# Governmental AI projects and perceived trustworthiness: evidence from two survey experiments on the limits of communicating trustworthiness towards citizens

Bjorn Kleizen, Wouter van Dooren, Koen Verhoest

University of Antwerp, GOVTRUST Centre of Excellence

# Starting point

Trust in governmental AI projects is important for legitimacy, take-up, cooperation, stability and to prevent projects from being discredited

We do not really know what citizens perceive to be trustworthy and what they support

Can we convince citizens on the project level that our AI project is trustworthy?

What is the role of pre-existing attitudes and perceptions?

# Designing trustworthy and fair big data and AI-based supervision

Different approaches, but EU HLEG on AI provides starting point through their guidelines!

Trustworthy AI should be (1) lawful, (2) ethical and (3) robust.

**7 principles:**

1. Human agency and oversight
2. Technical Robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, non-discrimination and fairness
6. Societal and environmental well-being
7. Accountability

# Communication noise?

**Insights from the interviews**

Governments (and their private partners) generally optimistic that trustworthiness of AI can be ensured by enhancing the trustworthiness of specific AI projects

- particular emphases on technical, ethical AI and legal dimensions

NGO's do not seem to evaluate specific projects, instead relying on general notions of how AI in government 'works'

- Trust breaches seem to act as salient cases, with respondents using information from these cases and extrapolating to their evaluation of other projects
- Technical, ethical and legal safeguards in specific projects are not emphasized strongly

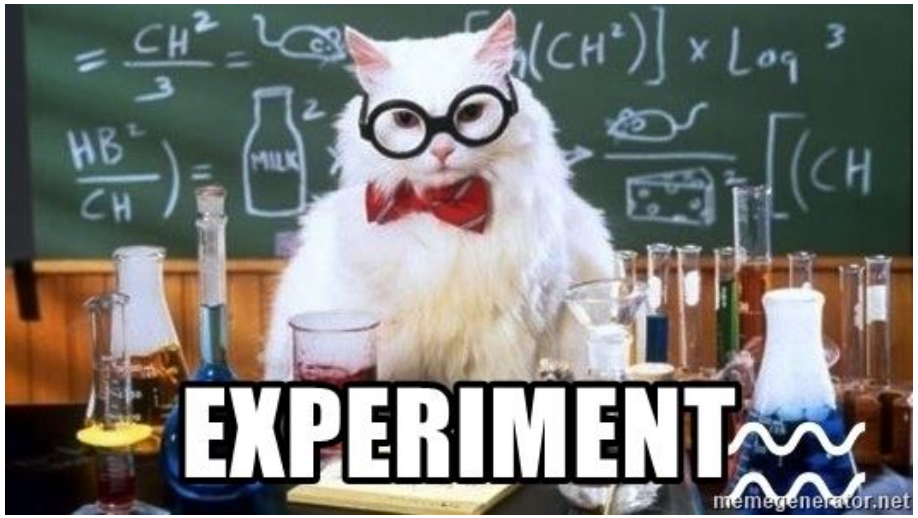
This process creates a mismatch, in which governments focus on internal measures to enhance trustworthiness, even though these effects do not play a large role in external evaluations of trustworthiness

# How to test this?

General information on three (hypothetical)

AI projects:

1. Road maintenance
2. Tax fraud detection
3. Visitor flows during events




Then:

| Experiment 1 | Experiment 2 |
|---|---|
| Information on legal compliance | Information on human-in-the-loop |
| Information on no-harm and explainability | Information on fairness & non-discrimination |
| Information on data-gathering | Information on technical robustness |

# Results experiment 1

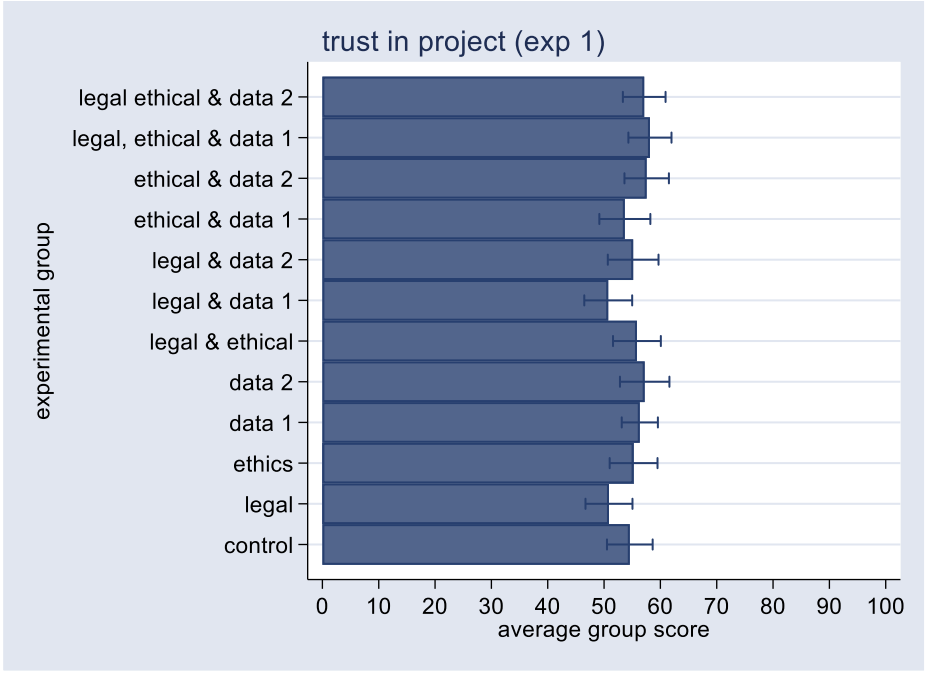
Randomize respondents

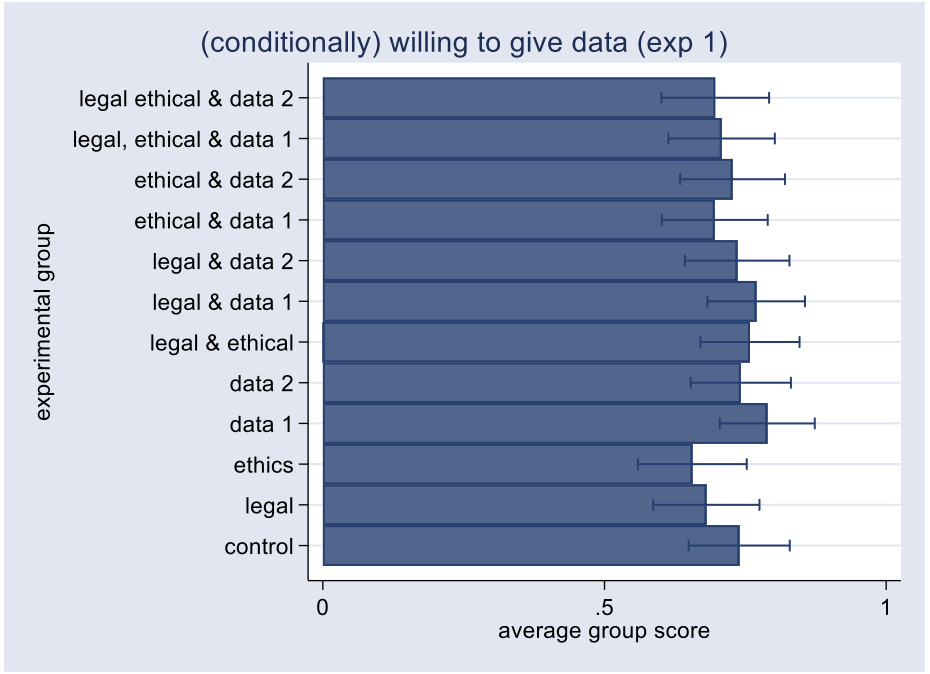
across 12 (exp. 1) or 8 (exp. 2) groups

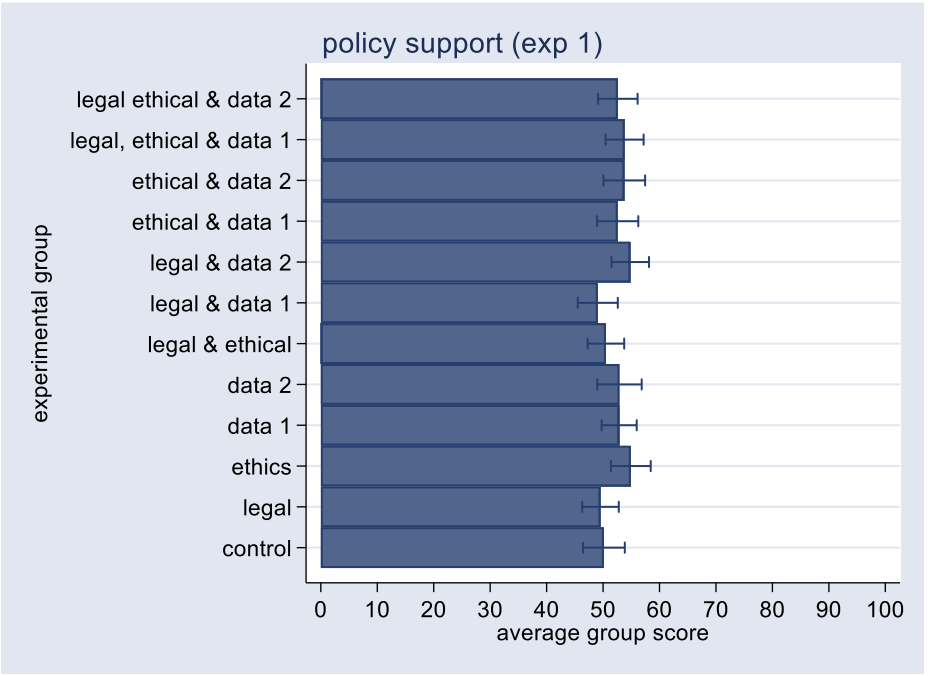
- Control (baseline)
- Legal (baseline + legal information)
- Legal & ethical (baseline + legal information + ethical AI information)

# Results experiment 1

- Little differences between experimental groups
- Information on fairness, legal compliance and data-gathering has little influence

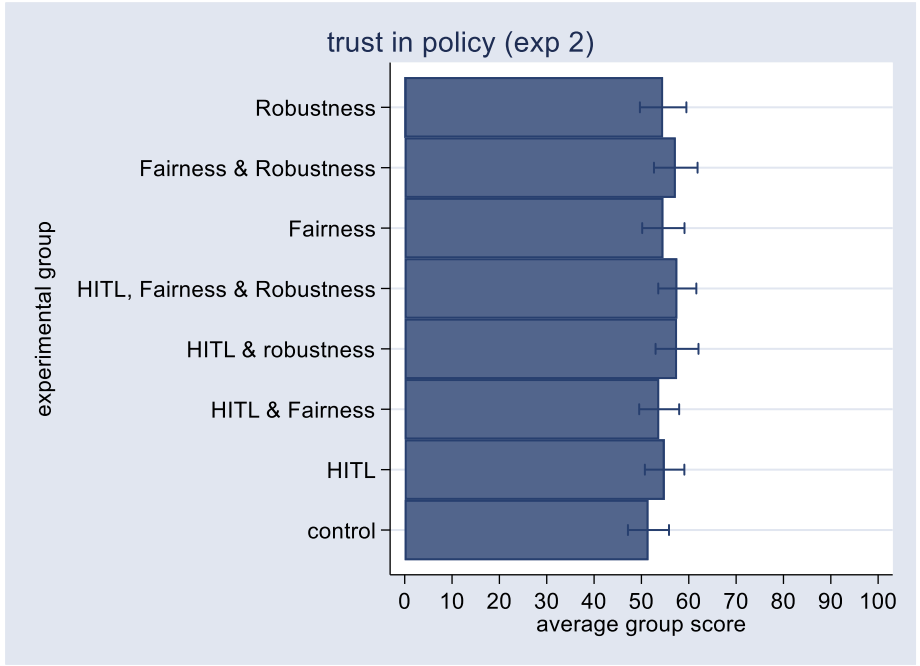


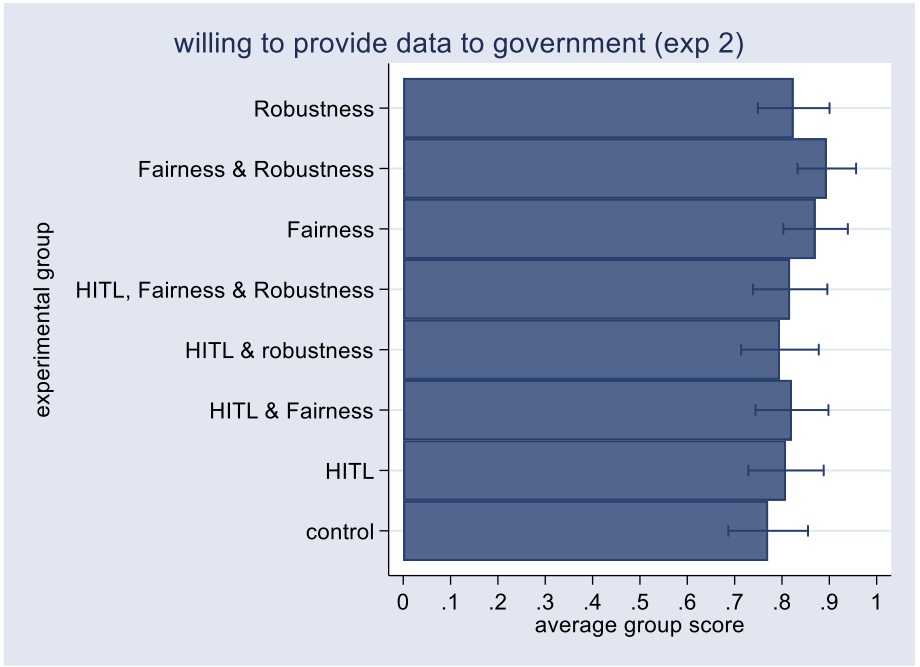



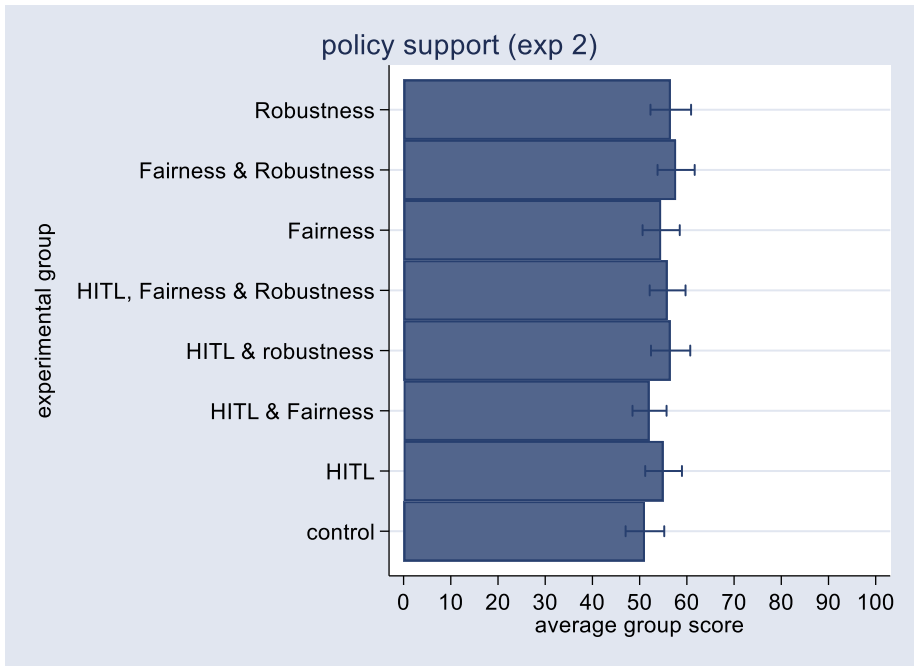



# Results experiment 2

- Very similar, although fairness & robustness tentatively suggests a minor positive effect, this result does not seem robust.



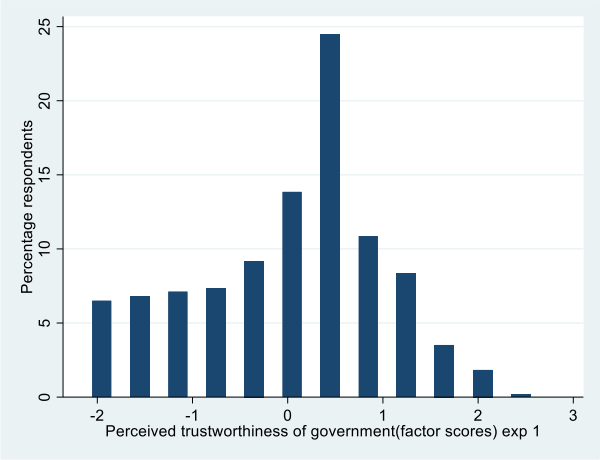
# Pre-existing attitudes, characteristics and perceptions

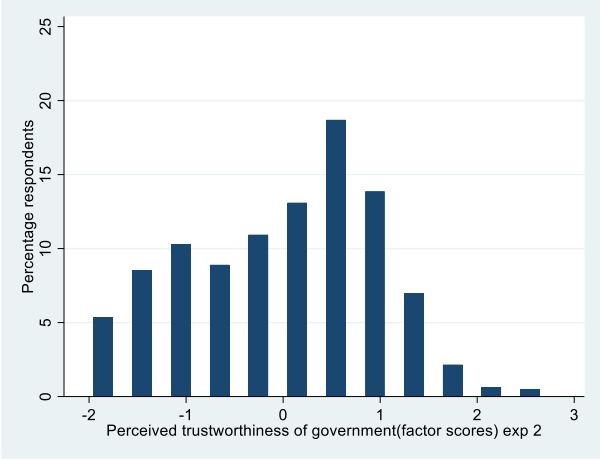
Attitudes are mostly explained by pre-existing attitudes and perceptions

| Intervention | Trust | Policy support | Giving data |
|---|---|---|---|
| Trust government | N.a. | N.a. | Consistently positive |
| Trust AI | N.a. | N.a. | Consistently positive |
| Privacy concern | Consistently negative | Consistently negative | Consistently negative |
| Ai use in job | Null | Mostly null (although some models do show significance) | Null |
| Discrimination | Consistently negative | Consistently negative | Null |

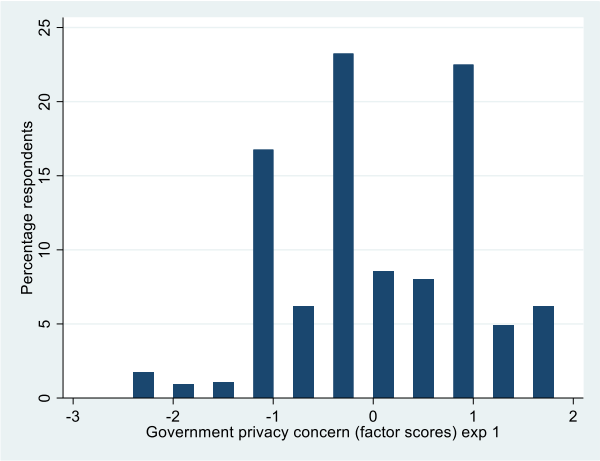
# The relevance of these variables

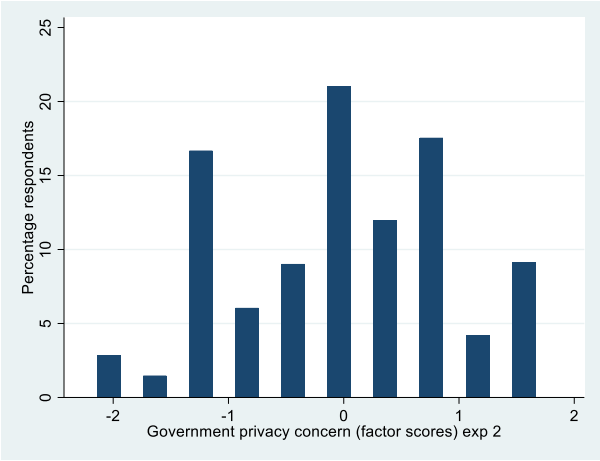
**Perc. trustworthiness of government**




**Privacy concern**




**Experiment 1**

**Experiment 2**

"I have more trust in this than what the private sector does"

"Privacy, the magic word government uses when it suits them"

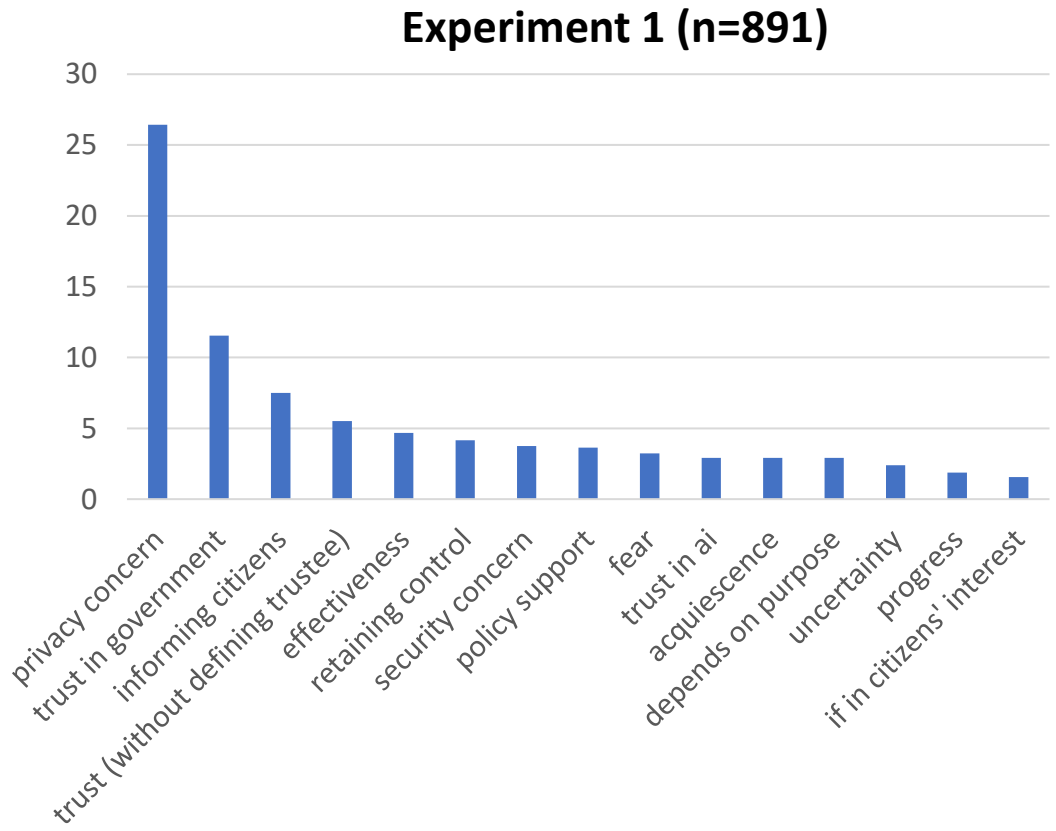
"Often automatic processes make you guilty until proven innocent … benefits were withheld from me because the computer had accidentally switched a date between me and my mother!"

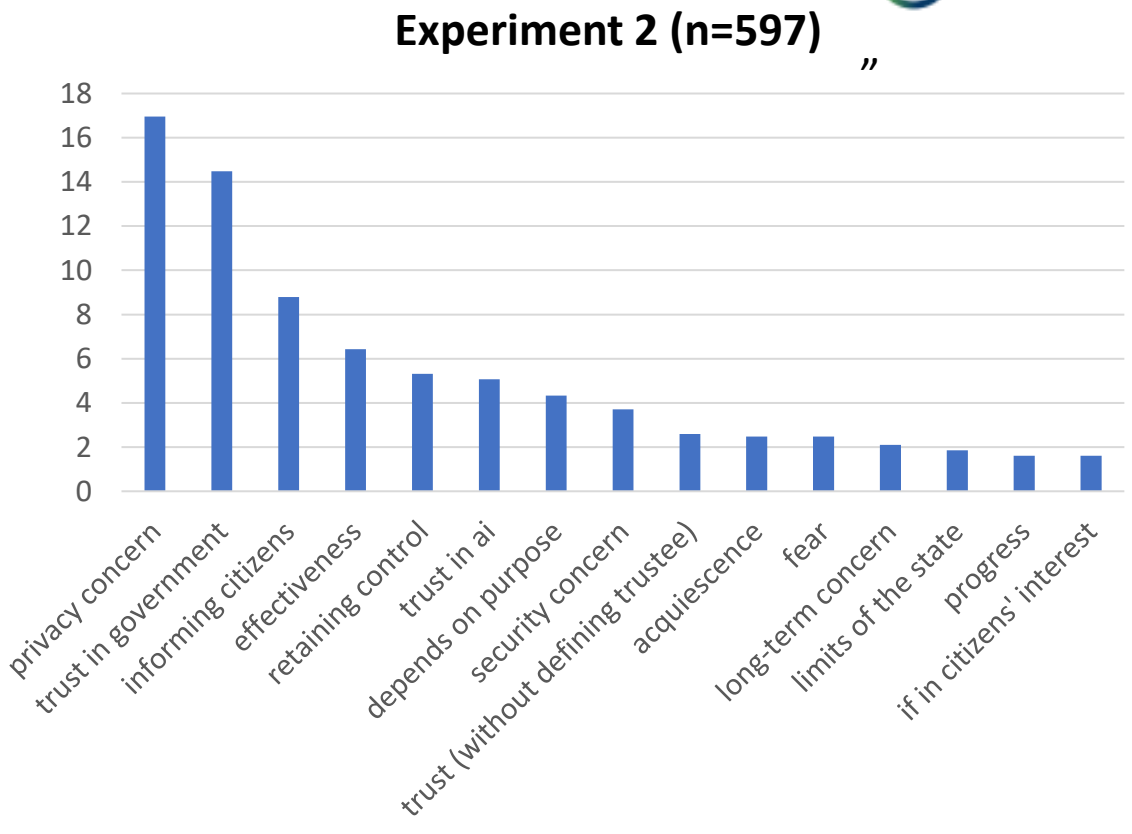
"I have nothing to hide"

"'Big brother' is watching"

**Open field**

# Some qualitative triangulation




1. Mostly pre-existing attitudes, opinions and experiences

2. Some promise of information provision (including defining purpose), retaining control and using AI to improve services for citizens

# In conclusion

Trustworthiness is a good aim from a good governance viewpoint

However, on the project- and organizational level our research has the following implications:

- do not expect short-term wonders of AI/project design in terms of citizen trust/support, as pre-existing attitudes provide a (far) better prediction of trust in/support for an AI project

- Although many people positive towards use of AI in government, there are also substantial groups in society with negative to very negative attitudes

# The way forward

1. Usefulness of AI guidelines and similar initiatives may lie in preventing trust breaches and fostering a long-term ‘ecosystem of trust’, instead of being a quick fix
2. Use of heuristics suggests that some coordination between public authorities is desirable
3. Take note of diverging opinions in society (e.g. role played by discrimination or ethical viewpoints on privacy and the ‘reach’ of governments)
4. Testable hypothesis: the more specific an NGO or civilian has an association with your project, the more likely you are to induce a trust response
    - This would suggest involving NGO’s and civil society in projects
    - XAI for immediate users
    - Do not forget that such measures are unlikely to produce an immediate effect among the general populace
5. It may be possible to pre-empt trust breaches by focusing on less salient applications (e.g. chatbots versus risk-profiling). Note the link with principles such as proportionality.
6. Open question: measurable role of (foreign) scandals?

| | Text vignette |
|---|---|
| Baseline information (presented to all respondents, in both experiments (including control groups)) | Artificial intelligence (AI) in federal government projects<br>Governmental organizations increasingly work digitally. To that end, the federal government has decided to focus on Artificial Intelligence (AI). This concerns multiple projects, including:<br>- A joint project with Wallonian and Flemish governments to recognize damage to roads using artificial intelligence. As the computer recognizes potholes in roads, maintenance can be organized more efficiently.<br>- The inspection of tax returns. Using various data, the probability that someone has committed fraud in his or her tax returns is predicted. By focusing on tax returns with a high probability of fraud, inspectors can more easily detect irregularities.<br>- Following streams of people through their mobile phones during events to predict where emergency services (such as ambulances) might be necessary. This helps emergency services to better anticipate swiftly changing situations.<br><br>However, independent experts are posing questions on privacy and data security. Also, due to the complexity of artificial intelligence it is not always clear on what basis a computer program takes a certain decision. |

| Interventions experiment 1 | |
|---|---|
| Legal information (legal) | The federal government acknowledges that there are legal concerns. To that end, the government has hired several independent data lawyers who will supervise the projects. A legal base that determines what governmental organizations can and cannot use AI for will also be established. |
| Ethical AI information (ethical) | - Governments must always ensure that the decisions of their artificial intelligence are completely explainable;<br>- Artificial intelligence must always be deployed in the interest of citizens. For instance, the federal government may not use artificial intelligence to detect minor mistakes made by citizens, but may do so for major fraud cases. |
| Data-gathering information 1, internal data-gathering (data 1) | Taking into account privacy considerations, the federal government limits itself to data it has gathered on its own. |
| Data-gathering information 2 (data 2), anonymized data from private parties | Taking into account privacy considerations, the federal government only uses anonymized data from private businesses. For instance, the following data is anonymously gathered through businesses for the projects concerning damage to roads, tax fraud and flows of visitors to events:<br>- Photographic material from private construction companies (such as businesses working on roads)<br>- Wage- and administrative data from employers<br>- Mobile location data from telecom service providers |

| Interventions experiment 2 | |
|---|---|
| Human-in-the-loop information (HITL) | To prevent mistakes, the federal government demands that humans will always be involved in decisions based on artificial intelligence. Civil servants that know how artificial intelligence works thoroughly evaluate the outcomes of the computer program. Only after this evaluation a final decision is taken. |
| Fairness & non-discrimination information (Fairness) | Artificial intelligence can unintentionally discriminate against vulnerable groups in society. To prevent this, civil servants extensively study each artificial intelligence project of the federal government. Should the chance of discrimination be high, then stringent extra checks are necessary to keep the projects honest. |
| Technical robustness (Robustness) | To safely store sensitive data, the federal government uses new technologies such as blockchain. This blockchain registers every attempt to access data in a permanent and tamper-proof way. This allows the federal government to always find out who had access to the data, so that citizens' data is protected better. |