Diagnosing and Addressing Emergent Harms in the Design Process of Public AI and Algorithmic Systems

Nouws, Sem; Martinez De Rituerto De Troya, Íñigo; Dobbe, Roel; Janssen, Marijn

DOI
10.1145/3598469.3598557

Publication date
2023

Document Version
Final published version

Published in
Proceedings of the 24th Annual International Conference on Digital Government Research - Together in the Unstable World

Diagnosing and Addressing Emergent Harms in the Design Process of Public AI and Algorithmic Systems

Sem J.J. Nouws
Delft University of Technology
s.j.j.nouws@tudelft.nl

Íñigo Martínez de Rituerto de Troya
Delft University of Technology
i.martinezderituertodetroya@tudelft.nl

Roel I.J. Dobbe
Delft University of Technology
r.i.j.dobbe@tudelft.nl

Marijn F.W.H.A. Janssen
Delft University of Technology
m.f.w.h.a.janssen@tudelft.nl

ABSTRACT

Algorithmic and data-driven systems are increasingly used in the public sector to improve the efficiency of existing services or to provide new services through the newfound capacity to process vast volumes of data. Unfortunately, certain instances also have negative consequences for citizens, in the form of discriminatory outcomes, arbitrary decisions, lack of recourse, and more. These have serious impacts on citizens ranging from material to psychological harms. These harms partly emerge from choices and interactions in the design process. Existing critical and reflective frameworks for technology design do not address several aspects that are important to the design of systems in the public sector, namely protection of citizens in the face of potential algorithmic harms, the design of institutions to ensure system safety, and an understanding of how power relations affect the design, development, and deployment of these systems. The goal of this workshop is to develop these three perspectives and take the next step towards reflective design processes within public organisations. The workshop will be divided into two parts. In the first half we will elaborate the conceptual foundations of these perspectives in a series of short talks. Workshop participants will learn new ways of protecting against algorithmic harms in sociotechnical systems through understanding what institutions can support system safety, and how power relations influence the design process. In the second half, participants will get a chance to apply these lenses by analysing a real world case, and reflect on the challenges in applying conceptual frameworks to practice.

KEYWORDS

Artificial Intelligence, data science, public sector, design process, system safety, institutional design, power analysis

ACM Reference Format:
Sem J.J. Nouws, Íñigo Martínez de Rituerto de Troya, Roel I.J. Dobbe, and Marijn F.W.H.A. Janssen. 2023. Diagnosing and Addressing Emergent Harms in the Design Process of Public AI and Algorithmic Systems. In *24th Annual International Conference on Digital Government Research - Together in the unstable world: Digital government and solidarity (DGO 2023), July 11–14, 2023, Gdańsk, Poland.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3598469.3598557

DGO 2023, July 11–14, 2023, Gdańsk, Poland

ACM ISBN 979-8-4007-0837-4/23/07.
https://doi.org/10.1145/3598469.3598557

1 INTRODUCTION

Algorithmic and data-driven systems are increasingly used in the public sector to improve the efficiency of existing services or to provide new services through their newfound capacity to process vast volumes of data. Unfortunately, certain systems have also resulted in negative impacts on citizens, in the form of discriminatory outcomes, arbitrary decisions, lack of recourse, and more. These have serious consequences for citizens ranging from material to psychological harms [3], [4]. These harms partly emerge from choices and interactions in the design process [6]. Scientific research does not seem to provide public organisations with practical leads to organise their design processes around the prevention and correction of algorithmic harms. Approaches to address citizen harms in AI literature either take a policy or technical perspective (e.g., governance frameworks [9] and bias detection tools [7]). Often, these approaches do not address whether and how they can effectively be implemented in practice, disregard the power dynamics between actors in design processes, and do not acknowledge lessons learned from automation in other disciplines. This workshop addresses these issues by integrating three perspectives that can support more reflective and critical design practices in public administration.

The three perspectives we will present in this workshop are system safety, institutional design, and power analysis. First, insights from systems safety - with its tradition in software-based automation and origins in engineering - can provide the right structure or institutions to attain more control on the design process of public AI systems. Second, systems safety components need to be situated in the public administration context of these systems. Therefore, we argue that institutions on responsibility, accountability and representation need to be incorporated in the design process. Finally, an awareness and treatment of power relations during design, development, and deployment is largely absent from existing systems engineering practices. While power relations are inherent to institutional dynamics – including in the definition of safety control structures –, they are not always made explicit. We argue that an analysis of power here is crucial for protecting citizens from algorithmic harms, and for affording different actors the possibility to bring to light and address potential system safety hazards. We will show that these themes are lacking in current

design practices and argue that they are necessary for building safe and just systems within public organisations.

The goal of this workshop is to develop the three perspectives and take the next step towards reflective design processes within public organisations. The workshop will be divided into two parts. In the first half we will elaborate the conceptual foundations of these perspectives in a series of short talks. Workshop participants will learn new ways of protecting against algorithmic harms in sociotechnical systems through understanding what institutions can support system safety, what their limitations are, and how power relations influence the design process. In the second half, participants will get a chance to apply these lenses by analysing a real world case, and reflecting on the challenges in applying conceptual frameworks to practice.

2 PART I: PERSPECTIVES ON SYSTEM SAFETY, INSTITUTIONAL DESIGN AND POWER RELATIONS IN DESIGN

In the workshop, we will consider design processes of public AI systems as sociotechnical processes. In other words, they reflect the interacting, sociotechnical components: institutions (i.e„ institutional design), human agents (i.e., power analysis), and technical artefacts (i.e., system safety). Each perspective emphasises a different component in the design process. Moreover, the perspectives are interconnected: institutional design and power analysis both consider the relationship citizen and government; institutional design and system safety both underline the need for the "right" institutional environment; and, power analysis and system safety both consider the effect of hierarchical structures in practice.

2.1 Institutions for public design processes

Existing design practices for AI systems in the public sector often lead to unintended consequences that can harm citizens [6]. A way to change these practices is to implement the appropriate institutional environment for a design process [5]. This environment should adhere to the fact that: (1) public AI systems are sociotechnical systems in nature, and that (2) they are embedded within public organisations. The first fact asks for institutions that fit sociotechnical design characteristics. The latter fact asks for a design process that adheres to the premises of democracy and the Rule of Law, which varies across jurisdictions and cultures. The premises of these concepts are known, but have yet to be translated to the practice of sociotechnical systems design.

2.2 Power relations in sociotechnical systems

Attention to how power relations affect sociotechnical systems design is largely absent from contemporary sociotechnical design practices, yet they underpin several aspects of system design. Who decides the purpose and scope of a system? Who determines the technical specifications? What issues are open to discussion and which ones are kept off the table? It has become commonplace to seek to foresee unintended consequences of sociotechnical systems by encouraging the participation of affected stakeholders. However, we must also ask to what extent participation is or can be meaningful and binding, and in what instances does it merely serve as a performative exercise that demands resources from participants but does not deliver on their demands [8]. Attending to power in the design process can help us understand how sociotechnical systems can entrench existing social relations, and to what extent they may empower or disempower not only the public, but also those involved in the development and deployment of these systems [1].

2.3 System safety

Ensuring safety in public AI systems is crucial to reducing algorithmic harms [2]. The fields of systems engineering and control theory have a long tradition in research on system safety. This research assumes that measures to reduce a system's harms cannot only be based on technical design choices on the model or algorithm alone. Instead, there is a need for an end-to-end hazard analysis and design frame that includes the context of use, impacted stakeholders, and the institutional environment in which the system operates. Safety and other values are then inherently socio-technical and emergent system properties that require design and control measures to instantiate these across the technical, social, and institutional components of a system. Apart from these measures, system safety principles are also important starting points for organising design processes and the subsequent processes of use, maintenance, and governance.

3 PART II: JOINT LEARNING ACTIVITY

In the second half of the workshop, participants will put these frameworks into practice through a think-pair-share exercise by applying them to one of several real world case studies of AI systems in the public sector: (i) the use of fraud risk profiling in the childcare benefit scandal in the Netherlands, (ii) COVID-19 contract-tracing apps, (iii) cross-border surveillance of migrants. These will be presented through a video presentation by researchers who have worked on the cases.

There will be ample room for feedback and opportunities to further co-develop the frameworks in situ. We will facilitate this joint learning activity through a few guiding questions regarding the translation of our perspectives (or other research on design processes) to practice. We would like to answer the following questions in the workshop:

What other perspectives for diagnosing and addressing algorithmic harm are missing in design process research in general or in this workshop?

What are obstacles to implementing reflective design practices for sociotechnical systems in public organisations?

Why have insights in our perspectives – that digital government research has touched upon for years – not trickled down to practice?

What possibilities do researchers have to bring their insights to practice – assuming that the development of frameworks or design science approaches may not be fully effective in changing practices in real-world design processes?

Through these questions, we aim to improve our research on design processes of public algorithmic systems, as well as our engagement with actors in the field. Moreover, the answers to these questions may bring forward insights for safer forms of digital government in general.

REFERENCES

[1] Barabas, C., Doyle, C., Rubinovitz, J., & Dinakar, K. (2020). Studying Up: Reorienting the study of algorithmic fairness around issues of power. In 2020 Conference on Fairness, Accountability, and Transparency (pp. 167-176).

[2] Dobbe, R. (2022, June). System Safety and Artificial Intelligence. In 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1584-1584).

[3] Eubanks, V. (2018) Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press.

[4] Frederik, J. (2021) De tragedie achter de toeslagenaffaire. De Correspondent, 15 January, 2021 https://decorrespondent.nl/11959/de-tragedie-achter-de-toeslagenaffaire/719468974741-fc85ca00

[5] Koppenjan, J. & Groenewegen, J. (2005). Institutional design for complex technological systems. Int. J. Technology, Policy and Management, 5(3), pp. 240-257.

[6] Nouws, S., Janssen, M., & Dobbe, R. (2022). Dismantling Digital Cages: Examining Design Practices for Public Algorithmic Systems. In: International Conference on Electronic Government (pp. 307-322). Springer, Cham.

[7] Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani. 2018. Aequitas: A bias and fairness audit toolkit. arXiv preprint arXiv:1811.05577 (2018).

[8] Sloane, M., Moss, E., Awomodo, O. & Forlano, L. (2020). Participation is not a Design Fix for Machine Learning. In Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

[9] Wirtz B.W. & Müller W.M. (2019). An integrated artificial intelligence framework for public management. Public Management Review, 21(7), pp. 1076-1100.