Contents lists available at ScienceDirect

Government Information Quarterly

journal homepage: www.elsevier.com/locate/govinf

Do citizens trust trustworthy artificial intelligence? Experimental evidence on the limits of ethical AI measures in government

Bjorn Kleizen [a,*], Wouter Van Dooren [a], Koen Verhoest [a], Evrim Tan [b]

[a] *University of Antwerp, Faculty of Social Sciences, Department of Political Science, GOVTRUST Centre of Excellence, Sint-Jacobstraat 2, 2000 Antwerpen, Belgium*
[b] *KU Leuven, Faculty of Social Sciences, Public Governance Institute, Parkstraat 45 - box 3609, 3000 Leuven, Belgium*

ARTICLE INFO

Keywords:
Trust in AI
Ethical AI
Trust in government citizen trust
Transparency
Policy support

ABSTRACT

This study examines the impact of ethical AI information on citizens' trust in and policy support for governmental AI projects. Unlike previous work on direct users of AI, this study focuses on the general public. Two online survey experiments presented participants with information on six types of ethical AI measures: legal compliance, ethics-by-design measures, data-gathering limitations, human-in-the-loop, non-discrimination, and technical robustness. Results reveal that general ethical AI information has little to no effect on trust, perceived trustworthiness or policy support among citizens. Prior attitudes and experiences, including privacy concerns, trust in government, and trust in AI, instead form good predictors. These findings suggest that short-term communication efforts on ethical AI practices have minimal impact. The findings suggest that a more long-term, comprehensive approach is necessary to building trust in governmental AI projects, addressing citizens' underlying concerns and experiences. As governments' use of AI becomes more ubiquitous, understanding citizen responses is crucial for fostering trust, perceived trustworthiness and policy support for AI-based policies and initiatives.

1. Introduction

Artificial intelligence (AI) in the public sector holds the promise of innovation in public sector services (Ingrams, Kaufmann, & Jacobs, 2021). At the same time, it also necessitates a rethinking of the citizen-government relationship (Busuioc, 2021; Andrews, 2019; Winfield & Jirotka, 2018). Even the most benign AI projects may face societal fears and questions on intrusiveness. These fears are exacerbated by recent crises such as the discriminatory application of facial recognition in the UK or the discriminatory results on recidivism chances in the US (Andrews, 2019; Koniakou, 2023; Meijer & Wessels, 2019; Winfield & Jirotka, 2018)). In response, academics and governments attempt to improve the trustworthiness of public sector AI projects. Various sets of ethical AI principles, data science techniques, and governmental guidelines on developing and implementing trustworthy AI have been developed by researchers and applied by governments (Gunning & Aha, 2019; Winfield & Jirotka, 2018; Floridi et al., 2018; Hagendorff, 2020; Stahl et al., 2021; Koniakou, 2023).

Although new solutions to ethical design challenges are thus deemed necessary (Winfield et al., 2018; Hagendorff, 2020; Busuioc, 2021; Veale, 2020), empirical insights on their efficacy is limited (Stahl et al., 2021, Choung, David, & Ross, 2022). This leads Stahl et al. (2021) and Choung et al. (2022) to call for more empirical research on the effects of ethical AI frameworks on societal trust in AI. Similarly, Gesk and Leyer (2022) note that research on citizen acceptance of AI in government remains limited. Our contribution attempts to address both gaps by experimentally examining whether ethical AI information provided on public authorities' websites and press-releases can affect citizen attitudes towards public sector AI projects in the short-term, or whether AI attitudes will largely be shaped by citizens' pre-existing attitudes towards government, AI and/or privacy.

We focus not on users, but on the general public. Although research into trust and trust-building efforts for direct or potential users has been conducted (Aoki, 2020; Grimmelikhuijsen, 2023; Logg, Minson, & Moore, 2019, Alon-Barkat & Busuioc, 2023; Sullivan, de Bourmont, & Dunaway, 2022), the perceptions of wider, non-user audiences remain under-researched from an empirical perspective (see Ingrams, Kaufmann, & Jacobs, 2021 and Gesk & Leyer, 2022 for recent exceptions). We argue this is an oversight, as the average citizen will not directly interact with most governments' AI systems. Instead, most citizens can

[*] Corresponding author.
E-mail addresses: bjorn.kleizen@uantwerpen.be (B. Kleizen), wouter.vandooren@uantwerpen.be (W. Van Dooren), koen.verhoest@uantwerpen.be (K. Verhoest), evrim.tan@kuleuven.be (E. Tan).

https://doi.org/10.1016/j.giq.2023.101834
Received 18 January 2023; Received in revised form 31 March 2023; Accepted 7 May 2023

Please cite this article as: Bjorn Kleizen et al., *Government Information Quarterly*, https://doi.org/10.1016/j.giq.2023.101834

only observe the government decisions that an AI has contributed to in the background (e.g., when an AI is integrated into a broader IT system, such as roadside cameras, or when a civil servant takes a decision based (in part) on AI predictions). As these citizens do not interact with the system directly, they will instead have to rely on a combination of heuristics and more general information (e.g., through the media, social norms or press releases) to fill in gaps (Bitektine, 2011; Kostka, Steinacker, & Meckel, 2023; Venkatesh & Davis, 2000). This would be in line with expectations from STS studies and studies on technology acceptance, which argue that technologies are socially embedded, as they are built in pursuit of the values of their designers and are evaluated based on stakeholder values and norms (Fischer & Wenger, 2021; Greene, Hoffmann, & Stark, 2019; Kleizen, Van Dooren, & Verhoest, 2022; Sullivan et al., 2022). Our study innovates by empirically focusing and theorizing on the way that the general audience responds to information on ethical AI.

The notion that general communication on ethical AI measures may improve trust among the general audience obtains some support from earlier research in other policy areas. For instance, studies into public communication have found that both transparency and symbological information may promote trust (Alon-Barkat, 2020). Others suggest that the opposite might be true: efforts to communicate trustworthiness may yield only limited or mixed effects on trust and policy support (Kleizen & Van Dooren, 2023). Citizen knowledge of public organizations and projects is often limited, which stimulates citizens to employ heuristics and to rely on general impressions of the trustworthiness of – in this case – a government's AI project (Bayram & Shields, 2021; Bitektine, 2011; Ingrams, Kaufmann, & Jacobs, 2021; Pétry & Duval, 2017). Therefore, a citizen may evaluate an AI project's trustworthiness more on the basis of pre-existing beliefs and attitudes than on information on trustworthiness by governments (Ingrams, Kaufmann, & Jacobs, 2021; Sullivan et al., 2022).

Given that extant literature thus provides no clearcut expectations, this article tests two sets of hypotheses through two survey experiments. First, we examine hypotheses that expect an effect of at least some trust-building practices on perceived trustworthiness, trust, and policy support. Innovatively, we will also study the plausibility that they have no effect (which we accomplish with equivalence testing). Secondly, we test hypotheses on attitudinal and perceptional predictors of perceived trustworthiness, trust, and policy support – i.e., whether differences between citizens' prior attitudes explain trust better than our experimental interventions. Experimental strategies are pre-registered onto OSF.[1] Finally, we use qualitative data from an open answer box in the survey (asking whether citizens would entrust their data to the projects listed in the experimental vignettes) to triangulate experimental findings. Results suggest that citizen attitudes towards AI projects in government are mostly pre-determined and that governmental communication does little in the short-term to alter these perceptions (i.e., a null result for the experimental variables). Instead, pre-existing attitudes such as privacy concerns, trust in government, and trust in AI are the most consistent predictors of our outcomes. Perceived discrimination and self-reported professional use of AI also seem to possess some predictive power.

2. Theoretical framework

2.1. *Trustworthiness, trust in government, and trust in AI*

Trust is important for public organizations. The absence of trust erodes the legitimacy of public action and stifles the cooperation of citizens in policy implementation (Bayram & Shields, 2021), leading to underperformance and political sanctioning (Thomas, 1998). Moreover, where trust is breached, citizens may experience profound feelings of injustice, providing an impetus for sanctions (Thomas, 1998). In the context of algorithmic governance specifically, a lack of trust may cause citizens to believe they are being monitored excessively by public authorities (Meijer & Wessels, 2019), which may adversely affect societal support for the use of predictive modelling.

To appropriately discuss trust, it is necessary to first discuss some of its conceptual complexities. There are important conceptual differences between trust, trustworthiness, and perceived trustworthiness. Regarding trust, Rousseau, Sitkin, Burt, and Camerer (1998) and Mayer, Davis, and Schoorman (1995) argue that most disciplines see accepting vulnerability in relationships between trustor and trustee as the concept's core component. The trustor understands that actions of others can cause harm, but chooses to accept this risk (Hamm, Smidt, & Mayer, 2019). In this context, the definition offered by Mayer et al. (1995) has become relatively dominant (Hamm et al., 2019). Their definition proposes that trust *"is the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party"* (Mayer et al., 1995). Seen as such, trust forms an essential "social lubricant", that allows the trustor to accept some degree of uncertainty in interactions, based on positive expectations.

If trust is the willingness to accept vulnerability, trustworthiness refers to the characteristics that make a trustee worthy of being trusted (Grimmelikhuijsen & Knies, 2017). These characteristics may not be perfectly evaluated, however, as the trustor only has limited information available. Perceived trustworthiness therefore refers to a trustor's beliefs regarding the trustworthiness of the trustee (Latusek & Hensel, 2022; Logg et al., 2019), which may or may not lie close to actual degrees of trustworthiness. Researchers often use Likert scales to measure perceptions of trustworthiness (e.g., whether the trustee is perceived to be honest) instead of the actual decision to place trust in a trustee. This has led to some criticism, as studies based on Likert scales often purport to measure trust even though their measurements lies closer to trustworthiness (Hamm et al., 2019). In our study, we will attempt to take both perceived trustworthiness and the decision to trust into account (willingness to be vulnerable by providing data to governmental AI projects).

Trust in government is not the only literature line to account for. Our study is also informed by theories on trust in technology and trustworthy AI. A key component of trust and trustworthiness definitions in standard *human-to-human* settings are the trustee's motivations, morality, and interests, which form the basis for a trustor's beliefs about trustee benevolence (Hamm et al., 2019). In the *human-to-machine* trust setting, such motivations and interests do not exist, as AI has no independent motivations. When placing trust in technology, it has therefore been argued that trust instead reflects the feeling of certainty that the technology in question will not fail (Montague, Kleiner, & Winchester III, 2009). Somewhat analogous to the ability, benevolence, and integrity dimensions of other forms of trust (such as trust in government), Lankton, McKnight, and Tripp (2015) argue that trust in technology is based on the technology's functionality (does the technology have the required functionality?), helpfulness (does the technology provide help responsively and adequately?) and reliability (will the technology operate properly and consistently?).

Trust in technology scholars also recognize that perceptions of technologies (including AI) are embedded in social structures. The Technology Acceptance Model (TAM), for instance, accounts for factors such as social norms (Choung et al., 2022; Venkatesh & Davis, 2000). Similarly, Science & Technology studies (STS) have argued that (AI) technologies are inherently value-laden, in part due to the norms and values of developers and contractors, and in part due to political decisions on their purpose audience (Fischer & Wenger, 2021; Greene et al., 2019). What is more, their purpose and values will also be subjectively evaluated by stakeholders such as users and the general audience (Kleizen et al., 2022; Yeung, 2018). Groups perceiving privacy and a limited state as normatively important may have completely different

[1] For experiment 1, see: https://osf.io/2bnh3. For experiment 2, see: https://osf.io/p3e7k

ideas on what constitutes an invasive AI than societal groups that place less emphasis on privacy than on values such as safety – affecting the acceptance of surveillance technologies (Bellanova & de Goede, 2022; Kleizen et al., 2022; Kleizen & Van Dooren, 2023). Moreover, as argued by Sullivan et al. (2022), the human mind is organized on the basis of pre-existing norms, values and experiences before it encounters new experiences, allowing for quick trust evaluations when new information is presented. Accordingly, what is seen as a non-intrusive or trustworthy act by governments may sometimes be seen as threatening by citizens (Ulbricht & Yeung, 2022). Bellanova and de Goede (2022) for instance note how travelers flagged as suspicious by algorithms tend to see their flagged status as a sanction in and of itself. Such critical approaches add the insights that, in cases of value and cognitive incongruence, (some) citizens may display low trust in technologies that are perceived as safe, ethical and trustworthy by their designers (Kleizen et al., 2022).

Given the socially complex and interlinked nature of trust evaluations, we use trust and perceived trustworthiness both as outcome and explanatory variables. As outcome variables, we will examine citizens' trust in specific governmental AI projects, both as trustworthiness perceptions (i.e. perceived characteristics of the trustee) and as decision to trust (i.e. to accept vulnerability). Moreover, we examine respondents' support of AI projects (i.e. policy support), an outcome that strongly correlates with trust evaluations according to prior research (Popelier, Kleizen, Declerck, Glavina, & Van Dooren, 2021), but which sheds light on the degree to which perceived trustworthiness also translates into legitimacy to use AI. We also use trust as an explanatory variable. Based on insights from Public Administration and Science & Technology studies, we expect that attitudes towards specific AI projects are in part driven by the levels of trust in government and trust in AI that citizens possess before seeing information on a specific AI project (Kleizen & Van Dooren, 2023; Sullivan et al., 2022). Therefore, we will also take pre-existing, general perceptions on the trustworthiness of AI and government into account when attempting to explain why respondents place trust in and support specific governmental AI projects.

2.2. Ethical and trustworthy AI? Current paradigms on ethics-by-design

The first crucial question of this contribution is whether information on ethical AI measures can build trust, perceived trustworthiness and support. We therefore first turn to a brief overview of the way data science and computer science have approached trust-building and AI models. The problem of making AI ethically sound and trustworthy is not new. Discussions of explainable algorithms, for instance, reach back decades (Xu et al., 2019). Yet, the widespread introduction of machine learning algorithms in business and government has given the field a new impetus (Xu et al., 2019). For instance, the US Defense Advanced Research Projects Agency (DARPA) initiated the XAI (Explainable AI) program that stressed the need for AI to be explainable to foster the trust of users (Gunning & Aha, 2019). XAI subsequently developed into a subfield of computer science and data science, and is concerned with attempts to explain to users of a system how and why a certain prediction was made (Gunning & Aha, 2019). Alongside the emergence of the XAI field, the fear of biases in AI models has resulted in a range of debiasing techniques in both the pre-processing and processing stages, with efforts ranging from bias-aware data-gathering to altering the weights of latent classes of groups during analysis (Ntoutsi et al., 2020).

The rise of XAI and debiasing techniques coincides with the emergence of a rich epistemological community of AI ethicists and legal scholars, studying the desirable traits of AI. Their research focuses on the impact of new legislation, such as the EU's GDPR, and ethical principles, such as the 'no-harm' principle and the retention of human autonomy. Various ethical AI guidelines have been developed across the world, such as the EU's High-Level Expert Group (HLEG) on AI Ethics Guidelines for Trustworthy AI (AI HLEG, 2019) or, more recently, UNESCO's Recommendation on the Ethics of AI (UNESCO, 2021). These guidelines combine ethical considerations, such as transparency, non-discrimination, human autonomy, and no-harm principles, with legal compliance. Developments in data science have also been incorporated into such documents. The EU's HLEG, for instance, includes both the explainability of AI and the avoidance of unfair bias as elements of their main principles (AI HLEG, 2019).

Academics and policymaker approaches to ethical and trustworthy AI strongly focus on the trustworthiness of AI models and projects (e.g., Floridi et al., 2018; Grimmelikhuijsen & Meijer, 2022; Gunning & Aha, 2019). The implicit assumption is that a trustworthy design of the project will lead to improved perceptions of trustworthiness on the part of stakeholders. For instance, Gunning and Aha (2019) state that "*The XAI program's goal is to create a suite of new or modified ML techniques that produce explainable models that, when combined with effective explanation techniques, enable end-users to understand, appropriately trust, and effectively manage the emerging generation of AI systems.*" The broader AI4People framework states that "*it is especially important that AI be explicable, as explicability is a critical tool to build public trust in, and understanding of, the technology* (Floridi et al., 2018)." Since governments increasingly develop AI based on frameworks such as the HLEG guidelines or the AI4People paradigm (see the next section) (Veale, 2020), we need to establish whether and under what circumstances trustworthy design of AI projects leads to perceived trustworthiness, trust and support among the public. Thus, the first hypothesis tested in this article is:

Hypothesis 1. *For specific governmental AI projects,* ***ethical AI measures*** *will positively affect 1) perceived trustworthiness, 2) trust, and 3) policy support.*

There are reasons to question the validity of this hypothesis. Citizens may not be able to assess the relevance of ethical AI measures accurately (Logg et al., 2019). Citizens' evaluations of AI in government are often not based on user experiences but on indirect information through the media or press releases. More specifically, we can distinguish between two groups of citizens, based on whether they directly interact with governmental AI (output). Direct users are close to the service agencies and inspectorates and may be confronted by a decision with or without a clear explanation of the AI's role in forming that decision (e.g., citizens or companies receiving a letter on a decisions that was informed by AI). These AI users can – to some degree – evaluate the AI and its output, as well as the policies in which the AI's activities are embedded. Direct users may, for instance, see that a decision was based on an AI prediction and may receive an explanation why the AI took a certain decision. With a direct stake in the decision-making process, they may be more susceptible to information on how AI works and is used. Recently, research on direct users suggests that transparency and concrete explanations of AI outcomes may improve trust (e.g., Aoki, 2020; Grimmelikhuijsen, 2023).

The second sphere is the general public. Here we find citizens who are not immediately interacting with the AI in question, but are part of a broader democratic system of holding government AI accountable (Busuioc, 2021). Our research focuses on this group, as trust dynamics for the general population remain relatively under-examined. For this second group, placing confidence in an AI system, or an organization using an AI system, must be done without specific information on that system's performance and decisions (Logg et al., 2019). Most citizens can be expected to only hear about governmental AI in the media or general government communication. Citizens with little specific performance information on AI will likely substitute that information by relying on existing values and heuristics (Bitektine, 2011; Ingrams, Kaufmann, & Jacobs, 2021; Sullivan et al., 2022). These heuristics can be powerful. Thus, when non-users are provided with limited information on the trustworthiness of an AI project, they might rely on the heuristics informed by their pre-existing values and attitudes, instead of the information provided (Sullivan et al., 2022). We will also study the evidence for the absence of an effect (a so-called "nil" instead of null finding for hypothesis 1) with equivalence tests.

2.3. Examining the role of pre-existing attitudes and traits

If the absence of an effect receives at least some support, it is interesting to consider what other factors explain (differences between) citizen attitudes towards AI projects. Citizens will read the information presented in our experimental vignettes with their pre-existing attitudes in mind. For that reason, we develop hypotheses on the impact of several pre-existing attitudes.

2.3.1. Pre-existing trust in government and trust in AI

Citizens' general perceptions of the degree to which governments and technologies are trustworthy may be antecedents that inform evaluations of specific governmental AI projects. General levels of perceived trustworthiness of the government have been shown to generate specific forms of policy support. Research into trust in the police, for instance, finds that perceived trustworthiness of the police stimulates cooperative attitudes towards specific officers or procedures (Hamm, Trinkner, & Carr, 2017; see also Murphy, 2013). Similarly, trust in government seems to positively affect vaccine uptake (Prickett & Chapple, 2021; Smith, Amlôt, Weinman, Yiend, & Rubin, 2017; Wynen et al., 2022) and – closer to current purposes – acceptance of public sector facial recognition applications (Kostka et al., 2023). Previous studies have also shown that general trust in technology also plays a critical role in the trust in and uptake of a specific application of that technology (Aoki, 2020; Kostka et al., 2023). We, therefore, expect that the perceived trustworthiness of government in general (i.e. as an attitude towards the entire government) and general levels of perceived trustworthiness of AI (i.e. of the technology as a whole) may be important predictors of trust attitudes towards in specific governmental AI projects. This leads to hypotheses 2 and 3:

Hypothesis 2. *For specific governmental AI projects, pre-existing levels of perceived* ***trustworthiness of government*** *as a whole positively affect trust in those AI projects.*

Hypothesis 3. *For specific governmental AI projects, pre-existing levels of perceived* ***trustworthiness of AI*** *positively affect trust in those AI projects.*

Here, pre-existing levels of perceived trustworthiness refer to general attitudes measured before respondents go into the experimental phase of the survey. Hence, the pre-existing levels cannot be influenced by the experimental treatment. Trust in AI projects refers to post-experimental evaluations of the presented AI projects (i.e. as a dependent variable). Note that we exclude the dependents perceived trustworthiness of specific AI projects and policy support for specific AI projects in hypotheses 2 and 3, as relating these dependents to pre-existing levels of trust in government and trust in AI may be specifically prone to exhibiting common measurement bias. Trust was measured as the decision to trust and was therefore less prone to such issues, still allowing us to test the impact of general, pre-existing attitudes towards government and AI (see methodology section).

2.3.2. Privacy concerns

Citizens' attitudes towards privacy are likely predictors of their views towards AI in government (Wirtz, Lwin, & Williams, 2007). One specific form of privacy is information privacy, which refers to whether individuals or groups can determine how intensely their information is communicated to others (Malhotra, Kim, & Agarwal, 2004). Privacy concerns, in this view, refer to the perceived fairness of data-sharing practices as well as to the degree to which someone can exercise control over data sharing (Malhotra et al., 2004). Some people are more privacy-minded than others and, therefore, more inclined to withhold data (Sundar & Marathe, 2010; Wirtz, Lwin, & Williams, 2007).

In the age of AI, privacy takes on a wider meaning. Even if people are unaware that AI is trained on (often large-scale) datasets, a limited degree of AI literacy may allow them to see that AI's could make potentially intrusive predictions (e.g. Kostka et al., 2023; Meijer & Wessels, 2019). Thus, in modern times, privacy concerns have moved beyond information into the realm of potentially intrusive predictions. Privacy concerns may therefore play a particularly important role in government projects that rely strongly on large-scale datasets and advanced data analytics techniques. Big data projects may activate a big-brother syndrome, a perception that government becomes too powerful if given too much data that can analyze and predict the behavior of citizens (Kostka et al., 2023). Therefore, we formulate hypothesis 4:

Hypothesis 4. *For specific governmental AI projects, higher self-reported* ***privacy concerns*** *negatively affect 1) perceived trustworthiness of AI projects, 2) trust in AI projects, and 3) policy support for AI projects.*

2.3.3. Perceived discrimination

We also study the role of previous experiences. Socio-cultural variables may make some individuals more likely than others to be skeptical of governmental AI projects. We focus specifically on perceived discrimination, as factors such as ethnicity have previously been established to be important predictors of attitudes towards government, including levels of trust in policing (Kääriäinen & Niemi, 2014). Similar to policing, reports of AI discrimination have appeared in the media worldwide. Facial recognition software, for instance, has performed worse among some ethnic groups (Steinacker, Meckel, Kostka, & Borth, 2020), and US judicial recidivism algorithms have reportedly been biased against ethnic minorities (Chouldechova, 2017). Given the proliferating discussions on the potential role that AI-based risk-profiling may play in areas such as supervision and enforcement (Grimmelikhuijsen, 2023; Busuioc, 2021), investigating whether such factors also determine attitudes towards governmental use of AI seems warranted. Thus, hypothesis 5 reads:

Hypothesis 5. *For specific governmental AI projects, a perception that one is a* ***member of a group that is discriminated*** *against negatively affects 1) perceived trustworthiness of AI projects 2) trust in AI projects and 3) policy support for AI projects.*

2.3.4. Familiarity with AI

Finally, knowledge of IT, data, and AI may also contribute to attitudes towards AI in government. Gefen (2000) defines familiarity in an e-commerce setting as "an understanding, often based on previous interactions, experiences, and learning of what, why, where and when others do what they do". Familiarity and affinity for technology has been argued to be a factor in predicting trust in multiple contexts, as it reduces uncertainty regarding outcomes and reliability – and, therefore, vulnerability (Gefen, 2000; Gulati & Sytch, 2008; Kostka et al., 2023). Familiarity with e-commerce websites, for instance, provides people with knowledge on how the website operates, but also the degree to which the operator is benevolent and integer (Gefen, 2000). Familiarity provides knowledge of previous interactions and experiences, which helps with trusting a trustee on the future actions (Gefen, 2000). Translated to the context of AI, we expect that familiarity with the design, uses and limits of AI models reduces uncertainty on how these models can be used by public authorities. Generally, familiarity is seen as a factor that builds trust in various technologies (Gefen, 2000; Komiak & Benbasat, 2006). One study, however, does find that familiarity with AI adversely impacted participants' willingness to accept an AI's advice (Berger, Adam, Rühr, and Benlian (2021). For now, we therefore formulate a positive hypothesis, although we acknowledge that other effects may also emerge:

Hypothesis 6. *For specific governmental AI projects, familiarity with AI positively affects 1) perceived trustworthiness of AI projects 2) trust in AI projects and 3) policy support for AI projects.*

Hypotheses 2–6 were also pre-registered.

3. Experimental design and data

We conducted two online survey experiments, with both using a between subjects design. Survey experiments allow us to test whether controlled exposure to information on a range of (combinations of) ethical AI measures affects citizen attitudes. Moreover, thanks to the (simple) random assignment of respondents, variables capturing the experimental vignettes are by construction exogenous, which strongly improves the internal validity of the study.

The first experiment contains interventions showing information on 1) legal compliance, 2) ethics-by-design information, and 3) information on how data was gathered by the governmental organizations in question. The second experiment incorporates interventions on 4) retaining a human-in-the-loop, 5) emphasizing fairness and non-discrimination, and 6) ensuring technical robustness. Both experiments used the same survey flow (see below for an explanation of each intervention). Respondents not assigned to the control group received information on at least one (hypothetical) measure taken by the Belgian federal government to increase the trustworthiness of their projects. Some respondents received combinations of two or even all three treatments. For experiment 1, the data-gathering condition had two variants: one in which the projects were committed to fully gathering data in-house and one in which the projects used data from private actors but anonymized these before use. Experiment 2 did not use variations on its three conditions. Experiment 1 has 12 possible combinations of interventions (or lack thereof), while experiment 2 has 8 (see Table 1).

3.1. Designing the experimental conditions

Designing realistic interventions is challenging. AI in government is a relatively new area of inquiry and Western European governmental AI projects are often in their starting phase. Therefore, we based the interventions for the experiments on 11 interviews and the existing scientific literature. The baseline information shown to both the control group and all intervention groups was hypothetical but informed by algorithms that currently exist or are being explored by public agencies. To avoid spurious effects induced by the different roles that algorithms may play in the public sector, the baseline information first includes information on a relatively non-invasive algorithm that detects damage to roads. Only then we introduce more controversial algorithms, with one focusing on risk profiling in tax returns and the other focusing on the prediction of citizen flows during events using mobile data. These two hypothetical AI projects were incorporated to include at least one form of fraud detection and to include an algorithm with a clear link to individuals' personal data (through mobile phones). In addition to the baseline information, each intervention group adds segments of information on ethical AI measures taken to enhance the trustworthiness of AI used by the federal government.

An explanation of the interviews, a pseudonymized list of respondents as well as the full vignettes, including baseline and interventions, is available in Appendix 1.

Table 1
Regression results experiment 1, robust standard errors in parentheses (* = 0.10, ** = 0.05*; *** = 0.01; **** = 0.001).

	Policy support (OLS)		Perceived trustworthiness of AI Project (OLS)		Behavioral trust (willing to provide data) (logistic)	
Experimental groups (control group is reference category)						
Legal	0.555	0.566	−3.282	−3.735	−0.165	−0.219
	(2.69)	(2.61)	(2.82)	(2.78)	(0.30)	(0.38)
Ethics-by-design	6.563**	5.589**	1.737	−0.235	−0.221	−0.493
	(2.64)	(2.62)	(2.80)	(2.86)	(0.30)	(0.35)
Data-sharing 1 (exclusively by public sector)	4.457*	3.769	2.253	0.346	0.385	0.111
	(2.48)	(2.42)	(2.82)	(2.58)	(0.32)	(0.39)
Data-sharing 2 (data is shared by private sector, but anonymized)	2.757	2.465	2.270	2.140	0.144	−0.091
	(2.74)	(2.75)	(2.81)	(2.99)	(0.31)	(0.38)
Legal & Ethics-by-design	3.123	3.027	1.702	1.365	0.329	0.158
	(2.63)	(2.68)	(2.81)	(2.95)	(0.32)	(0.42)
Legal & data-sharing 1	0.314	−1.410	−3.470	−5.926**	0.233	0.053
	(2.62)	(2.62)	(2.81)	(2.86)	(0.31)	(0.38)
Legal & data-sharing 2	4.408*	3.027	1.086	−1.310	0.092	−0.237
	(2.58)	(2.61)	(2.84)	(2.79)	(0.31)	(0.38)
Ethics-by-design & data-sharing 1	3.542	2.699	−0.573	−1.333	−0.135	−0.453
	(2.66)	(2.60)	(2.83)	(2.88)	(0.30)	(0.38)
Ethics-by-design AI & data-sharing 2	4.202	2.672	3.021	1.343	−0.059	−0.362
	(2.63)	(2.54)	(2.81)	(2.58)	(0.30)	(0.38)
Legal, Ethics-by-design & data-sharing 1	3.248	2.632	3.413	1.951	−0.026	−0.418
	(2.60)	(2.66)	(2.82)	(2.61)	(0.31)	(0.38)
Legal, Ethics-by-design & data-sharing 2	1.470	1.543	0.453	−1.114	−0.191	−0.556
	(2.71)	(2.77)	(2.82)	(2.86)	(0.30)	(0.37)
Privacy concern		−4.866****		−5.607****		−0.781****
		(0.65)		(0.74)		(0.10)
Perceived discrimination		−2.058****		−2.428****		0.028
		(0.57)		(0.68)		(0.09)
Professional use of AI		2.023**		2.637***		−0.096
		(0.85)		(0.95)		(0.13)
Perceived trustworthiness of government						0.181***
						(0.07)
Perceived trustworthiness of AI						0.702****
						(0.10)
Control variables	Excluded	Included	Excluded	Included	Excluded	Included
Constant	52.955****	59.273****	55.072****	63.868****	0.943****	−0.895
	(1.98)	(3.93)	(2.00)	(4.42)	(0.22)	(0.63)
Observations	1243	1100	1251	1107	1269	1117
F-test (OLS)/Chi2-test (logit)	1.240	6.253****	1.280	6.080****	9.200	178.270****
R2 (OLS)/McFadden's Pseudo-R2 (logit)	0.0112	0.147	0.0112	0.151	0.006	0.1927
Adjusted R2 (OLS)/n.a. (logit)	0.002	0.126	0.002	0.131	n.a.	n.a.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$

3.1.1. Experiment 1, condition 1: Legal information

Multiple respondents in public organizations (including all respondents from civil society organizations, as well as respondents from Belgian and Dutch entities in taxation, criminal law, transport, and social security) noted the relevance of strict implementation of the GDPR and other legal principles. This includes a strong Data Protection Officer (DPO) who actively safeguards GDPR principles that are aimed at maintaining societal trust, such as the requirement to have a legal basis, to limit the purposes for data processing, and to maintain a human-in-the-loop (Busuioc, 2021; Forcier, Gallois, Mullan, & Joly, 2019). One organization emphasized the importance of a strong legal base for AI programs, as this enhances the democratic legitimation of government action. Therefore, in the first condition, the government actively communicates about GDPR-compliance by ensuring a strong DPO function and having a legal basis for a project.

3.1.2. Experiment 1, condition 2: ethics-by-design

Many public organizations have invested in the ethical design of AI, stimulated by controversies (such as the Dutch SyRI case (see: Meuwese, 2020)) and the Ethics Guidelines of the EU. This implementation process frequently emphasizes that the design process of algorithms itself should already incorporate considerations on ethics, an approach known as ethics-by-design. Our vignette draws on two aspects highlighted by multiple interview respondents: ensuring algorithmic explainability and using AI for societal well-being. Model explainability was one of the most frequently cited methods to integrate ethics-by-design during the interviews and is also frequently discussed in the literature (Busuioc, 2021; Grimmelikhuijsen, 2023). The second was incorporated as it was a far-reaching yet realistic form of implementing ethical AI (reported by one interviewed organization), which may form the most likely way of triggering a trust response among citizens (see also Floridi et al., 2018).

3.1.3. Experiment 1, condition 3: Data-gathering

For the final intervention of experiment 1, we incorporate how public organizations gather data (see also Meuwese, 2020). In the interviews, data-sharing was seen as a particularly large risk by multiple respondents. The open formulation of the purpose limitation principle in the GDPR may be a particular problem in current legislation, in particular when there is a legal basis for sharing data beyond individual consent (for an academic treatment, see: Jasserand, 2018). We incorporate two variations of this condition, reflecting the multiple avenues public organizations may take regarding data-gathering strategies. In the first variant, we present information that the public organization will fully gather data in-house to safeguard privacy. In the second variant, information is given that the public organization uses data from private organizations, although this information is anonymized.

3.1.4. Experiment 2, condition 1: Human-in-the-loop

Interviews showed that several government organizations focus strongly on retaining human autonomy in some way, often by keeping a human in the loop. Research has also emphasized the importance of keeping humans in the loop to strengthen trust in AI (Aoki, 2021; Ingrams, Kaufmann, & Jacobs, 2021). The EU's high-level expert group on AI has incorporated human-in-the-loop as a component of its trustworthy AI guidelines. Therefore, the first intervention included in experiment 2 is information on whether there is a human in the loop during the decision-making stage of public service delivery.

3.1.5. Experiment 2, condition 2: Fairness & non-discrimination

Another theme frequently referred to in interviews with public sector actors is the prevention of unintentional biases that may creep into AI models that are trained on poor data. Such ethical AI measures have also received extensive attention in data science, resulting in design processes to make AI fairer and less biased (Greene et al., 2019). As with retaining humans in the loop, reducing bias has also been presented as a core component of the EU's High-level expert group Trustworthy AI guidelines. We incorporate fairness and non-discrimination through organizational processes as a second intervention in the second experiment.

3.1.6. Experiment 2, condition 3: Technical robustness

Finally, the technical robustness of AI systems (in particular in terms of data security) received attention in interviews with public sector actors and is included in the EU's High-level expert group guidelines on Trustworthy AI. Cybersecurity procedures should convince citizens that sensitive personal data will not fall into the wrong hands. Therefore, we create a third intervention on blockchain technology that contributes to technical data security. This intervention is inspired by the Estonian e-health blockchain systems, which register all data processing activities on Estonian health records in a tamper-proof way (Nortal, 2018).

3.2. Data-gathering procedure

Respondents for experiment 1 were retrieved from a Belgian market research company with a representative sample of Belgian citizens, based on gender and age. Recruitment continued until the desired sample size of 1200 completed surveys was attained (100 per experimental group). This process yielded 1269 observations, although some models draw on slightly fewer observations due to item non-response.

Experiment 2 made use of the University of Antwerp's *Burgerpanel* (literally translated as Citizen panel). The online survey was sent to 2.000 respondents, with 738 usable responses being returned (slightly fewer than the 100 per group we aimed for; and a response rate of about 37%). *Burgerpanel* recruitment is based on self-selection into the panel. Respondents who signed up were expected to participate in studies on political and public administration topics. As a result, the *Burgerpanel* includes respondents with some degree of interest in politics. The sample is therefore less representative of Flemish citizens than the one of the market research company. However, simple random assignment to the different groups (used in both experiments) should alleviate this issue in terms of endogeneity. While some generalizability issues may remain, the similarity of results (see section 4) suggests that differences between panels did not play a major role.

Budget restrictions for experiment 1 implied that we could not go beyond the ~1200 respondents here, while availability of slots in the *Burgerpanel* limited that study to 2000 respondents with around a 40% response rate. By fielding two experiments instead of one, we could test a greater array of possible ethical AI measures and combinations thereof.

Both surveys were designed and implemented in Qualtrics. Once respondents reached the vignettes comprising the experimental stage of the study (i.e., after preliminary survey questions were filled in), Qualtrics was used to automatically distribute respondents across groups using simple random assignment (i.e. each respondent had an equal chance to be assigned to any of the experimental or control groups).

3.3. Measurements

3.3.1. Dependent variables

Both experiments contain three dependent variables on the project level: perceived trustworthiness, behavioral trust, and policy support. Perceived trustworthiness was measured using a three-item version of the trust-in-government-scale developed and validated by Grimmelikhuijsen and Knies (2017), applied to AI projects. This scale assesses perceived trustworthiness based on the distinction between ability, benevolence and integrity. Behavioral trust was measured using a single item in which respondents indicated whether they would be willing to provide permission to the federal government to use their data. Respondents received three answer categories, 'yes', 'no', and 'only under certain conditions'. To facilitate analysis, these were recoded into a binary variable 'willing to give data' measuring whether a respondent indicated 'no' (=0), 'yes', *or* 'only under certain conditions' (=1). By measuring whether citizens are willing to provide data, we come closer to a measurement of willingness to be vulnerable. Policy support was measured through two items measuring whether the respondent believes that AI projects will enhance service delivery and whether the respondent supports the use of AI within governmental projects.

3.3.2. *Independent variables*

In addition to the variables reflecting the experimental manipulations, we measure several independent variables. Descriptive statistics are available in Appendix 2, while a full description of survey items is available in the pre-registration (see Appendix 4). Due to space constraints, we limit the discussion here to the variables operationalizing concepts contained in the hypotheses. General perceived trustworthiness of government is measured using the 9-item scale on perceived trustworthiness by Hamm et al. (2019), adapted slightly by incorporating two items from Grimmelikhuijsen and Knies (2017) and by excluding the items on trust (due to the Dutch translation being considered difficult to read by pre-test respondents). Factor analyses strongly support a single-factor interpretation. Trust in AI is measured using a scale adapted from trust in self-driving cars. The measure has functionality, helpfulness & reliability dimensions (although, once again, factor analysis support interpreting the concept as a single factor) (Nees, 2016). Privacy concerns are measured using a 3-item scale developed for this survey and oriented towards governments specifically. The items respectively reflect concerns regarding the processing of personal data, the processing of data from employers or previous employers, and the processing of an individual's prior interactions with governmental organizations (e.g., halted welfare, debts, fines, etc.). All items strongly load on a single factor. Perceived discrimination is measured using a single Likert-scale item on the degree to which respondents consider themselves a member of a group that is discriminated against in society, adapted from the European Social Survey (ESS, 2019). Finally, we measure familiarity with AI by taking a proxy that asks whether participants are involved to some degree in the application or development of AI. Such a measurement does not only capture statistics or programming skills relevant to AI and major datasets, but also captures familiarity more broadly in the sense of respondents having professional experience building datasets, using AI output, managing AI/data projects, etc.

Our experimental interventions are exogenous by design, given the random assignment of participants over groups. The same does not hold for the relations between other independent variables and the dependents. Therefore, as with most cross-sectional surveys, we aim to reduce endogeneity from model misspecification by controlling for potentially relevant characteristics and attitudes. The models with independent variables beyond the experimental interventions also incorporate education, age, gender, generalized trust, computer self-efficacy, and whether the respondent is employed in the public sector.

3.3.3. *Open answers*

Respondents received a behavioral trust item asking them to indicate whether they would be permitted by the federal government to use their data. Immediately after this item, respondents were also provided with a small open answer box asking them to explain their answers. We received 777 useful answers for experiment 1 and 569 useful answers for experiment 2. These answers were subsequently used to triangulate quantitative findings.

4. Analysis and results

The analysis section is split into several subsections. First, we test the experimental conditions in sections 4.1–4.3. Subsequently, we test the role of pre-existing attitudes in section 4.4 and triangulate findings using qualitative data from the comment box in section 4.5.

4.1. *Traditional inference tests on experimental manipulations*

The regression models (Table 1 and Table 2 for respectively experiments 1 and 2) show that most experimental groups do not differ significantly from the control group. In experiment 1, we obtain null findings on all dependent variables for the groups 'Legal information', 'Data-gathering information 1', 'Data-gathering information 2', and all combinations of interventions save 'Legal information & Data-gathering information 2'. In experiment 2, we obtain null findings for all outcome variables for the experimental groups 'Human-in-the-loop information' and 'Human-in-the-loop & Fairness information'. The only group that seems to have consistently significant results on outcomes is the 'Fairness and robustness information' group. Fairness and robustness information seems to slightly increase perceived trustworthiness, policy support, and behavioral trust. As can be expected, this result disappears when including independent variables for perceived trustworthiness and policy support.

Despite containing a small number of significant regressors, the models only using experimental manipulations as explanatory variables all display non-significant F-tests and Chi-square tests for the entire model, accompanied by extremely low R^2 and Pseudo-R^2 scores.[2] The absence of evidence for an effect does not seem to be caused by the lack of respondents' attention to the vignette. Robustness tests using a manipulation check (i.e., whether the respondent could correctly remember the AI projects outlined in the vignette) yield the same results (see Appendix 3). This suggests that the explanatory power of our experimental variables is exceedingly low, even when including the Fairness & Robustness groups for which some significant results were obtained. Therefore, we now proceed to equivalence testing to see whether there is evidence for the absence of effects.

4.2. *Equivalence testing on experimental manipulations*

Table 3 reports the results from traditional *t*-tests (for statistically significant differences) *and* the equivalence tests, determining the probability that the mean of the intervention group falls within an equivalence interval around the mean of the control group. Equivalence testing is needed to test whether the absence of significant findings also points towards absence of an effect (Dinno, 2017; Streiner, 2003). For equivalence testing, we need to establish which differences between control and intervention can be considered equivalent (i.e., what is an acceptable interval around the mean of the control group within which group means are considered equivalent?). To avoid arbitrary choices, we begin by testing for equivalence using an interval based on the mean ± 10% of that mean before narrowing down the interval to ±7,5 and ± 5%. As equivalence tests for binary outcomes require exceedingly high sample sizes (Dinno, 2017), the behavioral trust outcome could not be examined.

Experiment 1 displays some evidence of equivalence (i.e. nil findings). For this experiment, we find that the Ethics-by-design, Data-sharing 1, Legal & Data sharing 2, Legal & Ethics-by-design, Ethics-by-design & Data Sharing 1 and Legal, Ethics-by-design & Data sharing 2 groups are equivalent to the control group using a 10% equivalence interval (not to be confused with a confidence interval). Moreover, for policy support, we find some evidence that Legal information and Legal & Data-sharing 1 are equivalent on the 7,5% level. At the same time, Legal, Ethics-by-design & Data-sharing 2 is equivalent at the 10% interval. Legal & Data-sharing 1, while indeterminate from an equivalence perspective, yields a non-significant *negative* effect in some of the previously discussed regression models. Therefore, there seems to be relatively solid evidence that especially the legal information condition and

[2] Explained variance is so low that adjusted R^2 corrections for the amount of variables entered into the model actually causes a small negative value in some models containing only experimental groups

Table 2
Regression results experiment 2, robust standard errors in parentheses (* = 0.10, ** = 0.05*; *** = 0.01; **** = 0.001).

	Policy support - OLS		Perceived trustworthiness of AI projects – OLS		Behavioral trust (willing to give data) - logit	
Experimental groups (control group is reference category)						
Human in the loop (HITL)	4.192	−0.695	3.710	−0.046	0.227	−0.615
	(2.86)	(2.48)	(3.23)	(2.92)	(0.36)	(0.50)
HITL & Fairness	2.268	−0.711	2.246	0.451	0.310	0.299
	(2.85)	(2.46)	(3.21)	(2.83)	(0.36)	(0.50)
HITL & robustness	4.236	−0.150	6.308*	1.462	0.147	−0.521
	(3.06)	(2.69)	(3.25)	(2.95)	(0.35)	(0.44)
HITL, Fairness & Robustness	3.951	−0.947	6.576**	1.970	0.284	−0.142
	(2.98)	(2.66)	(3.27)	(2.78)	(0.36)	(0.48)
Fairness	3.718	1.724	2.859	1.155	0.697*	0.680
	(2.94)	(2.50)	(3.28)	(3.01)	(0.39)	(0.52)
Fairness & Robustness	6.699**	1.907	6.077*	−0.140	0.927**	1.007*
	(2.95)	(2.56)	(3.23)	(2.87)	(0.41)	(0.57)
Robustness	4.572	2.808	3.459	1.373	0.336	−0.347
	(3.23)	(2.74)	(3.23)	(2.95)	(0.36)	(0.46)
Privacy concern		−7.769****		−8.500****		−0.961****
		(0.73)		(0.92)		(0.16)
Perceived discrimination		−2.757****		−3.944****		−0.168
		(0.78)		(0.85)		(0.14)
Professional use of AI		2.070**		0.122		−0.325
		(0.93)		(1.17)		(0.20)
Perceived trustworthiness of government						0.253**
						(0.12)
Perceived trustworthiness of AI						0.845****
						(0.16)
Control variables	Excluded	Included	Excluded	Included	Excluded	Included
Constant	58.413****	73.437****	51.873****	62.526****	1.213****	−0.211
	(2.18)	(4.67)	(2.27)	(5.69)	(0.24)	(1.16)
Observations	738	659	737	655	756	671
F-test (OLS)/Chi^2-test (logit)	0.863	10.825	0.999	9.887	7.110	105.700****
R2 (OLS)/McFadden's pseudo-R2 (logit)	0.008	0.288	0.009	0.284	0.011	0.308
Adjusted R2 (OLS)/n.a. (logit)	−0.001	0.264	−0.000	0.260	n.a.	n.a.
* p < 0.10, ** p < 0.05, *** p < 0.01, **** p < 0.001						

its combinations with other manipulations are, at best, equivalent to the control group on policy information and perceived trustworthiness.

For experiment 2, results were less clear, with only the Human-in-the-loop condition obtaining equivalence on the policy support outcome. Most groups in experiment 2 thus yield a null finding, yielding no evidence of significant differences but also no statistical support for equivalence. Future studies with a higher sample size may be necessary to investigate whether the results are truly equivalent.

4.3. *Summarizing regression analysis and equivalence testing*

Legal information, in particular, seems to produce 'nill' effects compared to the control group (i.e. the true absence of an effect), suggesting that information on a legal base and strong DPO function does not influence perceived trustworthiness or policy support. Moreover, the other information types in experiment 1 provide null findings or tenuous significance findings, providing no evidence that information on no-harm & explainability (ethics-by-design) and safeguards in data-gathering (only using in-house data or anonymizing data from private suppliers) enhances perceived trustworthiness, policy support or behavioral trust (even if there is also no evidence of equivalence). Experiment 2 does provide some evidence that groups differ significantly from each other (in particular for the robustness and human-in-the-loop conditions), although effects remain limited and F-tests for overall model significance are also not significant.

We conclude that there is no robust evidence for hypothesis 1 (information on ethical AI measures will positively affect perceived trustworthiness, policy support, and behavioral trust). We also conclude that the results provide some support for equivalence (in particular for groups incorporating legal information as the only or one of their interventions), i.e. the absence of an effect. Combined, these results suggest that governments' ability to manage perceptions of trustworthiness, policy support, and behavioral trust among the general public by signaling that it has taken ethical AI measures may be rather limited, at least in the short-term.

4.4. *Results for non-experimental survey variables*

We found that actively managing policy support, behavioral trust, and public perceptions of project-level trustworthiness is difficult. We now turn to the hypotheses suggesting that citizen attitudes may be formed mainly by pre-existing attitudes, traits, and perceptions (for versions with control variables, see Appendix). We see in column 6 of Tables 1 and 2 that both the perceived trustworthiness of AI and the perceived trustworthiness of government, in general, have a strongly significant and positive effect on the willingness of citizens to make their data available for governments' analyses, i.e., their behavioral trust, thus supporting hypotheses 2 and 3. This suggests that individuals who, prior to seeing the vignettes, perceive governments and AI as relatively trustworthy, are also more likely to entrust their personal data to governmental AI projects. Common method bias is less of an issue for the behavioral trust dependent, as this variable is not construed from a Likert scale item (see, e.g., Jakobsen & Jensen, 2015).

Furthermore, our analyses consistently show a strongly significant negative effect of privacy concerns on the perceived trustworthiness of AI projects, policy support, and behavioral trust (columns 2, 4, and 6 of Tables 1 and 2). This indicates that individuals worried about how their data is used by governments are less inclined to be supportive of public sector AI projects, thereby supporting hypothesis 4. The fat tails of the distributions of the explanatory variables point to rather negative attitudes of citizens (Fig. 1). This distribution implies that relatively many participants perceive AI or the federal government to be untrustworthy and that relatively many participants were concerned regarding the way governments use their data. Given that our experimental manipulations

Table 3
Results equivalence tests experiment 1 & 2 (Δ denotes the value of the mean taken to construct the equivalence interval).

Intervention	Perceived trustworthiness of AI projects	Policy support
Experiment 1		
Legal	Indeterminate (Δ = 10%; p1 = 0.1797; p2 = 0.0010)	**Equivalence** (Δ = 7,5%; p1 = 0.0576; p2 = 0.0577)
Ethics-by-design	**Equivalence** (Δ = 10%; p1 = 0.0010; p2 = 0.0950)	Sig difference (Δ = 10%; p1 = 0.0000; p2 = 0.6118)
Data sharing 1	**Equivalence** (Δ = 10%; p1 = 0.0019; p2 = 0.0437)	Sig difference (Δ = 10%; p1 = 0.0000; p2 = 0.2835)
Data sharing 2	Indeterminate (Δ = 10%; p1 = 0.0032; p2 = 0.1163)	Indeterminate (Δ = 10%; p1 = 0.0016; p2 = 0.1239)
Legal & Data sharing 1	Indeterminate (Δ = 10%; p1 = 0.3038; p2 = 0.0002)	**Equivalence** (Δ = 7,5%; p1 = 0.0396; p2 = 0.0681)
Legal & Data sharing 2	**Equivalence** (Δ = 10%; p1 = 0.0068; p2 = 0.0599)	Indeterminate (Δ = 10%; p1 = 0.0001; p2 = 0.2944)
Legal & Ethics-by-design	**Equivalence** (Δ = 10%; p1 = 0.0134; p2 = 0.0452)	Indeterminate (Δ = 10%; p1 = 0.0.0005; p2 = 0.1685)
Ethics-by-design & Data sharing 1	**Equivalence** (Δ = 10%; p1 = 0.0476; p2 = 0.0112)	Indeterminate (Δ = 10%; p1 = 0.0005; p2 = 0.1807)
Ethics-by-design & Data sharing 2	Indeterminate (Δ = 10%; p1 = 0.0010; p2 = 0.1349)	Indeterminate (Δ = 10%; p1 = 0.0001; p2 = 0.2886)
Legal, Ethics-by-design & Data sharing 1	Indeterminate (Δ = 10%; p1 = 0.0007; p2 = 0.1413)	Indeterminate (Δ = 10%; p1 = 0.004; p2 = 0.1791)
Legal, Ethics-by-design & Data sharing 2	**Equivalence** (Δ = 7,5%; p1 = 0.0482; p2 = 0.0818)	**Equivalence** (Δ = 10%; p1 = 0.0049; p2 = 0.0551)
Experiment 2		
HITL	Indeterminate (Δ = 10%; p1 = 0.0029; p2 = 0.2854)	Indeterminate (Δ = 10%; p1 = 0.0003; p2 = 0.2469)
HITL & Fairness	Indeterminate (Δ = 10%; p1 = 0.0086; p2 = 0.1749)	**Equivalence** (Δ = 10%; p1 = 0.0025; p2 = 0.0827)
HITL & Robustness	Sig difference (Δ = 10%; p1 = 0.003; p2 = 0.6086)	Indeterminate (Δ = 10%; p1 = 0.0007; p2 = 0.2527)
HITL, fairness & robustness	Sig difference (Δ = 10%; p1 = 0.0001; p2 = 0.6217)	Indeterminate (Δ = 10%; p1 = 0.0007; p2 = 0.2152)
Fairness	Indeterminate (Δ = 10%; p1 = 0.0048; p2 = 0.2641)	Indeterminate (Δ = 10%; p1 = 0.0007; p2 = 0.1990)
Fairness & Robustness	Sig difference (Δ = 10%; p1 = 0.0005; p2 = 0.5941)	Sig difference (Δ = 10%; p1 = 0.0000; p2 = 0.5375)
Robustness	Indeterminate (Δ = 10%; p1 = 0.0071; p2 = 0.2695)	Sig difference (Δ = 10%; p1 = 0.0016; p2 = 0.2301)

suggest that perceived trustworthiness and support are difficult to 'win' by providing information on a governmental project, it may be that using AI for sensitive public tasks will face enduring legitimacy issues among some societal sub-groups.

Finally, we consider hypotheses 5 and 6 on the impact of perceived discrimination and professional contact with AI. Support for both hypotheses is mixed. Perceived discrimination appears to have a negative effect on perceived trustworthiness and policy support in both experiments, as shown in Tables 1 and 2, columns 2 and 4. However, there is no evidence of an effect on behavioral trust, i.e., willingness to provide data. It may therefore be that the perceived trustworthiness of AI or perceived trustworthiness of government mediate the effect between discrimination and all three outcome variables, although testing such mediation is beyond our data's limitations. Professionally encountering AI systems' development or implementation has a positive effect on policy support for experiments 1 and 2 and on the perceived trustworthiness of governmental AI projects in experiment 1. However, there seems to be no significant effect on willingness to provide data for both experiments and no significant effect on perceived trustworthiness for experiment 2.

4.5. *Triangulation through the open answers*

Our quantitative conclusions receive further support from a qualitative analysis of the open answers. Through an open answer box, respondents could explain why they would accept or refuse to provide their data to governmental AI projects. The answers allowed us to explore the drivers of trust beyond the survey items. All comments were sorted using open and axial coding. The results for the 15 largest categories of codes can be found in the table incorporated in Appendix 5. Note that the 13 out of 15 of the most popular codes overlap in both experiments, suggesting that we tapped into attitudes in a relatively reliable fashion across experiments.

Particularly relevant for our purposes is that answers seemed mostly based on the pre-existing attitudes privacy concerns (26% and 17% if comments for respectively experiments 1 and 2), trust in government (12–14% of comments), and more general descriptions of trust (~5% in both experiments), rather than responding to information provided in experimental treatments. Indeed, hardly any answers seemed to tie in directly to the vignettes. Only eight responses in experiment 1 and ten responses in experiment 2 recognizably referred to the vignettes. As emerged from the quantitative analyses, variations in responses are thus tied to prior perceptions instead of the ethical AI measures presented to citizens through the vignettes. Prior attitudes and perceptions (in particular privacy concerns and trust in government) are the most important determinants of trust in and support for a specific AI project in government.

Beyond privacy concerns and trust in government, four other categories were reflected in at least 5% of respondent comments of both experiments: informing citizens, retaining (some degree of) control over one's data, trust in AI, and the added value of AI in terms of the effectiveness of government policies. Typical responses regarding informing citizens and retaining control reflected the need for increased transparency and obtaining consent to use personal data. The relatively high frequency of the information and control retention categories was somewhat surprising. The vignettes that explicitly tested whether respondents would exhibit higher levels of trust when more information on AI projects and their safeguards is presented found no effects. Trust in AI is mentioned less frequently than issues relating to trust in government, tentatively suggesting that the latter may be more important for attitude-formation among citizens. Effectiveness was normally related to a positive expectation of the contribution of AI to government services, with many of these comments showing that there are ardent supporters of using AI in government as well.

5. Discussion & conclusion

We tested whether signaling ethical AI measures taken to enhance the trustworthiness in governmental AI projects increase perceived trustworthiness, policy support, and trust. We also tested whether such attitudes are better explained by differences between pre-existing perceptions and characteristics of citizens. We innovatively tested these factors among the general citizenry, which remains an under-investigated context (see Gesk & Leyer, 2022; Ingrams, Kaufmann, & Jacobs, 2021 and Aoki, 2021, for recent exceptions). This focus on (differences between) members of the general audience is a major strength of this study in an environment currently dominated by studies focusing on direct users of AI (e.g. Logg et al., 2019; Grimmelikhuijsen, 2023), and our research contributes by theorizing how ethical AI may affect non-users differently due to their distance to the algorithm. The results suggest that – at least in the short-term - it is difficult to influence the perceived trustworthiness of AI projects, policy support for AI projects, and behavioral trust in the form of citizens being willing to provide their data to governmental AI projects. Pre-existing attitudes and perceptions seem to be more important than providing information on trustworthiness. General perceived trustworthiness of government and generally perceived trustworthiness of AI have a positive effects. Privacy concerns had a negative effect on the evaluation of the AI project. Thus, at least in the short-term, pre-existing attitudes seem to be used to heuristically evaluate the trustworthiness of AI in government, limiting

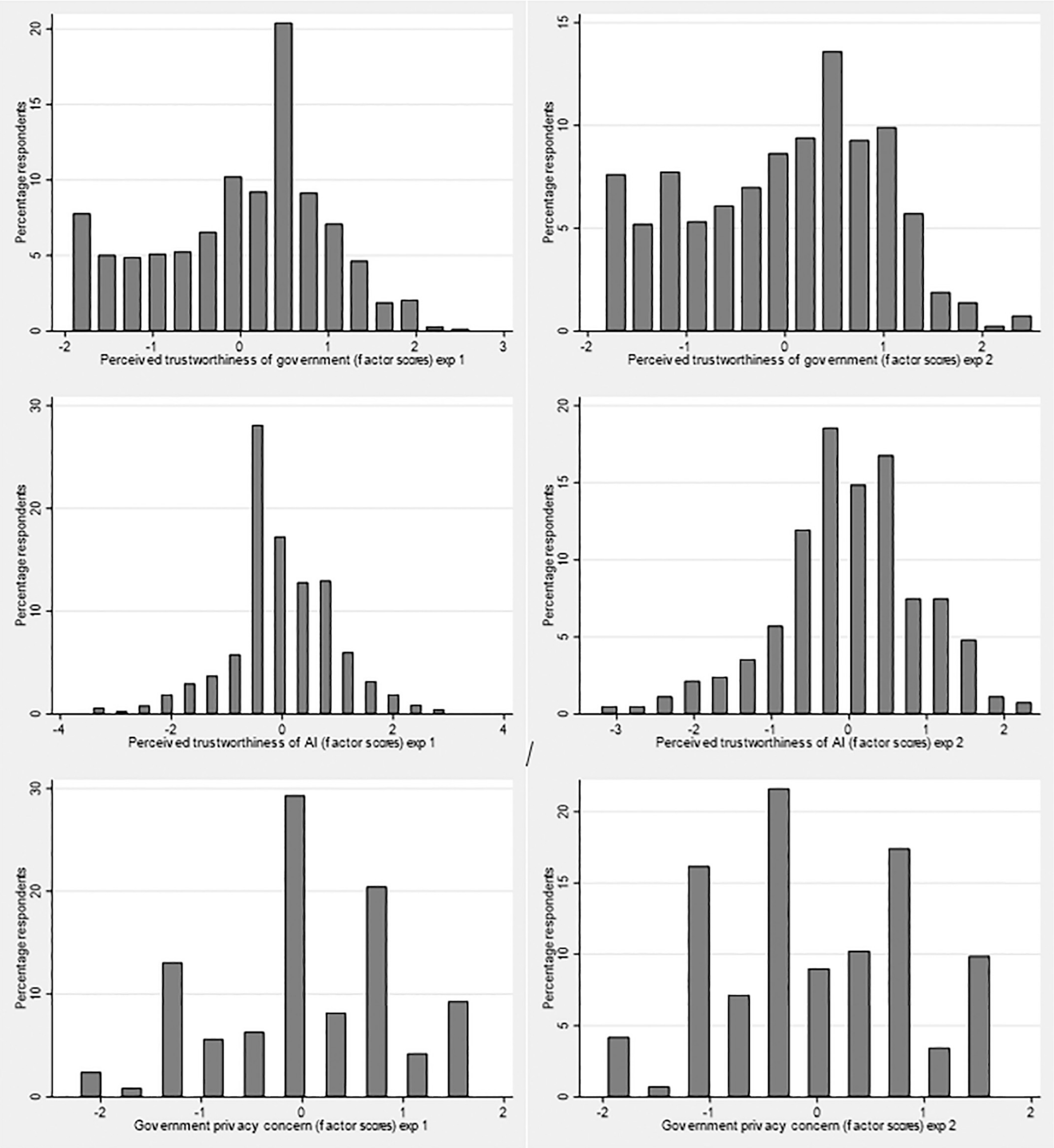

Fig. 1. Histograms for perceived trustworthiness of government, perceived trustworthiness of AI & privacy concern.

the impact of information on ethical AI among non-user citizens.

Our findings should warrant some modesty on the degree to which governments can quickly strengthen trust in their AI projects among citizens by signaling that it has taken ethical AI measures. The difficulty of managing pre-existing perceptions is probably underestimated in the Public Administration, data science, and AI ethics literature, as well as policy frameworks on trustworthy or ethical AI (e.g., the EU HLEG's trustworthy AI guidelines (AI HLEG, 2019)). Contributions and policy reports in these areas continue to point to AI design and project design as the most crucial factors in determining trustworthiness, assuming that actual trustworthiness will translate to perceived trustworthiness, trust, and – in turn – policy support among the general audience (e.g., Gunning & Aha, 2019; Floridi et al., 2018; Grimmelikhuijsen, 2023). In the short-term, policymakers and data scientists developing governmental AI projects may not be able to 'win' a major portion of the public's trust through signaling it has taken ethical AI measures. Instead, trust is likely formed on longer time-scales, implying that existing negative citizen attitudes should be seen as a strategic risk for their projects.

A focus on strategic risks could help in preventing trust breaches. As our contribution suggests that trust and support is difficult to build in the short-term, practitioners should instead shift to a long-term perspective. It is probably not enough to implement ethical AI with a checkbox mentality and on a project-by-project basis, as information ethical AI will not immediately change citizen attitudes towards potentially invasive projects. Instead, ethical AI procedures should be rolled out organization-wide or perhaps even on an entire governmental level, and with a strong focus on thoughtful, thorough and consistent application (see e.g. Busuioc, 2021). If a government's ethical AI measures 1) gradually allow society to build positive experiences with AI and 2) prevent major trust breaches, this may allow for trust formation in the long-term. Thus, the value of ethical AI does not lie in its immediate impact on citizen attitudes, but in its potential as a long-term safeguard.

Our research focuses on the general population. Antecedents for this group may differ from citizens who directly interact with the AI they are evaluating. Notably, Grimmelikhuijsen (2023) finds that ensuring explainability yields a positive effect on trust when respondents are confronted with a visa or welfare fraud decision addressed against them, while we do not find such an effect for the general population. Instead, differences in pre-existing attitudes and citizen characteristics provide a good explanation of citizen attitudes towards specific AI projects (more in line with Kostka et al., 2023). These findings do not oppose one another, as building trust through transparency may be easier when the stakes for an individual citizen are higher (Logg et al., 2019; Kleizen & Van Dooren, 2023). Therefore, future studies examining both immediate users and the general populace seem important. For practitioners, it implies focusing extensively on differences between sub-groups of citizens, and whether AI applications may provide legitimacy risks among those groups that already display low trust in government (Kleizen & Van Dooren, 2023).

Some important limitations remain. First, even though we obtained around 90 respondents per group for both experiments, the indeterminate results for the equivalence tests in experiment 2 suggest that larger group sizes would have been desirable (Dinno, 2017). Future studies can draw on effect sizes obtained here to perform power analyses, which was not yet possible for our study (power analysis requires comparable previous experiments to provide an idea of effect size). Second, while the experimental manipulations are exogenous by construction due to randomization, endogeneity remains an issue for all other survey variables related to the dependent variable. A panel data-based study would improve our ability to make causal statements. Third, while we studied a number of differences between citizens in terms of attitudes and personal characteristics, factors such as socio-economic status were not taken into account and should be tested in the future. Fourth, although vignettes were based on interviews to improve realism, the survey experiment as a method remains prone to questions regarding alternate operationalizations of conditions and external validity. Although we control for a number of demographic and personal characteristics, future studies could also explore whether specific sub-groups in society (e.g. based on differences in socio-economic status) are more receptive to information on ethical AI. Fifth, as results from two different methods of recruiting Belgian citizens support one another, we can be relatively confident of the generalizability within Belgium and similar Western European states. However, results may differ in other countries, in particular those with differing levels of trust in government (e.g. high trust societies such as the Scandinavian countries) or different paradigms on privacy and the role of the state (e.g. China (Kostka et al., 2023)). Sixth, while we attempted to include a wide variety of AI applications in our vignettes, presenting other types and applications of AI (e.g. a greater focus on intrusive surveillance, biometric identification and/or fraud detection) may still produce differing results.

Finally, while survey experiments are well-suited to detect short-term cognitive changes in assessments of trustworthiness, they are less optimal to study how information received over longer timespans can gradually change citizen attitudes and perhaps even values (Sullivan et al., 2022). Our central message is thus that ethical AI measures do not provide a silver bullet to improve trust and support on the short-term, but our analyses do not yet preclude that ethical AI may have long-term beneficial effects. If governments can facilitate the wide-spread rollout of ethical AI and thereby prevent (most) breaches of societal trust, the information space on how public authorities apply AI may gradually be capable of building trust. Whether this is true is an essential question for follow-up research.

Funding

This work was funded by Belspo (Belgian Science Policy Office), as part of the framework program BRAIN-be (Belgian Research Action Through Interdisciplinary Networks), contract nr. B2/191/P3/DIGI4FED.

CRediT authorship contribution statement

Bjorn Kleizen: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Wouter Van Dooren:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Koen Verhoest:** Conceptualization, Methodology, Writing – review & editing, Project administration, Funding acquisition. **Evrim Tan :** Conceptualization, Methodology and Funding Acquisition.

Declaration of Competing Interest

The authors have no competing interests.

Appendix 1. Interviews and vignettes

Interviews

Of the 11 public sector interviews performed, nine were held with six governmental organizations applying data analytics in Belgium and The Netherlands and two were held with external developers co-producing AI tools with a public sector organization. One public organization had a wider portfolio of data analytics projects and a larger number of departments involved than its counterparts, allowing us to perform four interviews in this organization (this is also why our sample contains a lower number of organizations than interviews). The organizations' policy domains range from tax and social security to policing, tourism, waste collection and transport. At the same time, applications of data analytics include data sharing platforms (non-AI), natural language processing of data (both on existing databases and new requests by citizens), the prediction of crowd movements and risk-score based fraud detection processes. Within the interviewed entities, interviews were usually held with respondents fulfilling either a data scientist or product owner role. However, the sample also incorporates two data protection officers (DPO) and two strategic advisers in the areas of data and AI. Most entities in the sample either applied data analytics tools such as AI for several years or were in an advanced development state (the proof of concept phase). The reasoning for this heterogeneous set of interviewees is that different organizations are likely to vary in terms of both policy goals and ethical AI measures used to secure their data analytics projects. Additionally, most entities working on advanced data analytics tools do so in moderately sized project groups, making it practically difficult to hold all interviews in one or two entities. See Table A1 for an overview.

Table A1
List of interview respondents.

Interview number	Actor	Reason for invitation
Public organizations		
1	Federal Belgian public organization in the area of finance and tax	Working on several AI projects in the proof of concept stage
2	Flemish public organization in the area of finance and taxes	Working on an AI project
3–6	Flemish public organization in the area of social protection (4 interviews held)	Organization attempts to implement AI and data-driven technologies in a widespread fashion and is working on ethics and privacy questions. Interviews held with project lead, strategic advisers, DPO and lead AI developer
7	Flemish public organization in the area of IT development and support	Developing several data-driven technologies
8	External developer, working for Belgian municipalities	Developing multiple AI systems for Belgian municipalities
9	Dutch municipality	Has implemented multiple AI systems, working on improving transparency and ethical design of AI and data-driven projects
10	Dutch public organization in the area of transport and infrastructure	Has developed multiple AI systems, working on transparency and ethical design
11	External developer, engaged in cooperative project with Dutch police	AI developer with an academic affiliation

Vignettes

Table A2 contains the baseline information and interventions that together form vignettes in experiments 1 and 2.

The full vignette combines the baseline vignette and the treatment(s) assigned to a respondent group. For instance, an experiment 1 respondent assigned to treatment groups for legal information and ethics-by-design information, as well as the internal data-gathering group, will receive the following information on a single screen:

1. Baseline information.
2. Legal information.
3. Ethics-by-design information.
4. Data-gathering information (data-gathering internally).

Table A2
Vignettes for experiments 1 and 2.

	Text vignette
Baseline information (presented to all respondents, in both experiments (including control groups))	**Artificial intelligence (AI) in federal government projects** Governmental organizations increasingly work digitally. To that end, the federal government has decided to focus on Artificial Intelligence (AI). This concerns multiple projects, including: - 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. - 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. - 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. 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 particular 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.
Ethics-by-design information (ethical)	- Governments must always ensure that the decisions of their artificial intelligence are completely explainable; - 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: - Photographic material from private construction companies (such as businesses working on roads) - Wage- and administrative data from employers - Mobile location data from telecom service providers

(*continued on next page*)

Table A2 (*continued*)

	Text vignette
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.

Citizen panels, fielding and pre-testing

Aside from containing different interventions, both survey experiments featured the same survey flow and baseline (control group) vignette. Before being fielded in citizen panels, the experimental flow was pre-tested through 5 read-out-loud interviews. Respondents read the survey and the vignettes to detect potential errors, difficult sentences, unclear elements, unexpected associations among respondents, etc. Subsequently, the first experiment (which was set out just before the second experiment) was soft-launched by the marketing company to detect whether elements may cause high attrition rates. The minor alterations made in the soft-launch process were also applied to the flow of experiment 2. These changes concerned dropping an 'are you sure you want to proceed without answering the questions' notification for an 'other' item and shortening the text on the opening screen. Both features lead to high dropout rates before implementing said changes.

Experiment 1

Experiment 1 was held online using a private marketing company panel to recruit a random sample of Flemish participants. The panel will approach its members until the goal of 1200 respondents is reached. Although recruitment into the panel is done through self-selection, selection of potential recruits will be done so that a representative sample of the Flemish population (based on age and gender and within a weighting factor of 3) is reached. All Bilendi panel members are 18 years or older. Bilendi panel members receive payment through points per survey completed, which they can subsequently spend on offers made by Bilendi. Under normal circumstances, conducting power analysis is considered desirable before fielding experiments. However, such analyses require at least some indication of expected effect sizes. As no similar experiments were found in the extant literature, no realistic indication on effect sizes was available and any assumptions on our part would thus be arbitrary. Instead of conducting a circumspect power analysis, we opted to simply set a target of 100 respondents per group (with 12 groups, this implies 12*100 = 1200 respondents). Recruitment continued until at least the target sample size was reached, with some oversampling eventually occurring. With item non-response reducing observations somewhat, the maximum available observations were 1268.

Experiment 2

Experiment 2 was fielded online through UAntwerpen *Burgerpanel*. Recruitment into the panel is done through self-selection, after which *Burgerpanel* members self-select into a specific survey by voluntarily responding to an invitation. All panel members are Belgian citizens and are 18 years or older. Panel members do not receive compensation for filling in a survey, although they receive a brief report informing them of the outcomes of their research. This may mean that responses for experiment 2 are skewed somewhat towards respondents with an interest in politics, government and/or AI, although randomization should ensure that the internal validity remains high for the experimental interventions. Two batches of 1000 citizens were approached given the panel's historical response rate of 40–50%, to obtain around 800 observations (again roughly 100 per group) (maximum amount of actual observations after item non-response is 756*)*.

Appendix 2. Descriptive statistics, variable descriptions, correlation table and full regression tables (including control variables)

Table A3
Descriptive statistics and correlations experiment 1

Variable	Obs.	Mean	Std. dev.	Min.	Max	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Policy support (1)	1243	55,846	17,984	−3513	93,318	1000													
Perceived trustworthiness of AI project (2)	1251	55,790	20,309	0,000	100,564	0,742	1000												
Behavioral trust (willingness to provide data) (3)	1269	0,724	0,447	0,000	1000	0,405	0,339	1000											
Perceived trustworthiness of government (general) (4)	1308	2605	1398	−0,138	6341	0,328	0,493	0,218	1000										
Perceived trustworthiness of AI (5)	1308	3602	0,960	0,034	6648	0,555	0,469	0,281	0,268	1000									
Privacy concern (6)	1343	2298	0,934	0,000	4045	−0,274	−0,275	−0,302	−0,260	−0,153	1000								
Perceived discrimination (7)	1237	0,869	0,927	0,000	4000	−0,130	−0,128	−0,059	−0,117	−0,050	0,043	1000							
Professional use of AI (8)	1381	0,424	0,666	0,000	2000	0,052	0,064	−0,005	0,101	0,203	−0,019	0,093	1000						
Education (9)	1417	2169	1188	1000	4000	0,083	0,038	0,060	0,093	0,146	−0,108	0,051	0,186	1000					
Age (10)	1418	2703	1604	0,000	5000	0,025	−0,006	0,063	−0,118	−0,134	0,097	−0,169	−0,337	−0,166	1000				
Gender (11)	1416	0,529	0,499	0,000	1000	−0,048	0,006	−0,002	0,070	−0,035	0,004	0,099	−0,127	0,009	−0,146	1000			

(*continued on next page*)

Table A3 (*continued*)

Variable	Obs.	Mean	Std. dev.	Min.	Max	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Generalized trust (12)	1383	2733	1486	0,000	6000	0,179	0,219	0,136	0,305	0,182	−0,179	−0,074	0,044	0,151	−0,058	−0,033	1000		
Computer self-efficacy (13)	1384	2324	1105	−0,716	4187	0,071	0,001	0,076	0,007	0,182	−0,078	0,049	0,182	0,168	−0,340	−0,183	0,118	1000	
Employed in the public sector (14)	1413	0,186	0,389	0,000	1000	0,015	0,030	0,011	0,106	0,018	−0,046	0,051	0,120	0,114	−0,249	0,048	0,036	0,081	1000

Table A4
Descriptive statistics and correlations experiment 2.

Variable	Obs.	Mean	Std. dev.	Min.	Max	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Policy support (1)	738	62,102	19,718	−0,980	94,749	1000													
Perceived trustworthiness of AI project (2)	737	55,748	22,142	0,000	100,912	0,722	1000												
Behavioral trust (willingness to provide data) (3)	756	0,825	0,380	0,000	1000	0,499	0,457	1000											
Perceived trustworthiness of government (general) (4)	786	2643	1462	−0,103	6415	0,388	0,535	0,299	1000										
Perceived trustworthiness of AI (5)	786	3830	1104	−0,067	6717	0,599	0,454	0,338	0,239	1000									
Privacy concern (6)	809	2208	1005	0,000	4048	−0,456	−0,434	−0,360	−0,373	−0,332	1000								
Perceived discrimination (7)	733	1150	0,918	0,000	3000	−0,246	−0,249	−0,163	−0,236	−0,171	0,202	1000							
Professional use of AI (8)	817	0,543	0,709	0,000	2000	0,054	−0,041	−0,049	−0,013	0,126	0,050	−0,070	1000						
Education (9)	838	2791	1154	1000	4000	0,160	0,113	0,113	0,212	0,181	−0,127	−0,137	0,146	1000					
Age (10)	839	3261	1430	0,000	5000	0,042	0,132	0,061	−0,021	−0,056	0,105	0,006	−0,253	−0,040	1000				
Gender (11)	835	0,319	0,466	0,000	1000	−0,038	0,040	0,028	0,119	−0,104	0,064	0,170	−0,014	0,025	−0,110	1000			
Generalized trust (12)	826	3036	1749	0,000	6000	0,210	0,206	0,113	0,289	0,186	−0,244	−0,154	0,114	0,191	−0,118	0,046	1000		
Computer self-efficacy (13)	822	2450	1039	−0,984	4278	0,144	0,078	0,114	0,025	0,240	−0,187	−0,132	0,138	0,104	−0,288	−0,155	0,150	1000	
Employed in the public sector (14)	837	0,154	0,361	0,000	1000	−0,025	0,003	0,002	0,090	−0,032	−0,007	0,006	0,069	0,003	−0,279	0,089	0,043	0,057	1

Table A5
Regression results experiment 1, robust standard errors in parentheses (* = 0.10, ** = 0.05*; *** = 0.01; **** = 0.001).

	Policy support (OLS)		Perceived trustworthiness of AI Project (OLS)		Behavioral trust (willing to provide data) (logistic)	
Experimental groups (control group is reference category)						
Legal	0.555	0.566	−3.282	−3.735	−0.165	−0.219
	(2.69)	(2.61)	(2.82)	(2.78)	(0.30)	(0.38)
Ethics-by-design	6.563**	5.589**	1.737	−0.235	−0.221	−0.493
	(2.64)	(2.62)	(2.80)	(2.86)	(0.30)	(0.35)
Data-sharing 1 (exclusively by public sector)	4.457*	3.769	2.253	0.346	0.385	0.111
	(2.48)	(2.42)	(2.82)	(2.58)	(0.32)	(0.39)
Data-sharing 2 (data is shared by private sector, but anonymized)	2.757	2.465	2.270	2.140	0.144	−0.091
	(2.74)	(2.75)	(2.81)	(2.99)	(0.31)	(0.38)
Legal & Ethics-by-design	3.123	3.027	1.702	1.365	0.329	0.158
	(2.63)	(2.68)	(2.81)	(2.95)	(0.32)	(0.42)
Legal & data-sharing 1	0.314	−1.410	−3.470	−5.926**	0.233	0.053
	(2.62)	(2.62)	(2.81)	(2.86)	(0.31)	(0.38)
Legal & data-sharing 2	4.408*	3.027	1.086	−1.310	0.092	−0.237
	(2.58)	(2.61)	(2.84)	(2.79)	(0.31)	(0.38)
Ethics-by-design & data-sharing 1	3.542	2.699	−0.573	−1.333	−0.135	−0.453
	(2.66)	(2.60)	(2.83)	(2.88)	(0.30)	(0.38)
Ethics-by-design & data-sharing 2	4.202	2.672	3.021	1.343	−0.059	−0.362
	(2.63)	(2.54)	(2.81)	(2.58)	(0.30)	(0.38)
Legal, Ethics-by-design & data-sharing 1	3.248	2.632	3.413	1.951	−0.026	−0.418
	(2.60)	(2.66)	(2.82)	(2.61)	(0.31)	(0.38)
Legal, Ethics-by-design & data-sharing 2	1.470	1.543	0.453	−1.114	−0.191	−0.556
	(2.71)	(2.77)	(2.82)	(2.86)	(0.30)	(0.37)
Privacy concern		−4.866****		−5.607****		−0.781****
		(0.65)		(0.74)		(0.10)
Perceived discrimination		−2.058****		−2.428****		0.028
		(0.57)		(0.68)		(0.09)
Professional use of AI		2.023**		2.637***		−0.096
		(0.85)		(0.95)		(0.13)
Perceived trustworthiness of government						0.181***
						(0.07)
Perceived trustworthiness of AI						0.702****
						(0.10)

(*continued on next page*)

Table A5 (*continued*)

	Policy support (OLS)		Perceived trustworthiness of AI Project (OLS)		Behavioral trust (willing to provide data) (logistic)	
Education ('primary or secondary education' is reference category)						
Vocational degree		−0.035		0.521		−0.037
		(2.19)		(2.63)		(0.32)
Bachelor's degree		0.100		−0.338		0.005
		(1.16)		(1.31)		(0.19)
Master's degree or higher		1.952		−1.261		0.112
		(1.51)		(1.67)		(0.23)
Age ('18–29' is reference category)						
Age 30–39		−2.544		0.234		−0.338
		(2.04)		(2.30)		(0.29)
Age 40–49		−2.361		−0.131		0.385
		(2.03)		(2.22)		(0.30)
Age 50–59		−3.513*		−3.111		0.632**
		(2.01)		(2.21)		(0.31)
Age 60–69		1.937		0.914		0.751**
		(2.10)		(2.34)		(0.33)
Age 70+		3.728*		3.007		1.229****
		(2.14)		(2.37)		(0.36)
Gender (0 = male, 1 = female)		0.192		1.570		0.221
		(1.08)		(1.23)		(0.17)
Generalized trust		1.506****		2.405****		0.041
		(0.39)		(0.43)		(0.06)
Computer self-efficacy		1.048**		−0.363		0.215***
		(0.51)		(0.58)		(0.08)
Employed in the public sector (0 = no, 1 = yes)		1.402		1.253		0.159
		(1.28)		(1.43)		(0.21)
Constant	52.955****	59.273****	55.072****	63.868****	0.943****	−0.895
	(1.98)	(3.93)	(2.00)	(4.42)	(0.22)	(0.63)
Observations	1243	1100	1251	1107	1269	1117
F-test (OLS)/Chi2-test (logit)	1.240	6.253****	1.280	6.080****	9.200	178.270****
R2 (OLS)/McFadden's Pseudo-R2 (logit)	0.0112	0.147	0.0112	0.151	0.006	0.1927
Adjusted R2 (OLS)/n.a. (logit)	0.002	0.126	0.002	0.131	n.a.	n.a.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$

Table A6

Regression results experiment 2, robust standard errors in parentheses (*=.10, **=0.05*; ***=0.01; ****=0.001)

	Policy support - OLS		Perceived trustworthiness of AI projects – OLS		Behavioral trust (willing to give data) - logit	
Experimental groups (control group is reference category)						
Human in the loop (HITL)	4.192	-0.695	3.710	-0.046	0.227	-0.615
	(2.86)	(2.48)	(3.23)	(2.92)	(0.36)	(0.50)
HITL & Fairness	2.268	-0.711	2.246	0.451	0.310	0.299
	(2.85)	(2.46)	(3.21)	(2.83)	(0.36)	(0.50)
HITL & robustness	4.236	-0.150	6.308*	1.462	0.147	-0.521
	(3.06)	(2.69)	(3.25)	(2.95)	(0.35)	(0.44)
HITL, Fairness & Robustness	3.951	-0.947	6.576**	1.970	0.284	-0.142
	(2.98)	(2.66)	(3.27)	(2.78)	(0.36)	(0.48)
Fairness	3.718	1.724	2.859	1.155	0.697*	0.680
	(2.94)	(2.50)	(3.28)	(3.01)	(0.39)	(0.52)
Fairness & Robustness	6.699**	1.907	6.077*	-0.140	0.927**	1.007*
	(2.95)	(2.56)	(3.23)	(2.87)	(0.41)	(0.57)
Robustness	4.572	2.808	3.459	1.373	0.336	-0.347
	(3.23)	(2.74)	(3.23)	(2.95)	(0.36)	(0.46)
Privacy concern		-7.769****		-8.500****		-0.961****
		(0.73)		(0.92)		(0.16)
Perceived discrimination		-2.757****		-3.944****		-0.168
		(0.78)		(0.85)		(0.14)
Professional use of AI		2.070**		0.122		-0.325
		(0.93)		(1.17)		(0.20)
Perceived trustworthiness of government						0.253**
						(0.12)
Perceived trustworthiness of AI						0.845****
						(0.16)
Education ('primary or secondary education' is reference category)						
Vocational degree		3.582		6.019		0.701
		(3.68)		(4.03)		(0.63)
Bachelor's degree		0.992		-0.644		0.400
		(1.79)		(1.99)		(0.32)
Master's degree or higher		3.628**		2.256		0.353
		(1.76)		(2.00)		(0.34)

(*continued on next page*)

Table A6 (*continued*)

	Policy support - OLS		Perceived trustworthiness of AI projects – OLS		Behavioral trust (willing to give data) - logit	
Age ('18-29' is reference category)		0.000		0.000		0.000
Age 30-39		-9.692***		-3.879		-0.624
		(3.47)		(4.21)		(0.68)
Age 40-49		-2.950		1.574		0.289
		(2.94)		(3.83)		(0.65)
Age 50-59		0.169		6.512*		0.462
		(2.76)		(3.66)		(0.60)
Age 60-69		1.150		9.745***		0.333
		(2.80)		(3.61)		(0.57)
Age 70+		3.332		12.872****		0.884
		(3.07)		(3.75)		(0.66)
Gender (0=male, 1=female)		1.054		5.329****		0.470
		(1.39)		(1.57)		(0.30)
Generalized trust		0.718*		1.046**		-0.031
		(0.42)		(0.47)		(0.08)
Computer self-efficacy		1.316*		1.073		0.140
		(0.69)		(0.74)		(0.14)
Employed in the public sector (0=no, 1=yes)		1.126		3.432		0.302
		(1.89)		(2.17)		(0.37)
Constant	58.413****	73.437****	51.873****	62.526****	1.213****	-0.211
	(2.18)	(4.67)	(2.27)	(5.69)	(0.24)	(1.16)
Observations	738	659	737	655	756	671
F-test (OLS)/Chi^2-test (logit)	.863	10.825	.999	9.887	7.110	105.700****
R2 (OLS)/McFadden's pseudo-R2 (logit)	.008	.288	.009	.284	0.011	0.308
Adjusted R2 (OLS)/n.a. (logit)	-.001	.264	-.000	.260	n.a.	n.a.

* p<0.10, ** p<0.05, *** p<0.01, **** p<0.001

Appendix 3. Analyses incorporating manipulation checks

In addition to the main analyses, we ran analyses that used the manipulation check items. Respondents were presented with 6 hypothetical AI projects, 3 of which were shown in the vignettes. Respondents were then asked to identify which of the 6 projects was shown earlier, to capture whether they could correctly recall elements of the vignettes.

As it may be possible that a respondent makes one or two errors without completely invalidating the experimental interventions, we opted not to exclude an observation when that respondent could not identify all AI projects. Instead, we weight observations based on the number of projects they have identified correctly. When all three projects are correctly ticked by the respondent, he/she is given a weight of 100%. One mistake results in the observation being weighted as 66,67%. Two mistakes reduce the weight to 33,33% and 0 correctly identified projects results in the observation being weighted as 0%.

The manipulation check was carried out well by the large majority of respondents in both surveys, suggesting that the interventions were administered effectively (see Table A2).

Table A7
Respondent scores for manipulation check.

Amount of correctly identified projects	Experiment 1	Experiment 2
3/3 projects	786	623
2/3 projects	162	84
1/3 projects	181	27
0/3 projects	120	14

Results for experiment 1 are available in Table A3, while results for experiment 2 are included in Table A4. Results do not differ substantially for analyses incorporating the manipulation check and analyses not including the manipulation check.

Table A8
Results for experiment 1 when including the manipulation check (check consists of using correctly identified AI projects to weigh observations: 3/3 correct projects are weighed 100%, 2/3 is weighed 66,7%, 1/3 is weighed 33,3%; 0/3 is weighed 0%.) Robust standard errors in parentheses.

	Policy support (OLS)		Perceived trustworthiness of AI Project (OLS)		Behavioral trust (willing to provide data) (logistic)	
Experimental groups (control group is reference category)						
Legal	−0.317	−0.290	−4.258	−4.472	−0.296	−0.439
	(3.00)	(2.79)	(3.14)	(2.95)	(0.34)	(0.44)
Ethics-by-design	6.280**	4.713*	0.259	−2.096	−0.173	−0.568
	(2.99)	(2.84)	(3.23)	(3.05)	(0.35)	(0.42)
Data-sharing 1 (exclusively by public sector)	3.210	1.857	0.105	−2.595	0.587	0.430
	(2.88)	(2.70)	(2.92)	(2.91)	(0.39)	(0.47)
Data-sharing 2 (data is shared by private sector, but anonymized)	1.252	0.235	1.240	0.160	0.243	−0.095
	(3.13)	(3.05)	(3.34)	(3.32)	(0.37)	(0.43)
Legal & Ethics-by-design	1.333	0.927	0.025	−0.451	0.227	−0.014
	(3.00)	(2.98)	(3.21)	(3.22)	(0.36)	(0.49)
Legal & data-sharing 1	−1.716	−3.649	−5.981*	−8.083***	−0.060	−0.223

(*continued on next page*)

Table A8 (*continued*)

	Policy support (OLS)		Perceived trustworthiness of AI Project (OLS)		Behavioral trust (willing to provide data) (logistic)	
	(2.98)	(2.86)	(3.19)	(3.03)	(0.35)	(0.43)
Legal & data-sharing 2	3.177	1.597	−1.476	−4.087	0.378	0.102
	(3.00)	(2.92)	(3.29)	(3.11)	(0.38)	(0.47)
Ethics-by-design & data-sharing 1	1.590	0.936	−3.717	−4.078	−0.040	−0.350
	(2.95)	(2.78)	(3.14)	(2.99)	(0.35)	(0.42)
Ethics-by-design & data-sharing 2	2.197	0.167	0.903	−1.534	0.022	−0.218
	(2.90)	(2.76)	(2.91)	(2.69)	(0.35)	(0.46)
Legal, Ethics-by-design & data-sharing 1	2.373	0.841	2.195	−0.350	−0.054	−0.373
	(2.99)	(3.01)	(2.95)	(2.80)	(0.35)	(0.44)
Legal, Ethics-by-design & data-sharing 2	−0.938	−1.186	−2.804	−4.270	−0.362	−0.730*
	(3.18)	(3.09)	(3.22)	(3.16)	(0.34)	(0.41)
Privacy concern		−5.308****		−6.174****		−0.794****
		(0.71)		(0.80)		(0.11)
Perceived discrimination		−2.237****		−2.826****		0.038
		(0.63)		(0.76)		(0.10)
Professional use of AI		2.257**		1.956*		0.031
		(1.01)		(1.15)		(0.16)
Perceived trustworthiness of government						0.207***
						(0.08)
Perceived trustworthiness of AI						0.834****
						(0.11)
Education ('primary or secondary education' is reference category)						
Vocational degree		−1.117		−1.507		0.218
		(2.67)		(3.07)		(0.39)
Bachelor's degree		−0.834		−1.675		0.099
		(1.29)		(1.45)		(0.22)
Master's degree or higher		1.424		−1.305		0.065
		(1.64)		(1.79)		(0.27)
Age ('18–29' is reference category)						
Age 30–39		−4.483**		−2.991		−0.290
		(2.25)		(2.55)		(0.33)
Age 40–49		−4.330*		−1.759		0.268
		(2.25)		(2.41)		(0.35)
Age 50–59		−4.917**		−4.453*		0.568
		(2.21)		(2.35)		(0.36)
Age 60–69		1.348		0.170		0.863**
		(2.26)		(2.49)		(0.39)
Age 70+		2.650		2.717		1.219***
		(2.26)		(2.49)		(0.41)
Gender (0 = male, 1 = female)		−0.501		1.265		0.251
		(1.19)		(1.36)		(0.19)
Generalized trust		1.050**		2.117****		0.016
		(0.44)		(0.48)		(0.07)
Computer self-efficacy		0.968*		−0.120		0.211**
		(0.58)		(0.65)		(0.10)
Employed in the public sector (0 = no, 1 = yes)		1.592		1.828		0.246
		(1.41)		(1.59)		(0.25)
Constant	55.748****	66.360****	57.245****	69.878****	1.218****	−1.214
	(2.28)	(4.02)	(2.20)	(4.61)	(0.25)	(0.78)
Observations	1115	1007	1121	1011	1128	1014
F-test (OLS)/Chi2-test (logit)	1.207	5.702****	1.300	5.585****	11.710	161.830****
R2 (OLS)/McFadden's Pseudo-R2 (logit)	0.0130	0.157	0.014	0.166	0.010	0.228
Adjusted R2	0.003	0.135	0.004	0.144	NA	NA

* p < 0.10, ** p < 0.05, *** p < 0.01, **** p < 0.001

Table A9

Results for experiment 2 when including the manipulation check (check consists of using correctly identified AI projects to weigh observations: 3/3 correct projects are weighed 100%, 2/3 is weighed 66,7%, 1/3 is weighed 33,3%; 0/3 is weighed 0%.) Robust standard errors in parentheses.

	Policy support (OLS)		Perceived trustworthiness of AI Project (OLS)		Behavioral trust (willing to provide data) (logistic)	
Experimental groups (control group is reference category)						
Human in the loop (HITL)	3.595	−0.594	3.682	−0.487	0.220	−0.637
	(2.91)	(2.49)	(3.13)	(2.85)	(0.36)	(0.52)
HITL & Fairness	2.722	−1.071	2.731	0.054	0.606	0.470
	(2.84)	(2.47)	(3.10)	(2.80)	(0.38)	(0.51)
HITL & robustness	4.132	−0.485	6.739**	1.945	0.142	−0.659
	(3.15)	(2.70)	(3.37)	(2.94)	(0.36)	(0.45)
HITL, Fairness & Robustness	5.755*	0.438	7.824**	2.617	0.553	0.091
	(2.96)	(2.62)	(3.14)	(2.78)	(0.39)	(0.50)

(*continued on next page*)

Table A9 (*continued*)

	Policy support (OLS)		Perceived trustworthiness of AI Project (OLS)		Behavioral trust (willing to provide data) (logistic)	
Fairness	3.254	0.962	2.328	0.421	0.788*	0.864
	(3.05)	(2.52)	(3.33)	(3.02)	(0.40)	(0.55)
Fairness & Robustness	6.225**	1.692	5.989*	0.049	0.936**	1.011*
	(3.00)	(2.56)	(3.32)	(2.83)	(0.42)	(0.59)
Robustness	5.667*	3.815	4.340	2.105	0.528	−0.214
	(3.27)	(2.73)	(3.45)	(2.91)	(0.38)	(0.51)
Privacy concern		−7.670****		−8.288****		−0.997****
		(0.73)		(0.94)		(0.17)
Perceived discrimination		−2.925****		−4.142****		−0.228
		(0.79)		(0.87)		(0.15)
Professional use of AI		2.281**		−0.034		−0.377*
		(0.93)		(1.22)		(0.21)
Perceived trustworthiness of government						0.219*
						(0.12)
Perceived trustworthiness of AI						0.922****
						(0.17)
Education ('primary or secondary education' is reference category)						
Vocational degree		4.442		6.255		0.331
		(3.95)		(4.46)		(0.57)
Bachelor's degree		0.609		−0.854		0.457
		(1.78)		(2.02)		(0.35)
Master's degree or higher		3.589**		2.333		0.341
		(1.77)		(2.04)		(0.36)
Age ('18–29' is reference category)						
Age 30–39		−8.986***		−5.096		−0.728
		(3.43)		(4.15)		(0.70)
Age 40–49		−1.799		0.363		0.148
		(2.96)		(3.80)		(0.68)
Age 50–59		1.825		6.563*		0.514
		(2.72)		(3.55)		(0.64)
Age 60–69		2.354		9.353***		0.280
		(2.78)		(3.59)		(0.61)
Age 70+		4.469		11.923***		0.797
		(3.05)		(3.71)		(0.70)
Gender (0 = male, 1 = female)		1.054		4.856***		0.473
		(1.38)		(1.60)		(0.32)
Generalized trust		0.665		0.883*		−0.041
		(0.41)		(0.48)		(0.08)
Computer self-efficacy		1.301*		1.166		0.044
		(0.69)		(0.75)		(0.14)
Employed in the public sector (0 = no, 1 = yes)		1.254		3.928*		0.367
		(1.81)		(2.14)		(0.38)
Constant	58.974****	72.683****	52.350****	63.549****	1.225****	0.106
	(2.23)	(4.64)	(2.25)	(5.53)	(0.25)	(1.20)
Observations	717	642	716	639	733	653
F-test (OLS)/Chi2-test (logit)	0.920	10.930****	1.273	9.538****	8.93	97.54****
R2 (OLS)/McFadden's Pseudo-R2 (logit)	0.010	0.291	0.012	0.281	0.014	0.316
Adjusted R2	0.000	0.266	0.002	0.255	NA	NA

* p < 0.10, ** p < 0.05, *** p < 0.01, **** p < 0.001

Appendix 4. Notes regarding implementation of pre-registrations

Both experiments were preregistered on the OSF preregistration platform and are available here:

Experiment 1: https://osf.io/2bnh3

Experiment 2: https://osf.io/p3e7k

The final analyses follow the preregistrations as closely as possible, although we opted to slightly alter several elements and removed one mistake from the preregistration in our submitted manuscript. Our reasoning for these alterations is outlined below.

Hypotheses

Originally, the hypotheses also referred to trust in technology as a dependent variable. This was a mistake in the original pre-registrations stemming from an older version of the text, as our dependent variables were simplified to test the trustworthiness of AI projects, behavioral trust and policy support (also evident later in the pre-registration where we discuss these three dependent variables). Reverse hypotheses were removed at the request of the reviewers.

Analyses

The analysis section of the pre-registration mentions that privacy concerns, trust in AI and general trust in government will be tested as moderators for all dependents. Tests for moderating effects (using interaction terms) were run but did not show relevant results, with the effects on dependent variables instead seeming direct. Results are available upon request. Moreover, in the final analysis, we did not regress trust in AI and general trust in government on our Likert scale dependents for policy support and trust in governmental AI projects. Correlations between these variables were relatively high, suggesting that common method bias may be a risk.

Furthermore, although we explored numerous interaction effects for other independent variables (as mentioned in the preregistrations), no significant interaction effects were uncovered.

For computer self-efficacy as a control variable, we utilize two general items that loaded well on a single latent dimension rather than the full scale (which did not load on the same underlying factor).

Finally, while the preregistration mentioned the use of Likelihood Ratio tests to determine whether model fit for a given dependent increases significantly upon introducing other variables of interest (beyond experimental interventions), we opted not to implement such tests after working with the actual data. F-test, Chi^2-test, and (pseudo)-R^2 results already show that model fit increases drastically when shifting from a specification with only experimental variables to a specification that also includes other variables of interest. The Likelihood ratio test, therefore, turns out to be redundant given the eventual data structure. At the same time, using the Likelihood Ratio test would prevent us from entering the different amounts of observations in every model (lowering power for models which would otherwise be able to use more observations).

Appendix 5. Table containing distribution of codes for qualitative triangulation

Table A10
Descriptions of 15 most prevalent codes for experiments 1 and 2.

Code	Experiment 1 (percentage)	Experiment 2 (percentage)	Description
Privacy concern	26,431	16,96	Comments reflecting the presence or absence of concerns relating to privacy
Trust in government	11,55	14,48	Comments relating to trust or trustworthiness of governments or politics (general or a subdimension of trust, such as ability, benevolence, or integrity)
Informing citizens	7492	8878	Comments suggesting that citizens should be informed about (how) the government uses data
Trust (without defining trustee)	5515	5074	Comments relating to trust, but not defining a specific trustee
Effectiveness	4683	6436	Comments related to the expectation that the use of AI or data will enhance the effectiveness or efficiency of government
Retaining control	4162	5322	Comments suggesting that the respondent would require some degree of control over their data before being willing to provide it
Security concern	3746	3713	Comments related to security concerns, such as hacking
Policy support	3642		Comments suggestive of high levels of policy support
Fear	3226	2475	Comments suggestive of fears regarding the consequences of AI
Trust in AI	2914	5074	Comments relating to some degree of trust in AI
Acquiescence	2914	2475	Comments suggesting that governments processing respondents' data is inevitable
Depends on purpose	2914	4332	Comments suggesting that willingness to provide data would depend on the purpose for which governments would use it
Uncertainty	2393		Comments denoting some degree of uncertainty on the subject matter
Progress	1873	1609	Comments suggesting that providing data is desirable to aid in the development of new technologies or services
If in citizens' interest	1561	1609	Comments suggesting that the respondent is okay with sharing data if it is in the interest of that respondent or society
Limits of the state		2104	Comments noting that the use of data and AI should adhere to the limits of the state
Long-term concern		1856	Comments suggesting concerns on how AI and data policies will develop in the future (e.g., 'sliding scale' concerns or concerns on how future actors with malevolent intent may abuse data)

For both experiments, the most frequently cited reasons for placing a certain degree of behavioral trust in AI projects are privacy concerns and trust in government. This is in line with the quantitative results, where privacy concerns and trust in government emerged as strongly significant and robust predictors of perceived project trustworthiness, policy support, and behavioral trust.[3] Although most open responses reflected either high levels of privacy concerns or low levels of trust, there were also relatively many examples of individuals reporting low privacy concerns, high levels of trust in government, or moderate levels of both categories. For privacy concerns, a large minority of responses reflect high levels of concern (experiment 1: 44,9%; experiment 2: 41,6%), and a smaller number of responses reflect moderate levels of concern (experiment 1: 34,6%; experiment 2: 29,9%) or low levels of concern (experiment 1: 20,5%; experiment 2: 28,5%). For trust in government, the large majority of comments reflected low trust (experiment 1: 66,7%; experiment 2: 64,1%), with lower amounts of comments reflecting moderate (experiment 1: 9,9%; experiment 2: 2,7%) and high levels of trust (experiment 1: 23,4%; experiment 2: 33,3%).

Typical high privacy concern responses include references to big brother ("big brother is watching" or "this is strongly approaching big brother!"), but some also combine privacy considerations with phrases that suggest a degree of mistrust in government ("privacy, the magic word government uses when it suits them"). Low trust in government responses may refer to various aspects, with dominant themes being that governments lack the competencies necessary to safely implement AI (i.e., low ability-based trust, such as "*no trust in the competencies of an unstable government")* or a lack of confidence in how governments will use citizens' data (i.e., low integrity-based trust, such as *"for 20+ years (as an external party) I have done projects for the government, and I know how they handle projects, and this makes me enormously mistrustful. The government changes specifications and rules as they go along")*. Conversely, low privacy concern responses frequently reflect the attitude that the respondent has nothing to hide from the government. An example is the following response, which combines low privacy concerns with the effectiveness of policies: *"I have nothing to hide, so if my contribution leads to a more efficient government, I will gladly go along with that"*. High trust responses relatively frequently refer to the perception that governments

[3] A notable difference between open answers provided in both experiments is, however, that privacy concern (whether that concern is low, moderate or high) is referred in over 30% of answers for experiment 1, whereas it is mentioned in roughly 17% of answers related to experiment 2.

will handle their data with integrity (i.e., integrity-based trust). A typical example here is, *"I am convinced that my information will not be used malignantly by the government."*

References

AI HLEG. (2019). Ethics Guidelines for Trustworthy AI. available at: https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html.

Alon-Barkat, S. (2020). Can government public communications elicit undue trust? Exploring the interaction between symbols and substantive information in communications. *Journal of Public Administration Research and Theory, 30*(1), 77–95.

Alon-Barkat, S., & Busuioc, M. (2023). Human–AI interactions in public sector decision making:"automation bias" and "selective adherence" to algorithmic advice. *Journal of Public Administration Research and Theory, 33*(1), 153–169.

Andrews, L. (2019). Public administration, public leadership and the construction of public value in the age of the algorithm and 'big data'. *Public Administration, 97*(2), 296–310.

Aoki, N. (2020). An experimental study of public trust in AI chatbots in the public sector. *Government Information Quarterly, 37*(4), Article 101490.

Aoki, N. (2021). The importance of the assurance that "humans are still in the decision loop" for public trust in artificial intelligence: Evidence from an online experiment. *Computers in Human Behavior, 114*, Article 106572.

Bayram, A. B., & Shields, T. (2021). WHO trusts the WHO? Heuristics and Americans' trust in the world health organization during the COVID-19 pandemic. *Social Science Quarterly, 102*(5), 2312–2330.

Bellanova, R., & de Goede, M. (2022). The algorithmic regulation of security: An infrastructural perspective. *Regulation & governance, 16*(1), 102–118.

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2021). Watch me improve—Algorithm aversion and demonstrating the ability to learn. *Business & Information Systems Engineering, 63*(1), 55–68.

Bitektine, A. (2011). Toward a theory of social judgments of organizations: The case of legitimacy, reputation, and status. *Academy of Management Review, 36*(1), 151–179.

Busuioc, M. (2021). Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review, 81*(5), 825–836.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163.

Choung, H., David, P., & Ross, A. (2022). Trust in AI and its role in the acceptance of AI technologies. *International Journal of Human Computer Interaction*, 1–13.

Dinno, A. (2017). tostt: Mean-equivalence t tests. Stata software package. Retrieved from: https://www.alexisdinno.com/stata/tost.html.

ESS. (2019). ESS9–2018 data download. retrieved on 15-11-2021 from: https://www.europeansocialsurvey.org/data/download.html?r=9.

Fischer, S. C., & Wenger, A. (2021). Artificial intelligence, forward-looking governance and the future of security. *Swiss Political Science Review, 27*(1), 170–179.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707.

Forcier, M. B., Gallois, H., Mullan, S., & Joly, Y. (2019). Integrating artificial intelligence into health care through data access: Can the GDPR act as a beacon for policymakers? *Journal of Law and the Biosciences, 6*(1), 317.

Gefen, D. (2000). E-commerce: The role of familiarity and trust. *Omega, 28*(6), 725–737.

Gesk, T. S., & Leyer, M. (2022). Artificial intelligence in public services: When and why citizens accept its usage. *Government Information Quarterly, 39*(3), Article 101704.

Greene, D., Hoffmann, A. L., & Stark, L. (2019). *Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning*.

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.

Grimmelikhuijsen, S., & Knies, E. (2017). Validating a scale for citizen trust in government organizations. *International Review of Administrative Sciences, 83*(3), 583–601.

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, 5*(3), 232–242.

Gulati, R., & Sytch, M. (2008). Does familiarity breed trust? Revisiting the antecedents of trust. *Managerial and Decision Economics, 29*(2–3), 165–190.

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine, 40*(2), 44–58.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*(1), 99–120.

Hamm, J. A., Smidt, C., & Mayer, R. C. (2019). Understanding the psychological nature and mechanisms of political trust. *PLoS One, 14*(5), Article e0215835.

Hamm, J. A., Trinkner, R., & Carr, J. D. (2017). Fair process, trust, and cooperation: Moving toward an integrated framework of police legitimacy. *Criminal Justice and Behavior, 44*(9), 1183–1212.

Ingrams, A., Kaufmann, W., & Jacobs, D. (2021). government decision making. In *, 14. Policy and Internet* (pp. 390–409). AI we trust? Citizen perceptions of AI.

Jakobsen, M., & Jensen, R. (2015). Common method bias in public management studies. *International Public Management Journal, 18*(1), 3–30.

Jasserand, C. (2018). Subsequent use of GDPR data for a law enforcement purpose: The forgotten principle purpose limitation. *European Data Protection Law Review, 4*, 152.

Kääriäinen, J., & Niemi, J. (2014). Distrust of the police in a Nordic welfare state: Victimization, discrimination, and trust in the police by Russian and Somali minorities in Helsinki. *Journal of Ethnicity in Criminal Justice, 12*(1), 4–24.

Kleizen, B., & Van Dooren, W. (2023). Is everything under control? An experimental study on how control over data influences trust in and support for major governmental data exchange projects. *Information Polity*, 1–23 (Preprint).

Kleizen, B., Van Dooren, W., & Verhoest, K. (2022). Chapter 6: Trustworthiness in an era of data analytics: What are governments dealing with and how is civil society responding?. In *The new digital era governance: How new digital technologies are shaping public governance* (pp. 563–574). Wageningen Academic Publishers.

Komiak, S. Y., & Benbasat, I. (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly*, 941–960.

Koniakou, V. (2023). From the "rush to ethics" to the "race for governance" in artificial intelligence. *Information Systems Frontiers, 25*(1), 71–102.

Kostka, G., Steinacker, L., & Meckel, M. (2023). Under big brother's watchful eye: Cross-country attitudes toward facial recognition technology. *Government Information Quarterly, 40*(1), Article 101761.

Lankton, N. K., McKnight, D. H., & Tripp, J. (2015). Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems, 16*(10), 1.

Latusek, D., & Hensel, P. G. (2022). Can they trust us? The relevance debate and the perceived trustworthiness of the management scholarly community. *Scandinavian Journal of Management, 38*(1), Article 101193.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*, 90–103.

Malhotra, N. K., Kim, S. S., & Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research, 15*(4), 336–355.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734.

Meijer, A., & Wessels, M. (2019). Predictive policing: Review of benefits and drawbacks. *International Journal of Public Administration, 42*(12), 1031–1039.

Meuwese, A. (2020). Regulating algorithmic decision-making one case at the time: A note on the Dutch'SyRI'judgment. *European Review of Digital Administration & Law, 1*(1), 209–212.

Montague, E. N., Kleiner, B. M., & Winchester, W. W., III (2009). Empirically understanding trust in medical technology. *International Journal of Industrial Ergonomics, 39*(4), 628–634.

Murphy, K. (2013). Policing at the margins: Fostering trust and cooperation among ethnic minority groups. *Journal of Policing, Intelligence and Counter Terrorism, 8*(2), 184–199.

Nees, M. A. (2016, September). Acceptance of self-driving cars: An examination of idealized versus realistic portrayals with a self-driving car acceptance scale. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 60, No. 1, pp. 1449-1453)*. Sage CA: Los Angeles, CA: SAGE Publications.

Nortal. (2018). Blockchain and healthcare: The Estonian experience. retrieved on 20-09-2021 from: https://nortal.com/blog/blockchain-healthcare-estonia.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., … Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), Article e1356.

Pétry, F., & Duval, D. (2017). When heuristics go bad: Citizens' misevaluations of campaign pledge fulfilment. *Electoral Studies, 50*, 116–127.

Popelier, P., Kleizen, B., Declerck, C., Glavina, M., & Van Dooren, W. (2021). Health crisis measures and standards for fair decision-making: A normative and empirical-based account of the interplay between science, politics and courts. *European Journal of Risk Regulation, 12*(3), 618–643.

Prickett, K. C., & Chapple, S. (2021). Trust in Government and Covid-19 vaccine hesitancy. *Policy Quarterly, 17*(3).

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review, 23*(3), 393–404.

Smith, L. E., Amlôt, R., Weinman, J., Yiend, J., & Rubin, G. J. (2017). A systematic review of factors affecting vaccine uptake in young children. *Vaccine, 35*(45), 6059–6069.

Stahl, B. C., Andreou, A., Brey, P., Hatzakis, T., Kirichenko, A., Macnish, K., … Wright, D. (2021). Artificial intelligence for human flourishing–beyond principles for machine learning. *Journal of Business Research, 124*, 374–388.

Steinacker, L., Meckel, M., Kostka, G., & Borth, D. *Facial recognition: A cross-national survey on public acceptance, privacy, and discrimination. arXiv preprint.* (2020). *arXiv:2008.07275*.

Streiner, D. L. (2003). Unicorns do exist: A tutorial on "proving" the null hypothesis. *The Canadian Journal of Psychiatry, 48*(11), 756–776.

Sullivan, Y., de Bourmont, M., & Dunaway, M. (2022). Appraisals of harms and injustice trigger an eerie feeling that decreases trust in artificial intelligence systems. *Annals of Operations Research, 308*, 525–548.

Sundar, S. S., & Marathe, S. S. (2010). Personalization versus customization: The importance of agency, privacy, and power usage. *Human Communication Research, 36*(3), 298–322.

Thomas, C. W. (1998). Maintaining and restoring public trust in government agencies and their employees. *Administration and Society, 30*(2), 166–193.

Ulbricht, L., & Yeung, K. (2022). Algorithmic regulation: A maturing concept for investigating regulation of and through algorithms. *Regulation & Governance, 16*(1), 3–22.

UNESCO. (2021). Recommendation on the ethics of artificial intelligence. https://en.unesco.org/artificial-intelligence/ethics#recommendation.

Veale, M. (2020). A critical take on the policy recommendations of the EU high-level expert group on artificial intelligence. *European Journal of Risk Regulation, 11*(1), 1–10.

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science, 46*(2), 186–204.

Winfield, A. F., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2133), 20180085.

Wirtz, J., Lwin, M. O., & Williams, J. D. (2007). Causes and consequences of consumer online privacy concern. *International Journal of Service Industry Management, 18*(4), 326–348.

Wynen, J., Op de Beeck, S., Verhoest, K., Glavina, M., Six, F., Van Damme, P., … Pepermans, K. (2022). Taking a COVID-19 vaccine or not? Do trust in government and trust in experts help us to understand vaccination intention? *Administration and Society, 54*(10), 1875–1901.

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019, October). Explainable AI: A brief survey on history, research areas, approaches and challenges. In *CCF international conference on natural language processing and Chinese computing* (pp. 563–574). Cham: Springer.

Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation & Governance, 12*(4), 505–523.

Bjorn Kleizen is a postdoctoral researcher at the University of Antwerp, and a professor in the Bachelor Social Economic Sciences. His research interests include trust in government, public management and AI in government.

Wouter Van Dooren is a professor of public administration in the Research Group Politics & Public Governance and the Antwerp Management School. His research interests include public governance, performance information, accountability and learning, and productive conflict in public participation.

Koen Verhoest is full research professor and leads the GOVTRUST Centre of Excellence and the Research group Politics and Public Governance at the University of Antwerp. He has published widely on autonomy, regulation and collaboration for public policies, and how these are conditioned by or affect trust and reputation.

Dr. Evrim Tan is a postdoctoral researcher at the KU Leuven Public Governance Institute. His research focuses on the use of digital technologies such as blockchain and AI in public governance.