of AI technology to advance equality for all.

References

Ansell, C., & Gash, A. (2008). Collaborative Governance in Theory and Practice. *Journal of Public Administration Research and Theory*, 18(4), 543–571.

Baert, P. (1991). Unintended Consequences: A Typology and Examples. *International Sociology*. 208

Barford, A. & Gray, M. (2022). The tattered state: Falling through the social safety net. *Geoforum*, Volume 137. 116

Browne, J., Drage, E., & McInerney, K. (2024). Tech workers perspectives on ethical issues in AI development: Foregrounding feminist approaches. Big Data Society, 11, Article 1.

Central Digital and Data Office, UK Cabinet Office (CDDO). (2024). Service Standard. https://www.gov.uk/service-manual/service-standard

Connolly, W. (2005). Pluralism, *Duke University Press*. 9

Department for Science, Innovation, & Technology (DSIT). (2023) A pro-innovation approach to AI regulation.

https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

Forester, J. (2009). Dealing with Differences : Dramas of Mediating Public Disputes. *Oxford University Press.* 180

Forester, J. (2018). Deliberative Planning Practices - Without Smothering Invention: A Practical Aesthetic View. *The Oxford Handbook of Deliberative Democracy*, Oxford Handbooks.

Frennert, S. (2021). Gender Blindness: On Health and Welfare Technology, AI and Gender Equality in Community Care. *Nursing Inquiry* 28.4. 12

Goldstein, J.(1999). Emergence as a Construct: History and Issues. *Emergence* vol 1 issue 1, *Lawrence Erlbaum Associates*. 53

Grimmelikhuijsen, S. (2023). Explaining Why the Computer Says No: Algorithmic Transparency Affects the Perceived Trustworthiness of Automated Decision‐Making. *Public administration review* 83.2. 241–262.

Human rights watch: E.U. (2021). Artificial intelligence regulation threatens social safety net. Targeted News Service. https://www.proquest.com/wire-feeds/human-rights-watch-e-u-artificial-intelligence/docview/2595941831/se-2

Jervis, R. (1997). System Effects: Complexity in Political and Social Life. Princeton University Press.

Sterman, J. D. (2001). System Dynamics Modeling: Tools for Learning in a Complext World. *California Management Review*, 43(4), 12

Valls, A. & Gibert, K. (2022). Women in Artificial Intelligence (AI). *Basel: Multidisciplinary Digital Publishing Institute*. 12

Vernon, D., Hocking, I., & Tyler, T. C. (2016). An Evidence-Based Review of Creative Problem Solving Tools: A Practitioner's Resource. *Human Resource Development Review*, 15(2), 230-259.