Fostering trustworthy AI in the public sector

Stephan Grimmelikhuijsen

Associate professor, Utrecht University School of Governance (NL)

June 8, 2022

@Stephangrim

Algorithm use & Dutch tax scandal

Two issues with algorithmic transparency

Accessibility

- Availability of code, model, data
- External experts/auditing
- Analyze bias and functionality

No access

Explainability

- Explained outcomes
- Understandable to human
- Gives reasons

Not explained

- Grimmelikhuijsen, S. (2022). Explaining why the computer says no: algorithmic transparency affects the perceived trustworthiness of automated decision-making. *Public Administration Review*. https://doi.org/10.1111/puar.13483

Can algorithmic transparency increase perceived trustworthiness of AI & bureaucrats?

Linking transparency and trust

- Representative group of ~1000 Dutch citizens read two vignettes
  - Imagine your visa is rejected
  - Imagine you get a house search for suspected welfare fraud
- Then asked if they trusted
  - The algorithm/computer system
  - The bureaucrat using the algorithm in decision
- Citizens were randomly assigned to one condition
  - With or without proper access AND
  - With or without proper explanations

- Grimmelikhuijsen, S. (2022). Explaining why the computer says no: algorithmic transparency affects the perceived trustworthiness of automated decision-making. *Public Administration Review*. https://doi.org/10.1111/puar.13483

Please read the following situation thoroughly. On the next page you will get a few questions.

Imagine: You frequently travel abroad for your work, such as to countries in the Middle East. For a new and important project, you request an online visa to access the United States. Your visa application is being rejected by the computer system. Now you have to take a day off to travel to the United States consulate in Amsterdam for a new application. You call with the visa department to ask why your application has been rejected.

The employee of the visa department says that the decision was based on a computer system.

[Random assignment to one of the conditions below]

- The underlying code (algorithm) of the computer system is not accessible: the functioning of the computer system cannot be determined. [low accessibility]
- Th cessible online to stem functions an

So… what did I find?

[Random

- The computer system only indicates that your visa is rejected but not what the reason is behind the rejection. [low explainability]
- The computer system indicates that your visa is rejected, because in the past five years you have travelled at least once to a 'suspect' country. [high explainability]

Findings

- Explainability increased trust in algorithm in both visa application & welfare fraud case
- Explainability increased trust in bureaucrat only in welfare fraud case
- Accessibility increased trust in algorithm only in welfare fraude case

Grimmelikhuijsen, S. (2022). Explaining why the computer says no: algorithmic transparency affects the perceived trustworthiness of automated decision-making. *Public Administration Review*. https://doi.org/10.1111/puar.13483

Algorithmic transparency matters…

- Accessibility matters for accountability, but not enough for trustworthiness
- We need explainable AI for perceived trustworthiness
- Particularly needed in more intrusive AI applications

… but is not enough

- *Transparency can be gamed, ignored or go unnoticed!*

- Regulation is needed but complicated
  - Enforcement & compliance
  - New tasks, new regulator?
  - Trust between regulators, regulatees and citizens
  - https://www.tigre-project.eu/info-corner/

TiGRE
Trust in Governance and Regulation in Europe

… but is not enough

- *Transparency can be gamed, ignored or go unnoticed!*

- Rethinking instutional safeguards
  - Strengthening democratic control over algorithm use
  - Participation in algorihmic design to prevent 'technical rationality'
  - Right to human contact
  - Cost-benefit analyses and periodic audits

- Grimmelikhuijsen, S., & Meijer, A. (2022). Legitimacy of Algorithmic Decision-Making: Six Threats and the Need for a Calibrated Institutional Response. *Perspectives on Public Management and Governance*. https://doi.org/10.1093/ppmgov/gvac008

Take home messages

1. Accessibility and explainability are two important compinents algorithmic transparency

2. We need accessibile AI for accountability, but in addition we need explainable AI for perceived trustworthiness

3. To achieve this, we need effective regulatory regimes and institutional re-calibration

Thank you for listening

- Contact:
  - S.g.grimmelikhuijsen@uu.nl
  - @Stephangrim (Twitter)

- More info:
  - www.uu.nl/staff/SGGrimmelikhuijsen (personal profile on UU website)
  - https://algopol.sites.uu.nl/ (project: algorithm use by the police)
  - www.tigre.eu (project: trust in regulation)
  - https://scholar.google.nl/citations?user=jDtNbekAAAAJ&hl=nl (overview of publications)