



Explanations Increase Citizen Trust in Police Algorithmic Recommender Systems: Findings from Two Experimental Tests

E. N. Nieuwenhuizen, A. J. Meijer, F. J. Bex & S. G. Grimmelikhuijsen

To cite this article: E. N. Nieuwenhuizen, A. J. Meijer, F. J. Bex & S. G. Grimmelikhuijsen (2025) Explanations Increase Citizen Trust in Police Algorithmic Recommender Systems: Findings from Two Experimental Tests, *Public Performance & Management Review*, 48:3, 590-625, DOI: [10.1080/15309576.2024.2443140](https://doi.org/10.1080/15309576.2024.2443140)

To link to this article: <https://doi.org/10.1080/15309576.2024.2443140>



© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.



[View supplementary material](#)



Published online: 27 Dec 2024.



[Submit your article to this journal](#)



Article views: 1640



[View related articles](#)







[View Crossmark data](#)



Citing articles: 4 [View citing articles](#)

Explanations Increase Citizen Trust in Police Algorithmic Recommender Systems: Findings from Two Experimental Tests

E. N. Nieuwenhuizen^a , A. J. Meijer^a , F. J. Bex^{a,b}  and
S. G. Grimmelikhuisen^a 

^aUtrecht University; ^bTilburg University

ABSTRACT

A long-standing question in e-government research is how to maintain citizen trust in digital encounters with the government. This question is even more pertinent as algorithmic recommender systems (such as chatbots) are now becoming an integral part of digital public service delivery. The literature suggests that the explanations that these systems provide for their recommendations are crucial to maintaining citizen trust in digital encounters, but so far the empirical research into this relationship is limited. To test the effects of various explanations provided by algorithmic recommender systems on citizen trust, we conducted two experimental studies. We developed a mock version of an actual algorithmic recommender system used by the Dutch police and tested it in two representative survey experiments. Study 1 ($n=717$) tested the effects of *procedural*, *rationale* and *combined explanations*. We found that providing any explanation increased trust and made citizens more likely to follow an algorithmic recommendation. Study 2 ($n=1005$) investigated whether providing a *directive explanation*—specific instructions for achieving a desired service outcome—increases trust, building a more nuanced understanding of the relationship between explanations and trust in algorithmic recommendations. We conclude that explaining algorithmic recommendations—in any form—strengthens trusting beliefs, trusting intentions and trust-related behavior in citizens receiving digital public services. This may suggest that trust in algorithmic recommendations increases when citizens see that governments make an effort to provide an explanation, regardless of the nature of this explanation.


KEYWORDS

Transparency; algorithmic governance; explanations; experiment; recommender systems; police

Introduction

As more interactions between government and citizens become “digital encounters” (Lindgren et al., 2019), the question of how the digitalization of public services affects citizen trust in these services becomes more

CONTACT Esther Nieuwenhuizen  e.n.nieuwenhuizen@uu.nl  Utrecht School of Governance, Utrecht University, Bijlhouwerstraat 6, 3511 ZC Utrecht, The Netherlands.

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/15309576.2024.2443140>.
© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

urgent (Welch et al., 2004). Here, we contribute to answering this question by focusing on a rapidly emerging trend in digital public service delivery: algorithmic recommender systems such as chatbots and decision support systems that provide suggestions to citizens in various service domains. In particular, we present experimental evidence for how the explanations provided by these systems can improve citizen trust in digital services.

Algorithmic recommender systems are increasingly introduced to facilitate digital encounters and, more specifically, to support decision-making by citizens (Androutsopoulou et al., 2019; Chen & Gasco-Hernandez, 2024). They are used, for example, to support citizens when they fill out online tax forms or file a report of online fraud to the police. The literature suggests that algorithmic systems can reduce administrative burden (Fatima et al., 2020) and improve decision making (Ojo et al., 2019). However, these algorithms can also pose challenges to public organizations. A key challenge is that these systems can negatively affect citizen trust (Giest & Grimmelikhuijsen, 2020), especially when the algorithms are “black boxes” that do not provide understandable explanations of why or how a recommendation is given (Lepri et al., 2018; Rader et al., 2018). Following the literature, we expect that the transparency of the digital encounter—the interaction between citizen and algorithmic recommender system—is important for building trust. Specifically, to foster trust in algorithmic recommendations, it seems crucial that recommender systems provide citizens explanations for their recommendations (Grimmelikhuijsen, 2023).

In spite of the rapid rise of algorithmic recommender systems in government, research into trust in algorithmic recommendation is scarce, non-conclusive, and primarily focused on the private sector. The few empirical studies on recommender systems for products or holidays (Tintarev & Masthoff, 2012; Wang & Benbasat, 2007) and art, books, or movies (Cramer et al., 2008; Tintarev & Masthoff, 2012) provide contradicting results. Some studies show an increase in trust by providing explanations (e.g., Kizilcec, 2016; Nothdurft et al., 2014; Wang & Benbasat, 2007) while others show a negative or non-existent effect of explanations on trust (e.g., Cramer et al., 2008; Tintarev & Masthoff, 2012).

Only recently has some empirical evidence on trust in algorithmic recommendations in the public sector been provided. Aoki (2020) found that citizens’ trust in algorithmic recommendations depends on the area or type of service. Schiff, Schiff and Pierson (2022) found that citizens lose trust in automated decisions when there is “transparency failure,” meaning that a decision is not understood by government officials themselves. Grimmelikhuijsen (2023) distinguished accessibility and explainability as components of algorithmic transparency and showed that explaining decisions trumps accessibility in terms of generating citizen

trust. These three studies, however, do not specify what kinds of explanations are needed for citizens to trust algorithmic recommendations.

Two recent papers offer relevant, but contrasting, insights into what kind of explanations strengthen (citizen) trust in algorithmic outcomes. According to a conceptual work by de Fine Licht and de Fine Licht (2020), providing *justifications for individual algorithmic decisions* is central to generating legitimacy and trust. Kizilcec (2016), on the other hand, showed that providing information to students about *the algorithmic decision procedure* regarding their grades was most effective in strengthening trust in the algorithmic outcome. Altogether, the type of explanation that is most effective in strengthening citizen trust has only been investigated to a limited extent and remains contested.

To fill this gap, we tested the effects of different types of explanations for algorithmic recommendations in the public sector on citizen trust. We aimed to answer the following research question: *What are the effects of explanations on citizen trust in algorithmic recommendations?* Our research builds on theories about government transparency and explainable artificial intelligence (XAI) to conceptualize algorithmic explanations and on psychological and e-commerce theories of trust to understand the effects of different types of explanations on citizen trust in algorithmic recommendations. For the empirical testing of the different explanations, we use an innovative *sequential factorial design* (Sniderman, 2018).

This study contributes to the literature on algorithmic transparency (de Fine Licht & de Fine Licht, 2020; Grimmlikhuijsen, 2023) by showing that citizen trust in algorithmic recommendations increases for all explanations. This indicates that trust increases when citizens see that governments make an effort to provide an explanation, regardless of the nature of this explanation. We conclude that maintaining citizen trust does not seem to depend on the exact content of the algorithmic explanation, but rather on the fact that some sensible explanation has been provided to citizens. We reflect on what this finding means for the relationship between governments and citizens in democratic societies.

Algorithmic recommendations

Explaining algorithmic recommendations

How digitization of public services affects citizen-state interactions has long been a question of interest to public administration scholars (Dunleavy et al., 2005; Welch et al., 2004), with a recent increased interest in the adoption of new (AI) technologies in public service delivery (Lindgren et al., 2019). For example, there has been a rapid rise in the use of algorithmic recommender systems (Androutsopoulou et al., 2019). Many of these recommender systems, especially in government services, rely on

algorithms that operate based on predefined rules and instructions, and do not adapt to the user or learn from data (Quijano-Sánchez et al., 2020). Other recommender systems use a broader range of modern AI technologies and capabilities. These systems utilize, for example, machine learning to adapt to the user over time, or natural language processing and generation to allow the user to interact with the system in a more natural way (Makasi et al., 2021; Nai et al., 2023).

There are various specific transparency challenges associated with the use of algorithmic recommender systems. One of these is the increasing use of complex, data-driven AI technologies, such as deep learning for natural language processing and generation (Klopfenstein et al., 2017). This reliance makes the reasoning and behavior of such systems hard to understand and predict even for their own designers (Burrell, 2016), let alone citizen users. There is an active field of Explainable AI (XAI) working to address the issue of (algorithmic) explainability (Arrieta et al., 2020; Miller, 2019), with techniques being developed to explain not only modern deep learning algorithms but also more traditional rule-based algorithms (Lacave & Diez 2004).

There are, however, transparency challenges that have nothing to do with the complexity or the inner workings of the algorithmic system itself. For instance, some government agencies may hesitate to explain the reasoning behind an algorithmic decision, fearing that people might attempt to “game the system,” as seen in fraud prevention scenarios (Mittelstadt et al., 2016). Similarly, commercial vendors may withhold information about how a system works due to business interests, as seen in cases like COMPAS recidivism prediction (Rudin et al., 2020). Unexplained algorithmic systems may erode citizen trust because they lack basic transparency (Grimmelikhuijsen, 2023; Meijer & Grimmelikhuijsen, 2020), which is an important element in creating impartial and trustworthy institutions (Bauhr & Grimes, 2012; Rothstein & Teorell, 2008).

Government transparency research indicates that governments can be transparent about the decision process—transparency in process—and about the motivation of a decision—transparency in rationale—or a combination of the two (de Fine Licht et al., 2014; Mansbridge, 2009). Transparency in rationale concerns information about the substance of the decision, such as the facts and reasons on which it was based. In contrast, transparency in process refers to transparency about the decision-making process, for example, which procedures were followed and which parties were heard. Following this, we apply the distinction between process and rationale transparency to explanations of algorithmic recommendations by distinguishing “procedural explanations,” “rationale explanations,” and “combined explanations.”

A *procedural explanation* of algorithmic recommendations contains information about the process that results in recommendations. De Fine Licht and

de Fine Licht (2020) equate such a procedural explanation with what Wachter et al. (2017) called a “system functionality” explanation, that is, information about the general functionality of a system. Kizilcec (2016) tested this type of explanation in a study about the grading process of students’ work. The procedural explanation contained information about how someone’s grade had been established—a combination of a peer review by fellow students and an algorithm. In this study, we take a similar view, considering a procedural explanation to contain general information about the algorithm and the steps that were taken in the process leading up to the recommendation.

A *rationale explanation* contains information about the reasoning behind a specific decision, such as the weighing of specific characteristics or individual circumstances (Kim & Routledge, 2018; Wachter et al., 2017). The main aim of a rationale explanation is to provide the underlying arguments for a specific decision (de Fine Licht & de Fine Licht, 2020). For example, Grimmelikhuijsen (2023) included vignettes in which specific reasons were provided for an algorithmic decision about the rejection of a visa application and a call for a welfare fraud investigation. Here, we see a rationale explanation as information about the individual circumstances and specific reasons behind the decision by the recommender system.

Finally, both types of explanations can be put together in a *combined explanation*, entailing information about both the decision procedure and the underlying arguments for a specific decision. Kizilcec (2016), for instance, provided students with a combined explanation, including information about the grading process as well as information about the reasoning behind a specific grade. In our research, the combined explanation consists of a procedural and a rationale explanation.

Note that the above types of explanations should be understood as being on a continuum rather than as completely separate explanations. For example, as de Fine Licht and de Fine Licht (2020, pp. 918–919) argue, it is difficult to present the process leading up to a decision without giving some insight into the reasons on which the decision is based, in the same way that it is difficult to provide the reasons for a decision without mentioning something about the process that led up to the decision.

Although far from conclusive, the literature suggests that these different types of explanation may result in different effects on citizen trust. To investigate these effects, we will present a conceptualization of citizen trust in the next section and develop hypotheses about the expected effects of the different explanations on citizen trust.

Trusting algorithmic recommendations

Citizens’ trust in government in general and in digital public encounters specifically is an important goal in itself and also has instrumental value.

First, citizen trust is an important public value to be pursued by government organizations (Moore, 1995) and can be seen as a goal in itself. Since citizens have no realistic “exit” option for public services (Hirschman, 1970), they are dependent on a single service provider. Ensuring citizen trust in service providers they are “stuck with” should therefore be at the core of (digital) public service delivery. Second, trust may serve an instrumental value (Hoff & Bashir, 2015; Mei & Zheng, 2024). Public organizations have initiated algorithmic recommendations, for instance, to increase the efficiency or quality of public service delivery. Increased trust will also increase citizens’ willingness to follow algorithmic advice, such as acting according to a recommendation (McKnight et al., 2002). In the following sections, we conceptualize trust in algorithmic recommendations and describe how different explanations might affect trust in algorithmic recommendations, leading to four hypotheses.

Researchers in computer and information sciences often use the concept of human-computer trust (HCT) when investigating users’ trust in algorithmic recommendations. HCT is “the extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid” (Madsen & Gregor, 2000, p. 1). Two elements of trust are captured in this definition. First, the *confidence* of a user in the recommendations, actions, and decisions of the decision aid. Second, the *willingness* of the user to act on those recommendations, actions, and decisions of the decision aid. Literature on e-commerce elaborates extensively on the central role trust plays in making consumers comfortable acting on vendor advice and making purchases. Similar ideas about trust—as confidence and willingness to act—can be found in this research area.

McKnight, Choudhury, and Kacmar (2002) developed an influential model of trust in information systems, consisting of trusting beliefs, trusting intentions, and trust-related behaviors. *Trusting beliefs* refer to the perception that the trustee (the one who is to be trusted, e.g., an algorithmic recommendation system or a public organization) has attributes that are beneficial to the truster (the one who trusts, e.g., a citizen). The most cited trusting beliefs are competence (ability of the trustee to do what the truster needs), benevolence (trustee’s caring and motivation to act in the truster’s interests), and integrity (trustee’s honesty and promise keeping) (McKnight et al., 2002, 337). Eventually, trusting beliefs lead to trusting intentions, as Vidotto et al. (2012, p. 576) describe: “Trusting beliefs are a solid conviction that the trustee has favorable attributes to induce trusting intentions.” In our context, this would mean that an algorithmic recommendation by a public organization has attributes that are beneficial to a citizen, leading to trusting intentions.

Trusting intentions refer to the intention to engage in trust-related behaviors. It means that the truster is “securely willing to depend, or intends

to depend, on the trustee” (McKnight et al., 2002, 337). In this paper, this includes citizens’ willingness to follow the advice in an algorithmic recommendation.

Trust-related behaviors refer to actual behavior: acting according to an algorithmic recommendation (McKnight et al., 2002). If trusting beliefs and intentions are present, trust-related behaviors are likely to follow (e.g., Matook, Brown, & Rolf, 2015; Moody, Galletta, & Lowry, 2014). In this paper, we see trust-related behaviors as the actual behavior of citizens when they follow an algorithmic recommendation. We adopt these three dimensions to investigate trust in algorithmic recommendations in digital public service delivery.

While recognizing that trust in government algorithmic recommender systems is influenced by various factors, as shown by Aoki (2020), we specifically focus on understanding the impact of explanations. Aoki’s work demonstrates that clear communication of purposes that directly benefit citizens, such as ensuring consistent response quality and timely interactions, plays a significant role in building trust in chatbots. Additionally, her study reveals interesting variations in initial trust levels when a government announces the use of AI chatbots in different areas (parental support and waste separation). While these factors contribute to trust, our paper’s primary aim remains to thoroughly investigate how explanations influence trust in algorithmic recommendations. The next section delves into various ways in which explanations may affect citizen trust in algorithmic recommendations.

How explanations might affect trust in algorithmic recommendations

Many scholars claim that transparency mechanisms, such as providing explanations, should be implemented to increase user trust in algorithmic recommendations. On the basis of the few empirical studies on algorithmic transparency (Grimmelikhuijsen, 2023; Kizilcec, 2016) and the variety of conceptual publications on the topic (Ananny & Crawford, 2018; de Fine Licht & de Fine Licht, 2020; Meijer & Grimmelikhuijsen, 2020), we formulate four hypotheses regarding the relationships between different types of explanations and citizen trust in algorithmic recommendations. Grimmelikhuijsen (2023) found that explaining algorithmic decisions generates citizen trust, but he did not specify what types of explanations are needed for trust to increase. Kizilcec (2016), on the other hand, tested the effects of different explanations but did so outside the public sector context. Providing explanations has been found to be effective only when expectations have been violated (Kizilcec, 2016; Rader et al., 2018). We take this notion into account by only looking into decisions that violate citizens’ expectations.

Recent studies in public administration have found positive effects of algorithmic transparency, specifically explainability, on citizen trust (Grimmelikhuijsen 2023; Schiff et al., 2022), but they don’t examine or specify types of explanations. A longer tradition of researching different

types of explanations for algorithmic outcomes exists in computer science. We can draw on insights from this field to understand what effects different explanations might have on citizen trust in algorithmic recommendations. For instance, Nothdurft et al. (2014) have investigated the effects of two types of explanations on trust in human-computer interactions. First, the authors looked at “transparency explanations,” which focus on increasing the understanding of users regarding how a technical system works. This is similar to a procedural explanation. Second, they looked at “justifications,” which include a motivation for a specific decision. The goal of this explanation is “to provide support for and increase confidence in a given system advices or actions” (Nothdurft et al., 2014, p. 53). This type of explanation corresponds with the rationale explanation. They found both explanations to be effective in strengthening user trust.

These positive relationships also appeared in other studies. We will discuss findings for the procedural explanations first and then elaborate on the rationale explanations. Studies by Kizilcec (2016), Nothdurft et al. (2014) and Rader et al. (2018) found support for a positive relationship between a procedural explanation and trust in an algorithmic decision but provide different reasons for the mechanism behind this effect. Kizilcec (2016, p. 2391) referred to procedural justice theory (Tyler, 1990), which posits that individuals can be satisfied with negative outcomes as long as the underlying procedure is considered to be just. Nothdurft et al (2014) and Rader et al. (2018) provided alternative reasonings for the positive effects on user trust of explaining the decision procedure. The mechanism, as Nothdurft et al. (2014) explained, is that a procedural explanation helps users to understand how a system works and reasons and thereby changes the user’s perception of the process from a black-box model to a comprehensible system. This allows the users to build a mental model of the system, including its underlying reasoning processes (Nothdurft et al., 2014, p. 53). Rader et al. (2018) followed a similar line of reasoning. They argued that knowing how an algorithmic recommender system produces a recommendation helps users to understand and act upon the recommendation. A procedural explanation helps to fill the gap between users’ intuitions about a system and the actual internal processes of a system. In fact, this makes explanations “white box” descriptions of the inputs, outputs, and steps a system takes to arrive at a specific outcome (Rader et al., 2018, p. 2), which increases user trust in the system. Despite disagreement about the underlying mechanisms, all studies show the positive effects that a procedural explanation can have on trust in algorithmic outcomes. We therefore expect a positive relationship between a procedural explanation and trust in an algorithmic recommendation. Trust will be higher with a procedural explanation than when no explanation is provided.

Hypothesis 1: Citizen trust (i.e., trusting beliefs, intentions, and trust-related behaviors) in algorithmic recommendations with a procedural explanation is higher than in algorithmic recommendations without an explanation.

To investigate the effects of the rationale explanation, we draw on studies from Hind (2019) and Kim and Routledge (2018), who argued that a rationale explanation could lead to an increase in trust in algorithmic decisions because people understand *why* they receive what they receive. In addition, de Fine Licht and de Fine Licht (2020, p. 924) argued that if citizens know the reasons behind algorithmic decisions, they should have a greater understanding of why the decisions were made. A rationale explanation allows citizens to form an opinion about the desirability of the decision. Similarly, Rader et al. (2018, p. 2) described how users feel more comfortable and satisfied with a recommendation and are more willing to accept it when they believe they understand why the recommendation was made. Our second hypothesis therefore expects trust to be higher with a rationale explanation than when no explanation is provided about the algorithmic recommendation.

Hypothesis 2: Citizen trust (i.e., trusting beliefs, intentions, and trust-related behaviors) in algorithmic recommendations with a rationale explanation is higher than in algorithmic recommendations without an explanation.

While these two hypotheses provide expectations for procedural and rationale explanations separately, the literature provides mixed insights into the comparative strengths of the effects. Some scholars argue that a rationale explanation ensures a better understanding of a decision outcome compared to a process explanation (de Fine Licht & de Fine Licht, 2020). For a procedural explanation to be understandable, they argue, the code would likely need to be much simpler than it is in current systems, assuming that the public needs a chance to truly grasp it. A simplified code would potentially make it easier to manipulate the system and give fewer incentives to innovate (e.g., Lepri et al., 2018; Zarsky, 2016). Demands for transparency would therefore likely result in decisions of inferior quality and, as a result, less trust (de Fine Licht & de Fine Licht, 2020, p. 921).

Empirical work on algorithmic trust from in nonpublic contexts, however, paints a different picture. Rader et al. (2018) found no statistical differences between procedural and rationale explanations about Facebook's News Feed algorithm but acknowledged that this might be the result of too little difference between the two types of explanations in the operationalization of their study. In another study, Kizilcec (2016) examined the effect of different explanations (no, procedural and combined) on trust in an algorithmic interface. He found that an explanation outlining how an algorithm worked (i.e., procedural explanation) had the strongest

positive effect on trust in the outcome. This can be understood using procedural justice theory, which posits that individuals can be satisfied with negative outcomes as long as they consider the underlying procedure to be just (Tyler, 1990). Opaque algorithmic procedures, therefore, may erode trust in algorithmic recommendations (Rudin, 2019). In sum, this means that there is scarce, and contradictory, evidence for which type of explanations work better in terms of trust. Since Kizilcec's experiment more closely aligns with the context of algorithmic recommendations, we hypothesize that procedural explanations are most effective and therefore trump rationale explanations.

Hypothesis 3: Citizen trust (i.e., trusting beliefs, intentions, and trust-related behaviors) in algorithmic recommendations with a procedural explanation is higher than in algorithmic recommendations with a rationale explanation.

Our fourth and final hypothesis concerns combining procedural and rationale explanations, which could perhaps provide the best of both worlds. At the same time, there is a risk that people will receive so much information that they do not read the message properly, which leaves them dissatisfied with the final recommendation or decision. This is also referred to as an *information overload* (Hosseini et al., 2015). An information overload in an explanation can lead to a decrease in trust. Finding the relevant information would be, as de Fine Licht and de Fine Licht (2020, 922) noted, “as difficult as finding the proverbial needle in a haystack”. Making all information available (full transparency) would therefore not be more beneficial than providing a single explanation (partial transparency). This expectation is supported by Kizilcec's experiment (2016), which found that providing a combination of rationale and procedural explanations decreased trust. This has led to the fourth and final hypothesis, which expects a combination of a procedural and a rationale explanation to be less effective in strengthening citizen trust in algorithmic recommendations than a single explanation.

Hypothesis 4: Citizen trust (i.e., trusting beliefs, intentions, and trust-related behaviors) in algorithmic recommendations with a combined explanation is lower than in algorithmic recommendations with only a procedural or a rationale explanation.

In the next section, Study 1, we test these four hypotheses.

Study 1: Initial test of the four hypotheses

The intelligent crime reporting tool

We developed a mock version of an actual algorithmic recommender system, namely the Intelligent Crime Reporting Tool (ICRT) of the Dutch police (Odekerken et al., 2022). Our research therefore addresses a recent

call by Zuiderwijk et al. (2021) to move beyond the generic focus on AI in public administration research, focusing on a specific AI tool (ICRT) in a specific domain (police), in a specific country (the Netherlands).

The ICRT can be accessed via the police website¹ by citizens who believe they have been victims of online fraud, such as scams involving fake web shops and malicious secondhand traders on platforms like eBay, where they order and pay for a product but never receive it. Through the ICRT, citizens can describe what happened to them, with the ICRT asking follow-up questions when necessary. The ICRT then automatically assesses whether it is likely that the citizen has been a victim of fraud, and, if so, recommends that they proceed with filing an official report online. If the ICRT assesses that fraud did not occur, it can provide recommendations for other actions the citizen can take (e.g., contacting the trading platform). Our mock version of the ICRT asked participants the same questions as the real ICRT of the Dutch police.

The ICRT primarily relies on a rule-based legal model of the fraud domain, in combination with some basic natural language processing for analyzing the free-text descriptions of (alleged) fraud, provided by citizens. It is an example of digital public services, defined as “public services provided using internet-based technologies wherein a citizen’s interaction with a public organization is mediated partly or completely by an IT-system” (Lindgren et al., 2019, 429). With online shopping being increasingly commonplace, the Dutch police receive tens of thousands of complaints regarding online trade fraud every year. In these complaints, it is not always clear whether a case is fraud or not, particularly since citizens often do not know the nuances of the law and hence tend to have a hard time identifying which facts are relevant from a legal standpoint. So, in addition to allowing citizens to file a report of a crime online, the ICRT also acts as an official source of information, explaining why a citizen’s case is (or is not) a case of fraud and recommending further actions to take besides filing an official report. We believe that asking for advice on and possibly reporting a crime to the police through the ICRT is a very relevant case to investigate when learning about trust in algorithmic recommender systems because it sheds light on a digital public service that many citizens are likely to use.

Materials and methods

Experimental setting and procedure

To examine the previously stated research question, we designed an online survey experiment in which we randomly varied the type of explanation that followed the recommendation that participants received. The design is shown schematically in Figure 4 in the Appendix. This fully randomized

experiment has the advantage of high internal validity (Shadish et al., 2002). Therefore, we can draw firm cause-effect conclusions about the effects of different types of explanations on trust in algorithmic recommendations.

As illustrated by [Figure 4 in the Appendix](#), the experiment started by asking all participants three demographic questions in order to ensure a representative sample of the Dutch population in terms of gender, age, and education. Next, participants were told that they would read a hypothetical case of possible online fraud and that they should use the Intelligent Crime Reporting Tool (ICRT) to file a report of this case.

After using the ICRT (going through all the steps and questions illustrated in [Appendix D](#)) to report their case of online fraud, all participants received a recommendation not to file an official report of online fraud. This recommendation went against their expectations, as they were told in the hypothetical situation that they suspected to be a victim of online fraud. All participants received an explanation about the recommendation. Participants were randomly assigned to a group in which an experimentally varied explanation was presented. The potential fraud situation, the recommendation, and the experimental vignettes used for the explanation can be found in the appendix.

After reading the recommendation and the explanation, participants were first asked whether they wanted to file a report of online fraud, which allowed us to measure their actual behavior. Then, they were asked about their trusting beliefs and intentions regarding the recommendation they received.

Sample and data collection

We conducted the survey experiment online using Qualtrics in June 2021. Prior to data collection, we preregistered the experiment using the Open Science Foundation format (Bowman et al., 2020), and the study received ethical approval from the institutional Ethical Review Committee.² Data was collected using the sample-only service of *Dynata*, a renowned global recruitment firm. Dynata has a large respondent pool, which they use for distributing surveys. Respondents are free to choose whether they want to apply for Dynata's participant pool and could therefore decide for themselves whether they wanted to participate in our experiment. Dynata provided a sample of 717 respondents for our experiment with the following parameters: 1) Dutch speakers living in the Netherlands, and 2) participants that represented the country's population in terms of age, gender and level of education. Respondents were reimbursed for their participation in the survey-experiment upon completion by Dynata. Using the software program G*Power, we conducted an *a priori* power analysis

to calculate the estimated sample size (Power = .9 and $\alpha = .008$).³ The sample for the experiment was $n = 717$. We used stratified sampling methods to ensure we had a representation of the Dutch population. The background variables of the sample are reported in [Appendix E](#). The sample resembles the Dutch population regarding three key background variables: education, gender, and age. We took these background variables into account as control variables to carry out balance checks. Participants in our sample were somewhat older compared to the Dutch population (see [Appendix E](#)).

Experimental conditions

We have four different explanation conditions in our experiment (see [Appendix G](#) for the exact wording of the explanation manipulations). Participants in the control condition received no explanation for the statement that, based on their story, they did not need to file a report to the police. The procedural condition included information about the decision procedure: how the ICRT used various algorithms to analyze their text and arrive at the conclusion that the webshop was trustworthy. Participants in the rationale condition were told the specific reason why the webshop was trustworthy: it was affiliated with a quality mark, which guaranteed an extensive screening procedure for all associated webshops.⁴ The combined condition included both the procedural and the rationale explanations.

Measures

The trusting beliefs of participants were measured after the experiment by means of a questionnaire, using Gulati, Sousa, and Lamas's (2019) human-computer trust scale that investigates user trust in human-computer interactions. We used cognitive interviews to test the scale (DeVellis, 2017). By asking eight people unfamiliar with the topic what they understood the items to be about and how they would formulate responses to the items, we were able to identify confusion about vocabulary and concepts as well as misunderstandings related to response options that we had overlooked. This led to using four items that measure general feelings of trust. We have adapted these items according to the research needs of the current study. Trusting beliefs were thus measured using four items ($\alpha = .96$) on a scale from 1 ("totally disagree") to 7 ("totally agree").

Similar to participants' trusting beliefs, trusting intentions were measured using a questionnaire. The scale for the intention to act according to the recommendation was derived from the 5-item scale measuring intention to follow vendor advice by McKnight et al. (2002) and adapted to the context of the ICRT. Cognitive interviews resulted in dropping one item. Trusting intentions were therefore measured using four items ($\alpha = .96$).

on a scale from 1 (“totally disagree”) to 7 (“totally agree”). The items used for measuring trusting beliefs and intentions can be found in [Appendix A](#).

Before questioning the participants about their trusting beliefs and intentions, participants were asked whether or not they wanted to report the crime (trust-related behaviors), measured as “Don’t file a report”/“File a report.” This allowed us to record the actual behavior of participants based on the recommendation.

The survey was extensively pre-tested prior to its implementation in three ways. First, we asked two developers of the real ICRT from the Netherlands police to revise the experimental vignettes in order to increase mundane realism. Second, to ensure measurement validity, we carried out eight face-to-face cognitive interviews to ensure survey questions were well understood, which led to some revisions of the survey items. Third, we conducted a pilot study with 82 people to test the reliability of the new scales of trusting beliefs and intentions. Furthermore, we used a factual manipulation check (FMC) in the pilot study to test the experimental vignettes. Findings from the FMC in the pilot study resulted in some changes in the vignettes to make them more distinctive in Study 1. The results of the FMC of both the pilot study and Study 1 can be found in [Appendix H](#).

Analyses

For the outcome variables *trusting beliefs* and *trusting intentions* we used independent samples *t*-tests to compare means between groups. For *trust-related behaviors* we used *z*-tests to compare proportions between groups. We used *p*-values with a significance level of $\alpha = 0.05$ for all tests. However, we used a Bonferroni adjustment to correct for an inflated chance of Type 1 errors due to multiple testing. For each outcome variable, we conducted six comparisons of means or proportions between groups. Therefore, we multiplied every *p*-value by six. To calculate effect sizes, we used Cohen’s *d* for the outcome measures trusting beliefs and intentions, and Cohen’s *h* for the outcome measure trust-related behaviors (Cohen, 1988).

Results

Descriptive statistics

Means and standard deviations for trusting beliefs and intentions can be found in [Table 1](#). Participants that received no explanation for the recommendation by the intelligent crime reporting tool scored on average between 3 (fairly disagree) and 4 (neither agree nor disagree) on a scale from 1-7 for their trusting beliefs and intentions. Participants that received a procedural, rationale, or combined explanation scored on average between 4 (neither agree nor disagree) and 5 (fairly agree) on a scale from 1-7 for their trusting beliefs and intentions. Furthermore, for trust-related

Table 1. Descriptive statistics of trusting beliefs, intentions, and trust-related behaviors in the algorithmic recommendation for different types of explanations.

	<i>n</i>	Trusting beliefs		Trusting intentions		Trust-related behaviors
		Mean	SD	Mean	SD	% that did not report the crime
No explanation	177	3.59	1.54	3.61	1.67	42.29%
Procedural explanation	177	4.75	1.48	4.84	1.52	74.58%
Rationale explanation	179	4.86	1.30	4.88	1.50	75.98%
Combined explanation	184	4.87	1.41	4.92	1.58	76.09%

behaviors we reported the percentages of participants that did not report the crime in Table 1, thus the percentage of participants that followed the algorithmic recommendation not to report the crime. Less than half of the participants that received no explanation followed the recommendation not to report the crime, compared to approximately three quarters of the participants that received any type of explanation.

Analyses

We tested four hypotheses, the results of which are visualized in Figure 1. We found support for the first hypothesis (procedural > no explanation). Participants that received a procedural explanation had significantly higher *trusting beliefs* than participants that received no explanation, $t(351.49) = 7.18$, $p < .001$. This is a medium effect ($d = .76$). In addition, participants that received a procedural explanation had significantly higher *trusting intentions* than participants that received no explanation, $t(348.76) = 7.25$, $p < .001$, with a medium effect ($d = .77$). Finally, the proportion of participants with a procedural explanation that followed the recommendation not to report the crime (*trust-related behaviors*) was significantly higher than the proportion of those with no explanation, $z = 6.05$, $p < .001$, with a medium effect ($h = .66$).

We also found support for the second hypothesis (rationale > no explanation). Participants who received a rationale explanation had significantly higher *trusting beliefs* than participants that received no explanation, $t(343.28) = 8.34$, $p < .001$, with a large effect ($d = .89$), and significantly higher *trusting intentions*, $t(349.22) = 7.58$, $p < .001$, with a large effect ($d = .80$). Finally, for *trust-related behaviors*, the proportion of participants with a rationale explanation that followed the recommendation not to report the crime was significantly higher than the proportion of those with no explanation, $z = 6.35$, $p < .001$. This is a medium effect ($h = .68$).

We did not find any support for the third hypothesis (procedural > rationale). Participants with a procedural explanation did not have significantly higher *trusting beliefs* than participants with a rationale explanation, $t(347.25) = -.74$, $p = 1$, nor did they have significantly higher *trusting intentions*, $t(353.58) = -.30$, $p = 1$, or *trust-related behaviors*, $z = .31$, $p = 1$.

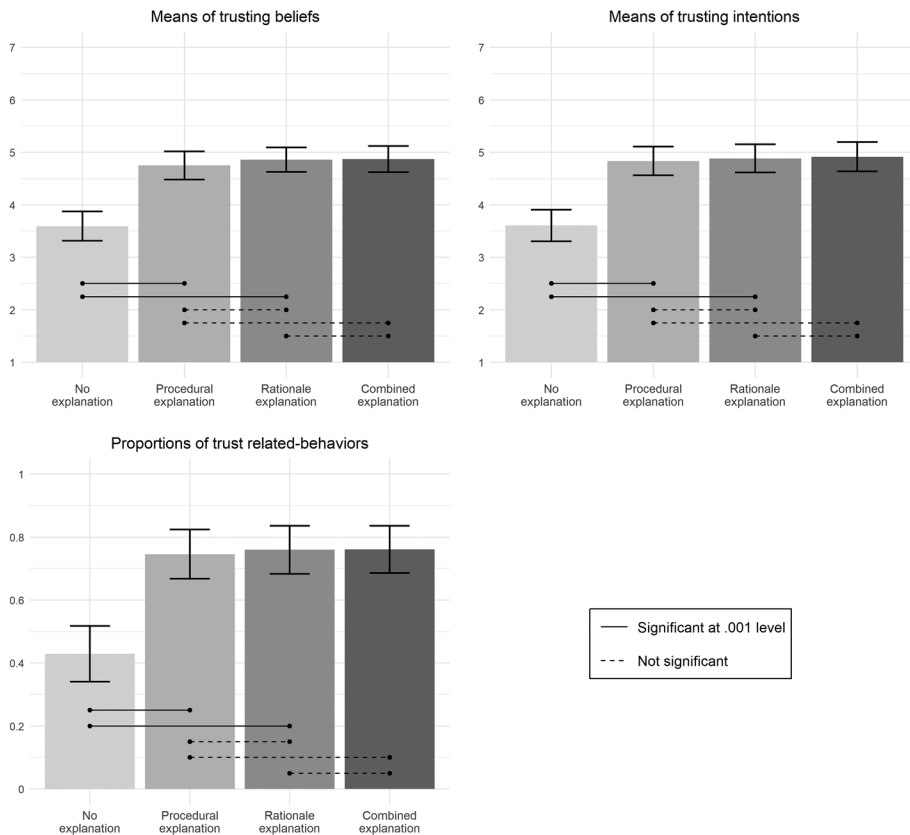


Figure 1. Analyses of differences between explanation conditions.

Note. Error bars are confidence intervals with Bonferroni-adjusted standard errors.

Lastly, we did not find support for the fourth hypothesis (combined < procedural or rationale). Participants with a combined explanation did not have significantly lower *trusting beliefs* ($t(356.2) = .81, p = 1$), *trusting intentions* ($t(359) = .50, p = 1$), or *trust-related behaviors* ($z = .33, p = 1$) than participants with a procedural explanation. Similarly, participants with a combined explanation did not have significantly lower *trusting beliefs* ($t(360.05) = .10, p = 1$), *trusting intentions* ($t(360.83) = .20, p = 1$) or *trust-related behaviors* ($z = .02, p = 1$) than participants with a rationale explanation. A sensitivity analysis excluding participants that failed the factual manipulation check did not alter the results of our analyses (i.e., no significant effect became nonsignificant, and no nonsignificant effect became significant).

Conclusion Study 1

To answer our research question “What are the effects of different types of explanations on citizen trust in algorithmic recommendations?” we examined

four theoretical expectations. In line with the first two hypotheses, our findings show that citizen trust in algorithmic recommendations is significantly higher with a procedural or a rationale explanation than without an explanation. Our results showed that providing any type of explanation will increase trust in algorithmic recommendations compared to providing no explanation. Regarding the third and fourth hypotheses, there were no significant differences in trust between a procedural and a rationale explanation, nor were there any significant differences in trust between providing a combined explanation instead of merely a procedural or a rationale explanation. Using sequential factorials, which are “a model of a sequence of experiments to progressively deepen and draw out the implications of a line of reasoning” (Sniderman, 2018, p. 266), we tested a possible interpretation for these null findings in Study 2. This means that the results of the first experiment determined the design and input of the follow-up experiment, which helped to deepen our understanding of which explanations increase citizen trust in algorithmic service provision. In line with our sequential factorial design, we identified a new concept, a directive explanation, that seemingly plays a key role in interpreting findings that cannot be understood by concepts previously discussed in the literature. This concept led to a fifth hypothesis that was tested in Study 2.

Study 2: Follow-up test with a directive explanation as an additional manipulation

In line with the principles of sequential factorials design (Sniderman, 2018), we explored the literature to identify a new concept which could help us to better understand the relationship between explanations and citizen trust. In contrast to the hypotheses formulated in Study 1, we did not find any significant differences between different types of explanations. While this could be a true null effect (i.e., there is no effect of explanation type), it is also possible that this lack of differentiation is an artifact of our experimental design and the way the explanations were formulated. Specifically, every treatment contained the following sentence: “Check the website of the quality mark (www.webshopqualitymark.nl) to see what you can do to get your money back.” Providing such a sentence offers citizens a concrete path forward with actions to undertake in order to achieve their desired outcome, i.e., to get their money back. After all, an important reason for victims of online fraud to report their case is to get their money back (Cross, 2018).

According to Singh et al. (2021a), explaining which steps to take in order to achieve the desired result is covered by a so-called directive explanation. A *directive explanation* lists specific actions or interventions an individual needs to take to achieve a desired outcome (Singh et al.,

2021a, p.1). If a recommendation is detrimental to a citizen (e.g., advising against filing a report of fraud due to its low likelihood of success), then a directive explanation provides information about how the citizen could obtain their desired outcome, if possible.

A directive explanation can be seen as a layer of explanation on top of other types of explanations to justify algorithmic decisions that affect people personally (Singh et al., 2021b). It increases an individual's sense of control over a decision outcome, which could strengthen trust. In the context of our research, this could mean that citizen trust in algorithmic recommendations is higher when a directive explanation is provided, because it potentially brings a recommendation more in line with an individual's preference. We therefore hypothesize that providing a *directive explanation* is a specific element in an explanation that may foster trust.

Hypothesis 5: Citizen trust in algorithmic recommendations with a directive explanation is higher than in algorithmic recommendations without a directive explanation.

We cannot formulate hypotheses about the effect of a directive explanation on specific types of explanations, such as a procedural, rationale, or combined explanation because it has not been empirically examined before. Thus, in the next section and main analysis, we look at the effect of a directive explanation in general. This allows us to contribute to theory development about explaining public sector algorithmic recommendations.

Materials and methods

Experimental setting and procedure

Our second study builds on what we learned from the findings and limitations of our first survey experiment. The experimental setting and procedure are the same as the first experiment, except for the experimental treatments. As illustrated in Figure 5 in the Appendix, Study 2 has eight experimental groups instead of the four in Study 1. Study 2 investigates whether citizen trust in algorithmic recommendations is higher if a directive explanation is provided.

The *directive explanation* is operationalized as follows: “Check the website of the quality mark (www.webshopqualitymark.nl) to find out what you can do to get your money back.” This element provides participants with concrete steps forward to achieve their desired outcome (i.e., get their money back). In our study we use the element of the directive explanation, as operationalized above, to test a greater degree of

directiveness versus little to no directiveness. See [Appendix J](#) for the experimental vignettes used in Study 2. A subjective manipulation check (SMC) shows that participants experienced the treatment as we intended (see [Appendix L](#)).

Sample and data collection

We conducted the survey experiment online on Qualtrics in September 2021 using the sample-only service of the recruitment firm *Dynata*. Prior to data collection, we preregistered the experiment using the Open Science Foundation format (Bowman et al., 2020).⁵ Additionally, the study was approved by the institutional Ethical Review Committee. We conducted an *a priori* power analysis to calculate the estimated sample size (Power = .9 and $\alpha = .05$) using the software program G*Power.³ The sample for the experiment was $n = 1005$. The sample resembles the Dutch population regarding level of education, gender, and age, although the participants in our study were slightly older in comparison to the Dutch population (see [Appendix E](#)).

Analyses

Based on the hypothesis, we compared the means of trusting beliefs and intentions of Groups 1-4 with the means of trusting beliefs and intentions of Groups 5-8 using *t*-tests with planned contrasts. For trust-related behaviors we used a *z*-test to compare proportions between groups. We compared the proportion of trust-related behaviors of Groups 1-4 with the proportion of trust-related behaviors of Groups 5-8 using planned contrasts. We used *p*-values with a significance level of $\alpha = .05$. To calculate effect sizes, we used Cohen's *d* for the outcome measures trusting beliefs and intentions and Cohen's *h* for the outcome measure trust-related behaviors (Cohen, 1988).

Results

Descriptive statistics

[Table 2](#) shows descriptive statistics for trusting beliefs, intentions, and trust-related behaviors. The descriptive statistics for the two main sets of participants (with and without a directive explanation) are shown in bold. All participants, receiving a directive explanation or not, scored on average between 4 (neither agree nor disagree) and 5 (fairly agree) on a scale from 1-7 for their trusting beliefs and intentions. Furthermore, more than two-thirds of all participants followed the recommendation not to report the crime. [Table 2](#) also shows the descriptive statistics for the explanation subsets (Groups 1 to 4) that received no directive explanation and the subsets (Groups 5-8) that did receive a directive explanation.

Table 2. Descriptive statistics of trusting beliefs, intentions, and trust-related behaviors in the algorithmic recommendation with or without a directive explanation.

	<i>n</i>	Trusting beliefs		Trusting intentions		Trust-related behaviors
		Mean	SD	Mean	SD	% that did not report the crime
No directive explanation	497	4.60	1.47	4.69	1.57	69.22%
G1: No explanation	125	4.03	1.52	3.96	1.60	55.20%
G2: Procedural explanation	123	4.44	1.46	4.43	1.57	64.23%
G3: Rationale explanation	123	4.98	1.28	5.15	1.35	79.67%
G4: Combined explanation	126	4.95	1.42	5.22	1.40	77.78%
Directive explanation	508	4.69	1.41	4.75	1.53	72.24%
G5: No explanation	129	4.25	1.48	4.25	1.62	58.91%
G6: Procedural explanation	125	4.92	1.29	4.99	1.35	74.40%
G7: Rationale explanation	124	4.98	1.20	5.11	1.34	83.06%
G8: Combined explanation	130	4.64	1.53	4.66	1.63	73.08%

Analyses

We did not find support for our fifth hypothesis that citizen trust in algorithmic recommendations with a directive explanation is higher than in algorithmic recommendations without a directive explanation. Participants that received the directive explanation manipulation did not have significantly higher *trusting beliefs* than participants that received no directive explanation ($t(998.64) = 1.06, p = .144$). In addition, no significant effect was found on *trusting intentions* ($t(1000.6) = .64, p = .260$) or *trust-related behaviors* ($z = 1.06, p = .854$). Overall, the main analysis showed that participants that received an explanation with a directive explanation did not have significantly higher trust in algorithmic recommendations than participants that received an explanation without a directive explanation.

Post-hoc analyses

We introduced a directive explanation as a new concept that could be an important element in explaining algorithmic decisions. Due to its novelty, we could not build on theoretical foundations to develop specific theoretical expectations. In other words, we were only able to formulate a hypothesis for the main effect of a directive explanation on citizen trust. To examine whether a directive explanation has an interaction effect, meaning different effects on specific explanations, we conducted four post-hoc analyses.

Following H5, we expected that the groups with a directive explanation had higher trust in algorithmic recommendations than the groups without a directive explanation. We compared all groups with equal explanation treatments separately, so we compared Groups 1 and 5, 2 and 6, 3 and 7, and 4 and 8. All comparisons were directional, expecting that the groups with a directive explanation would have higher trust scores than the groups without a directive explanation, as stated in H5. As in Study 1, to adjust for the inflated chance of Type 1 errors, we used a Bonferroni adjustment.

For these post-hoc analyses we multiplied the p -value by four because we conducted four comparisons between groups, with an alpha of .05. The full results of all four post-hoc comparisons can be found in [Appendix M](#).

Overall, the post-hoc analyses indicated that a directive explanation had an effect only on the procedural explanation. In other words, only the comparison between Group 6 (procedural explanation *with a directive explanation*) and Group 2 (procedural explanation *without a directive explanation*) resulted in significant results, as illustrated in [Figure 2](#).

Because a directive explanation has an interaction effect on the procedural explanation *only*, it is interesting to see if the effect of a directive explanation has mitigated possible differences between the procedural explanation and other types of explanations without a directive explanation. After all, this was a potential reason for not finding differences between types of explanations in Study 1. To test this, we compared the group of participants that received a procedural explanation without a directive explanation with all the other groups without a directive explanation, so we compared Groups 2 and 1, 2 and 3, and 2 and 4. The comparisons were nondirectional since we did not have theoretical expectations for these comparisons. To adjust for the inflated chance of Type 1 errors, we used a Bonferroni adjustment of three times the p -value, with an alpha of .05. The full results of these comparisons can be found in [Appendix N](#).

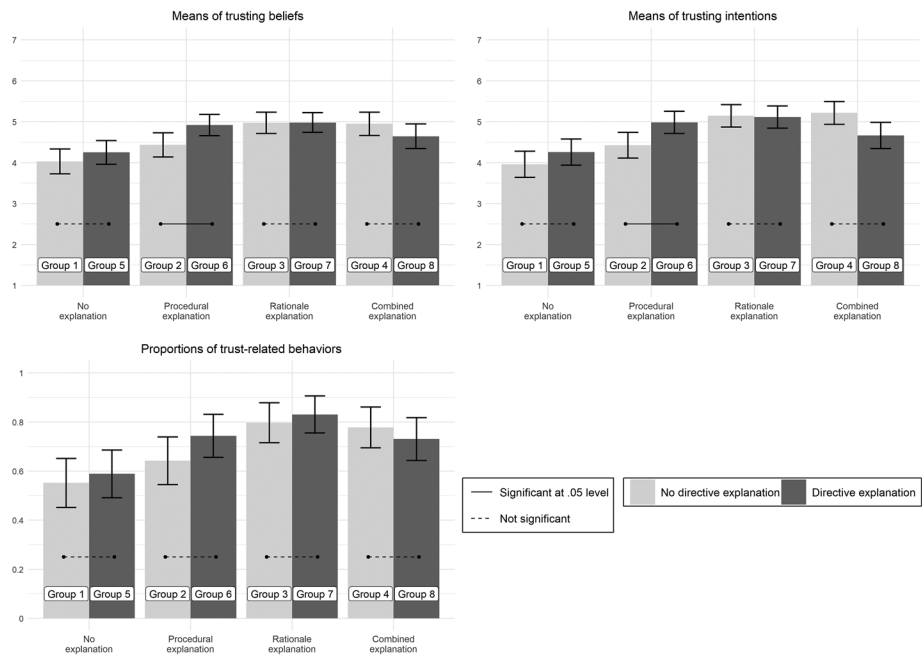


Figure 2. Post-hoc analyses of directive explanation differences within equal explanation conditions.

Note. Error bars are confidence intervals with Bonferroni-adjusted standard errors.

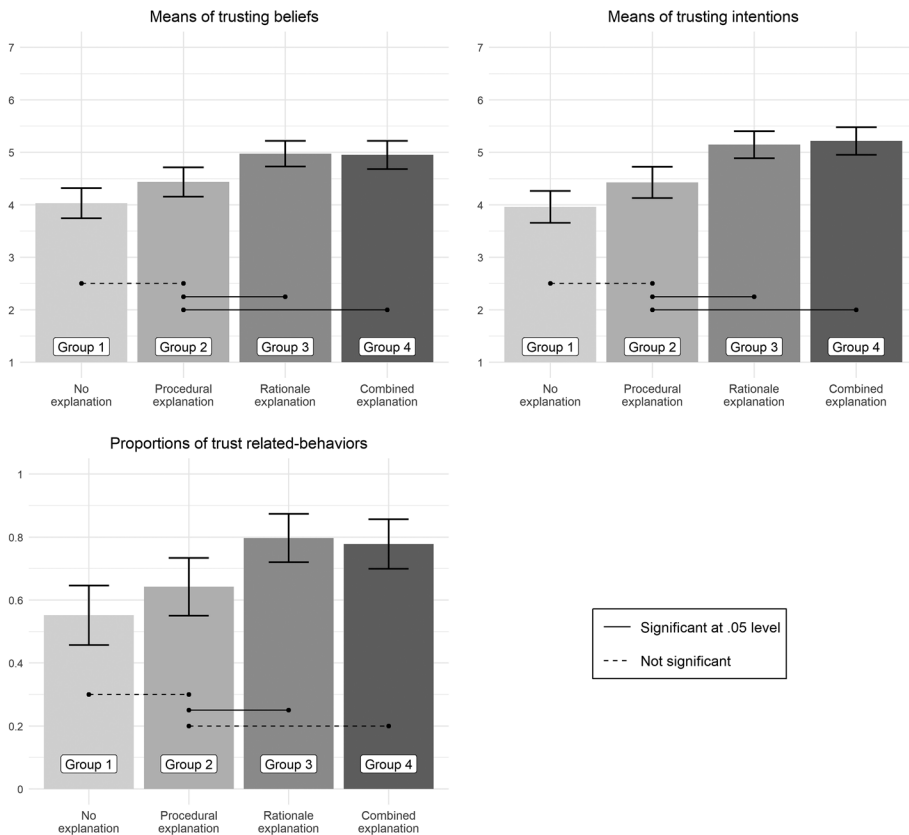


Figure 3. Post-hoc analyses of differences between the procedural explanation condition without a directive explanation and all the other explanation conditions without a directive explanation.

Note. Error bars are confidence intervals with Bonferonni-adjusted standard errors.

These post-hoc analyses resulted in three important findings that are visualized in Figure 3. First, we found no significant differences in trust between participants that received a procedural explanation without a directive explanation (Group 2) and participants that received no explanation without a directive explanation (Group 1). Second, the results showed that participants receiving a rationale explanation without a directive explanation (Group 3) had significantly higher trust in the algorithmic recommendation than participants with a procedural explanation without a directive explanation (Group 2). Third, participants with a combined explanation without a directive explanation (Group 4) had higher trust in algorithmic recommendations than participants with a procedural explanation without a directive explanation (Group 2).

Conclusion Study 2

In Study 2, we focused on a novel element in explaining algorithmic recommendations, namely a directive explanation that lists specific actions

a citizen needs to take to achieve a desired outcome. We examined the main effect of a directive explanation on trust in algorithmic recommendations. We did not find support for our fifth hypothesis, which expected that citizen trust in algorithmic recommendations would be higher when a directive explanation was provided compared to when no directive explanation was provided. Overall, there was no main effect of a directive explanation on citizen trust in algorithmic recommendations.

However, a post-hoc analysis suggested that adding the element of a directive explanation to a procedural explanation does, in fact, increase trust in algorithmic recommendations. In addition, we found that a directive explanation mitigated the differences between a procedural explanation and the other three experimental groups that received no directive explanation in two ways. First, the analyses showed that participants that received no explanation with no directive explanation (Group 1) did not have significantly higher or lower trust than participants with a procedural explanation without a directive explanation (Group 2). Statistically, this means that when no directive explanation is provided, a procedural explanation is as effective as no explanation in strengthening citizen trust in algorithmic recommendations.

Second, we found that participants that received a procedural explanation without a directive explanation had less trust in the algorithmic recommendation than participants with a rationale or a combined explanation without a directive explanation. Thus, when no directive explanation is provided, a rationale and a combined explanation are more effective than a procedural explanation in strengthening citizen trust in algorithmic recommendations. [Table 3](#) summarizes the results of the main findings of both Study 1 and Study 2.

Discussion

Our research has important implications for academic and societal debates on algorithmic transparency in digital service delivery and raises a new set of questions for further research into algorithmic explanations.

Table 3. Overview of results per hypothesis of Study 1 and Study 2.

Hypothesis			Result
H1: Procedural	>	No explanation	Supported
H2: Rationale	>	No explanation	Supported
H3: Procedural	>	Rationale	Rejected
H4a: Combined	<	Procedural	Rejected
H4b: Combined	<	Rationale	Rejected
H5: Directive explanation	>	No directive explanation	Rejected

Academic implications

We found that the type of explanation matters, but only to an extent. Indeed, while some argue that a rationale explanation (explaining *why*) leads to the most pertinent increase in citizen trust (e.g., de Fine Licht & de Fine Licht, 2020) and others show that a procedural explanation (explaining *how*) is most effective in strengthening trust (e.g., Kizilcec, 2016), we found that generally similar effects can be achieved using all types of explanations. Our manipulation checks showed that, generally, people were able to distinguish and recognize the different explanation types. It is therefore unlikely that the lack of differences is an artifact of our research design. One way to explain these findings is by looking at social psychology literature on what constitutes a persuasive message. One of the most-used persuasion models is the Elaboration Likelihood Model (ELM) (Petty & Briñol, 2011; Petty & Cacioppo, 1986). The ELM describes two routes to persuasion. On the central route, persuasion will likely occur based on someone's careful and thoughtful consideration of the benefits of the information. On the peripheral route, on the other hand, persuasion is not based on the intrinsic merits of the information but on peripheral cues such as format and visual presentation.

People are likely to take the peripheral route when the stakes are low and when there is little involvement with the given information. Individuals generally have limited time and capacity for information processing and simplify the options and information provided in order to make decisions. Therefore, the choice to trust a government agency may not necessarily be conscious or rational. Based on the ELM, Grimmelikhuijsen and Meijer (2014) argue that if citizens have high prior knowledge about a topic, providing information will not affect a government organization's perceived trustworthiness. Citizens' trust is "cognition based," meaning that their prior knowledge becomes the main driver for perceived trustworthiness along with the information provided to them (Grimmelikhuijsen & Meijer, 2014, p. 153).

In addition, Alon-Barkat (2020) shows that citizens do not rationally decide to trust the government based on the information provided to them. Using the ELM, he found that symbols increase citizens' trust by making citizens pay less attention to logically unpersuasive information, and thus counteracting its negative effect. Even though they were carried out in other contexts, these studies show that the ELM is relevant for explaining unexpected and non-rational trusting attitudes and behaviors of citizens based on the information provided to them.

Despite our expectations that differences would exist between explanations, our studies showed that all types of explanation led to similar levels of trust, which could suggest that participants were taking the peripheral

route to process the information they received. The experiments involved a hypothetical fraud case in which shoes were not delivered after being purchased online. This could be an example of a case where the stakes were low for most participants, meaning that results of attitude or behavioral changes between the types of explanations were limited, but between no explanation and any explanation, the results were significant. Future research is necessary to better understand how the ELM model can be used in theorizing the relationship between explanations (transparency) and trust in the context of algorithmic recommendations by public services. The model could be used to examine whether higher stakes cause participants to take the central route, which might result in differences between the types of explanations. Still, we must pay attention to the implications of this interpretation.

From a democratic point of view it might be harmful that citizens do not critically evaluate the explanations that are provided to them, but that they find *any* explanation satisfactory. As explained above, accepting any explanation contrasts with the normative expectation that citizens in a democracy should form their opinions on the government based on rational critical thinking and healthy skepticism (Norris, 2022). Uncritical citizens may be more vulnerable to manipulation and government abuse in public service delivery. Therefore, we argue that having correct and understandable explanations is not enough. There should be checks and balances in place for the use and explanation of algorithmic recommendations for public organizations. The latter concern, if valid, necessitates the development of algorithmic scrutiny or accountability mechanisms (see also Grimmelikhuisen & Meijer, 2022; Wieringa, 2020).

Societal implications

Based on the finding that explanations are highly important for trustworthy digital public services, we argue for including explanations as a design requirement for algorithmic recommendation systems in public services. This is in line with the Transparency by Design principles of Felzmann et al. (2020), which highlight areas where system designers need to address transparency concerns regarding artificial intelligence during the design process. Moreover, requiring explanations for algorithmic recommendations connects with the work of Rudin (2019), who states that the way forward is to design algorithmic models that are inherently interpretable, rather than creating methods for explaining black box models. Including explanations as a core element of the design process will ensure the development of algorithms that contribute to citizen trust. This requires tool designers to ensure that a recommender system not only provides a recommendation but also (procedural and rationale) explanations for it. This is crucial for

a representative government that is, or can be, called to account for its actions. Moreover, it stimulates the designers to think about the pros and cons of the tool, thus leading to higher-quality decisions that are reflected in an AI tool (de Fine Licht & de Fine Licht, 2020). While embracing greater openness in digital public service delivery, it is important to recognize that openness might incur substantial costs by diminishing operational efficiency, posing a delicate balancing act for public managers (Halachmi & Greiling, 2013). However, this paper showed that considering explanations as an integral part of the AI design process in the public domain is crucial for fostering citizen trust in digital encounters and should therefore be prioritized.

Research limitations and future research

This study is subject to some limitations that provide future research directions. An important point of discussion is to what extent our findings are generalizable to other contexts. This question touches upon three elements of our study: the design, the sample, and the experimental scenario. First, regarding the design, we used a mock version of a real algorithmic recommender system with, in contrast to Kizilcec's (2016) experiment, understandable text for all explanation conditions. Recall that the actual Intelligent Crime Reporting Tool (ICRT) of the Dutch police, on which our experiment is based, arrives at recommendations via an inherently interpretable, rule-based algorithm, which is based on legal and policy rules of the police (Borg & Bex, 2021; Odekerken et al., 2022). The rule-based algorithm makes it possible to provide rationale explanations in terms of these rules and the input provided by the citizen, making these explanations understandable for the average citizen. This might be a reason why our results differ from those of Kizilcec (2016), whose rationale explanation consisted of more complex statistical calculations of students' grades and possible biases. Thus, the challenge of explanation will become more pronounced in systems that use less transparent AI techniques such as deep learning or machine learning (Grimmelikhuijsen & Meijer, 2022). The type of explanations for such systems (if any) are very often based on statistical patterns that are difficult to understand for lay users such as citizens.

Furthermore, as our explanations were designed with the police based on real-world situations, they might have been less distinctive as explanations used in experiments with fictional AI tools. This could have been a reason why not all participants correctly identified the type of explanation they read. A sensitivity analysis, however, showed that excluding participants that failed the factual manipulation check did not alter our results. Alternatively, and in line with our interpretation of the results, participants'

incorrect responses could have also been due to the fact that individuals generally have limited time and capacity for information processing and therefore simplify the information provided to them in order to make decisions. Having any type of explanation (in comparison with no explanation) would have been enough to change their attitudes and behaviors. Moreover, it has been proven difficult to make a strict distinction between what is a directive explanation and what is not.

In the second study, we targeted what we believed were the desired outcomes of citizens who reported their case of online fraud, that is, specific actions they needed to take to achieve financial compensation. Nevertheless, there was an element in the rationale explanation that could have been interpreted as a directive, since it mentioned that the quality mark mediated the dispute with the web shop. To the best of our knowledge, this research is the first to empirically examine directive explanations for public sector algorithmic systems. Future researchers may learn from the limitations of our study and further specify and test directive explanations in other settings. In any case, examining these different types of explanations for public sector algorithmic recommender systems is necessary to better understand their effects in different settings.

Second, we used a mock version of the tool in our experiments with a random sample of citizens, rather than those who had actually been victims of internet fraud and used the police's tool. Despite this, we consistently focused on safeguarding high mundane realism by closely mimicking a real chatbot in our experimental scenario. This experimental scenario was developed together with police employees to make the vignettes as realistic as possible. In addition, we asked participants how well they could empathize with the fraud situation they were given. Participants in both experiments scored relatively high on this question: $M=5.46$ in Study 1 and $M=5.49$ in Study 2 on a 7-point Likert-scale. This suggests that even though hypothetical, our experiment was highly realistic to participants. Evaluations of the actual crime reporting tool performed by the police both support and counter some of our findings: while a large number of citizens trusted the recommendation not to file a report, there is an equally large group that ignored the (explained) recommendation and still filed an official report even when recommended not to. Of course, these people paid actual money for a product they did not receive or did not like, so it can be imagined that they were more emotionally invested in the case than our participants. In any case, in future research we may study real end-users of an algorithmic recommender system in, for instance, a field experimental setting.

Third, we focused on one specific system in one organization, so the question of how well our findings transfer to other contexts remains. However, we expect that other public organizations deal with similar

questions. Usually, citizens do not have a realistic exit option for public services, and recommendations may therefore be much more forceful than in private-sector organizations. In addition, public organizations are generally subject to greater demands for transparency and accountability than commercial enterprises, which leads to stronger demands on explainability of algorithms in government (Busuioc, 2020; Meijer & Grimmelikhuijsen, 2020). The findings of this study may therefore be helpful for public organizations that deal with similar explainability questions. The type of service the ICRT provides—allowing citizens to seek justice, i.e., get advice on and report on a possible crime—is perhaps more abstract than in cases where the citizen directly requests something of more tangible value from the government, such as a residence permit or welfare benefits. Additionally, the type of algorithmic public service examined in this paper, assisting citizens in detecting fraudulent activities of online vendors, is a beneficial service, instead of a service that could potentially violate civil rights, such as a risk assessment tool used for bail decisions. It is important to consider the potential implications of the findings when applying them to different situations or contexts (Aoki, 2020). Nevertheless, upholding the law in an effective and transparent way and giving its citizens access to justice are among the core tasks of good government (Holmberg et al., 2009), and the police have every reason not to damage citizens' trust in them. That said, we encourage scholars to empirically examine the effects of providing explanations for algorithmic recommendations in other scenarios where other aspects of government (e.g., access to health care) and other type of services (e.g., requesting welfare benefits or predictive policing) play a role.

Furthermore, and expanding on the points made in the preceding paragraph, there might be a context-specific nature to the effectiveness of different types of explanations in enhancing trust (Aoki, 2020; Grimmelikhuijsen, 2023; Schiff et al., 2022). It is important to acknowledge that trust dynamics can vary significantly across countries and regions. In the Netherlands, known for its high levels of trust in both law enforcement and government overall (European Commission, Brussels, 2020), this elevated baseline trust might impact the reception of explanations of police algorithms differently compared to countries with lower levels of baseline trust in the police. Nevertheless, in countries with low trust in the police, procedural justice becomes crucial, as it shapes attitudes toward the entire criminal justice system (Nix et al., 2015). While procedural explanations for police algorithms may help in low-trust regions, the evidence about the effectiveness of procedural justice in police work is mixed, with some studies suggesting potential adverse effects (Murphy, 2017). This fits with our findings from Study 2, where procedural explanations had a less pronounced effect compared with rationale explanations. Hence, additional

research is necessary to comprehend the applicability of our findings in various contexts, particularly those characterized by low trust in police.

Conclusion

We investigated the question *What are the effects of explanations on citizen trust in algorithmic recommendations?* by conducting a sequential factorial design with two consecutive survey experiments. On the basis of our findings, we draw two main conclusions. First, explanations—in general—have a significant and positive effect on all dimensions of citizen trust. Explanations increase citizens' trusting beliefs, enhance their intention to act accordingly, and even increase the likelihood that they will change their behavior to follow up on algorithmic recommendations in digital public service delivery. The second main conclusion is that the type of explanation does not seem to matter in terms of the level of citizen trust. Providing information about a decision procedure, the rationale behind a decision, or a combination of these two had no distinguishable effect. This may suggest that trust increases when citizens see that governments make an effort to provide an explanation, regardless of the nature of this explanation.

At the same time, the post-hoc analyses add some nuance to the conclusions. These exploratory analyses showed that the devil might be in the details. If no directive explanation—listing which specific actions an individual needs to take to achieve their desired outcome—is provided, having a rationale element, explaining why a specific action has been recommended, seems to be more effective in strengthening citizen trust. Thus, a rationale explanation or a combined explanation are possibly more powerful than a procedural explanation when no directive explanation is provided.

Finally, the post-hoc analyses provided evidence that a directive explanation has an effect only on a procedural explanation. Adding a directive explanation, that specifies which actions the individual should perform to obtain their desired outcome (if possible), to a procedural explanation leads to an increase in trust. If explaining why a citizen receives a specific recommendation is not possible or feasible — for example, to prevent “gaming the system” (such as tax avoidance) or to protect national security — providing a procedural explanation with a directive element can increase citizen trust.

Notes

1. See <https://aangifte.politie.nl/iaai-preintake/#/>.
2. Open Science Foundation Registration: https://osf.io/d862z/?view_only=59da5dfefa934dccb39ad73a0cdcae74

3. We used p -values with a significance level of $\alpha = 0.05$ for all tests. However, we used a Bonferroni adjustment to correct for an inflated chance of Type 1 errors due to multiple testing. For each outcome variable, we conducted six comparisons between groups. Therefore, in our analysis we multiplied every p -value by six. G*Power, however, doesn't allow adjustment of the p -value, so we performed the power analysis with the equivalent adjustment of α (.05) divided by six.
4. The quality mark that the ICRT refers to, claims to be the biggest web shop quality mark in the Netherlands and Europe (WebshopKeurmerk, [n.d.](https://www.webshopkeurmerk.nl/)). It is therefore likely to assume that participants have heard of this quality mark before.
5. Open Science Foundation Registration: https://osf.io/hyuxm/?view_only=377a4f29c7ab414d897652fd2f789180; The terminology has been changed from "action perspective" to "directive explanation."

Acknowledgements

The authors would like to thank two anonymous reviewers for their constructive feedback. In addition, we thank Ines Mergel for her valuable feedback at the European Group of Public Administration Conference. Furthermore, we thank the following people for their comments on previous versions of the paper: Barbara Vis, Lars Brummel, Resie Hoeijmakers, Noortje de Boer, Bert George and Benjamin Tidå.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Ethical review

Study 1 and Study 2 have been approved by the Faculty's Ethical Review Committee of Utrecht University.

Funding

This work was supported by the Dutch Research Council under Grant 406.DI.19.011. The funding source was not involved in any phase of this research.

Preregistration reports

Study 1: Open Science Foundation Registration; <https://osf.io/d862z>

Study 2: Open Science Foundation Registration; <https://osf.io/hyuxm>

Notes on contributors

Esther Nieuwenhuizen is a PhD-candidate at the Utrecht University School of Governance (The Netherlands). She mainly focuses on the influence of transparency on citizen trust in algorithm use by the police. Her research is part of a project on responsible and trustworthy algorithmic policing in the Netherlands (ALGOPOL).

Albert Meijer is a full professor of public innovation at the Utrecht School of Governance and chair of the Public Governance and Management Group. His research focuses on

public administration in an information age and he has published on topics as diverse as public innovation, smart cities, democracy and social media in the public sector.

Floris Bex is full professor of Data Science and the Judiciary at the Tilburg Institute for Law, Society and Technology (Tilburg University), as well as associate professor of AI and scientific director of the National Police Lab AI at Utrecht University. His research is on AI & Law, developing new AI techniques for legal and police professionals, and investigating the legal aspects of AI technologies.

Stephan Grimmelikhuijsen is an associate professor at the Utrecht University School of Governance, The Netherlands. His research centers on technology in government, citizen-state interactions and behavioral public administration.

ORCID

E. N. Nieuwenhuizen  <http://orcid.org/0000-0003-0199-8848>

A. J. Meijer  <http://orcid.org/0000-0001-8532-7894>

F. J. Bex  <http://orcid.org/0000-0002-5699-9656>

S. G. Grimmelikhuijsen  <http://orcid.org/0000-0002-1553-6065>

Data availability statement

We chose to make the data underlying the two studies open prior submission. We deposited the data at the Open Science Foundation. The data can be accessed through this URL: https://osf.io/ksa8h/?view_only=1271438d14274efb8e6b701cabea0273

References

- 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. <https://doi.org/10.1093/jopart/muz013>
- Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973–989. <https://doi.org/10.1177/1461444816676645>
- Androutsopoulou, A., Karacapilidis, N., Loukis, E., & Charalabidis, Y. (2019). Transforming the communication between citizens and government through AI-guided chatbots. *Government Information Quarterly*, 36(2), 358–367. <https://doi.org/10.1016/j.giq.2018.10.001>
- Aoki, N. (2020). An experimental study of public trust in AI chatbots in the public sector. *Government Information Quarterly*, 37(4), 101490. <https://doi.org/10.1016/j.giq.2020.101490>
- Arrieta, A., Barredo, Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Bauhr, M., & Grimes, M. (2012). *What is government transparency? New measures and relevance for quality of government*. 1–27.

- Borg, A., & Bex, F. (2021). Explaining Arguments at the Dutch National Police'. In: *AI Approaches to the Complexity of Legal Systems XI-XII (Lecture Notes in Computer Science)*. Ed. by Víctor RodríguezDoncel, Monica Palmirani, Michał Araszkiewicz, Pompeu Casanovas, Ugo Pagallo, and Giovanni Sartor. Cham: Springer International Publishing. 183–197.
- Bowman, S., DeHaven, A., Errington, T., Hardwicke, T. E., Mellor, D. T., Nosek, B. A., & Soderberg, C. K. (2020). *OSF Prereg template*. MetaArXiv <https://doi.org/10.31222/osf.io/epgjd>
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 1–12. <https://doi.org/10.1177/2053951715622512>
- Busuioc, M. (2020). Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review*, 81(5), 825–836. <https://onlinelibrary.wiley.com/doi/full/10.1111/puar.13293>
- Chen, T., & Gasco-Hernandez, M. (2024). Uncovering the results of AI Chatbot use in the public sector: Evidence from US state governments. *Public Performance & Management Review*, 1–26. https://www.tandfonline.com/doi/full/10.1080/15309576.2024.2389864?casa_token=eiQgkqWa1-QAAAAA%3AOAW9QbRsr-N4uR5VviHFinQ--IXhTjacPo4sa0bjbe1rvy8PwjxW0rJovS_sCuVi1DByE5FB49Z16#abstract
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Cramer, H., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, N., Aroyo, L., & Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, 18(5), 455–496. <https://doi.org/10.1007/s11257-008-9051-3>
- Cross, C. (2018). Victims’ motivations for reporting to the ‘fraud justice network. *Police Practice and Research*, 19(6), 550–564. <https://doi.org/10.1080/15614263.2018.1507891>
- de Fine Licht, J., Naurin, D., Esaiasson, P., & Gilljam, M. (2014). When Does Transparency Generate Legitimacy? Experimenting on a Context-Bound Relationship: When Does Transparency Generate Legitimacy? *Governance*, 27(1), 111–134. <https://doi.org/10.1111/gove.12021>
- de Fine Licht, K., & de Fine Licht, J. (2020). Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy. *AI & Society*, 35(4), 917–926. <https://doi.org/10.1007/s00146-020-00960-w>
- DeVellis, R. F. (2017). *Scale development: Theory and applications*. (Fourth edition) SAGE.
- Dunleavy, P., Margetts, H., Bastow, S., & Tinkler, J. (2006). New public management is dead—Long live digital-era governance. *Journal of Public Administration Research and Theory*, 16(3), 467–494. <https://academic.oup.com/jpart/article-abstract/16/3/467/934257?redirectedFrom=fulltext>
- European Commission, Brussels. (2020). *Eurobarometer 92.4 (2019)Eurobarometer 92.4 (2019): Attitudes of European citizens toward the Environment, Corruption, and Attitudes toward the impact of digitalization on daily lives: Attitudes of European citizens toward the Environment, Corruption, and Attitudes toward the impact of digitalization on daily lives (1.0.0)* [dataset]. GESIS Data Archive. <https://doi.org/10.4232/1.13652>
- Fatima, S., Desouza, K. C., & Dawson, G. S. (2020). National strategic artificial intelligence plans: A multi-dimensional analysis. *Economic Analysis and Policy*, 67, 178–194. <https://doi.org/10.1016/j.eap.2020.07.008>
- Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, 26(6), 3333–3361. <https://doi.org/10.1007/s11948-020-00276-4>

- Giest, S., & Grimmelikhuijsen, S. (2020). Introduction to special issue algorithmic transparency in government: Towards a multi-level perspective. *Information Polity*, 25(4), 409–417. <https://doi.org/10.3233/IP-200010>
- 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. <https://doi.org/10.1111/puar.13483>
- Grimmelikhuijsen, S., & Meijer, A. J. (2014). Effects of transparency on the perceived trustworthiness of a government organization: Evidence from an online experiment. *Journal of Public Administration Research and Theory*, 24(1), 137–157. <https://doi.org/10.1093/jopart/mus048>
- Grimmelikhuijsen, S., & Meijer, A. J. (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. <https://doi.org/10.1093/ppmgov/gvac008>
- Gulati, S., Sousa, S., & Lamas, D. (2019). Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology*, 38(10), 1004–1015. <https://doi.org/10.1080/0144929X.2019.1656779>
- Halachmi, A., & Greiling, D. (2013). Transparency, E-government, and accountability: some issues and considerations. *Public Performance & Management Review*, 36(4), 572–584. <https://doi.org/10.2753/PMR1530-9576360404>
- Hind, M. (2019). Explaining explainable AI. *XRDS: Crossroads, The ACM Magazine for Students*, 25(3), 16–19. <https://doi.org/10.1145/3313096>
- Hirschman, A. O. (1970). *Exit, voice, and loyalty: Responses to decline in firms, organizations, and states*. (Vol. 25) Harvard university press.
- Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3), 407–434. <https://doi.org/10.1177/0018720814547570>
- Holmberg, S., Rothstein, B., & Nasiritousi, N. (2009). Quality of government: What you get. *Annual Review of Political Science*, 12(1), 135–161. <https://doi.org/10.1146/annurev-polisci-100608-104510>
- Hosseini, M., Shahri, A., Phalp, K., & Ali, R. (2015). *Towards engineering transparency as a requirement in socio-technical systems* [Paper presentation]. 2015 IEEE 23rd International Requirements Engineering Conference (RE), 268–273. <https://doi.org/10.1109/RE.2015.7320435>
- Kim, T. W., & Routledge, B. R. (2018). *Informational privacy, a right to explanation, and interpretable AI* [Paper presentation]. 2018 IEEE Symposium on Privacy-Aware Computing (PAC) (pp. 64–74). <https://doi.org/10.1109/PAC.2018.00013>
- Kizilcec, R. F. (2016). *How much information?: Effects of transparency on trust in an algorithmic interface* [Paper presentation]. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 2390–2395). <https://doi.org/10.1145/2858036.2858402>
- Klopfenstein, L. C., Delpriori, S., Malatini, S., & Bogliolo, A. (2017). *The rise of bots: A survey of conversational interfaces, patterns, and paradigms* [Paper presentation]. Proceedings of the 2017 Conference on Designing Interactive Systems (pp. 555–565).
- Lacave, C., & Diez, F. J. (2004). A review of explanation methods for heuristic expert systems. *The Knowledge Engineering Review*, 19(2), 133–146. <https://doi.org/10.1017/S0269888904000190>
- Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology*, 31(4), 611–627. <https://doi.org/10.1007/s13347-017-0279-x>

- Lindgren, I., Madsen, C. Ø., Hofmann, S., & Melin, U. (2019). Close encounters of the digital kind: A research agenda for the digitalization of public services. *Government Information Quarterly*, 36(3), 427–436. <https://doi.org/10.1016/j.giq.2019.03.002>
- Madsen, M., Gregor, S. (2000). Measuring human-computer trust. *Proceedings of the 11 Th Australasian Conference on Information Systems* (pp. 6–8).
- Makasi, T., Tate, M., Desouza, K. C., & Nili, A. (2021). Value-based guiding principles for managing cognitive computing systems in the public sector. *Public Performance & Management Review*, 44(4), 929–959. <https://doi.org/10.1080/15309576.2021.1879883>
- Mansbridge, J. (2009). A “selection model” of political representation*. *Journal of Political Philosophy*, 17(4), 369–398. <https://doi.org/10.1111/j.1467-9760.2009.00337.x>
- Matook, S., Brown, S. A., & Rolf, J. (2015). Forming an intention to act on recommendations given via online social networks. *European Journal of Information Systems*, 24(1), 76–92. <https://doi.org/10.1057/ejis.2013.28>
- McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, 13(3), 334–359. <https://doi.org/10.1287/isre.13.3.334.81>
- Mei, H., & Zheng, Y. (2024). How M-government services build relative trust? The mediating roles of value creation and risk perception. *Public Performance & Management Review*, 47(6), 1327–1355. <https://doi.org/10.1080/15309576.2024.2370935>
- Meijer, A., & Grimmelikhuijsen, S. (2020). Responsible and accountable algorithmization: How to generate citizen trust in governmental usage of algorithms. In *The algorithmic society* (pp. 53–66). Routledge.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 205395171667967. <https://doi.org/10.1177/2053951716679679>
- Moody, G. D., Galletta, D. F., & Lowry, P. B. (2014). When trust and distrust collide online: The engenderment and role of consumer ambivalence in online consumer behavior. *Electronic Commerce Research and Applications*, 13(4), 266–282. <https://doi.org/10.1016/j.elerap.2014.05.001>
- Moore, M. H. (1995). *Creating public value: Strategic management in government*. Harvard university press.
- Murphy, K. (2017). Challenging the ‘invariance’ thesis: Procedural justice policing and the moderating influence of trust on citizens’ obligation to obey police. *Journal of Experimental Criminology*, 13(3), 429–437. <https://doi.org/10.1007/s11292-017-9298-y>
- Nai, R., Meo, R., Morina, G., & Pasteris, P. (2023). Public tenders, complaints, machine learning and recommender systems: A case study in public administration. *Computer Law & Security Review*, 51, 105887. <https://doi.org/10.1016/j.clsr.2023.105887>
- Nix, J., Wolfe, S. E., Rojek, J., & Kaminski, R. J. (2015). Trust in the police: The influence of procedural justice and perceived collective efficacy. *Crime & Delinquency*, 61(4), 610–640. <https://doi.org/10.1177/0011128714530548>
- Norris, P. (2022). *In praise of skepticism: Trust but verify*. Oxford University Press.
- Nothdurft, F., Richter, F., & Minker, W. (2014). *Probabilistic human-computer trust handling* [Paper presentation]. Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL) (pp. 51–59). <https://doi.org/10.3115/v1/W14-4307>
- Odekerken, D., Bex, F., Borg, A., & Testerink, B. (2022). Approximating stability for applied argument-based inquiry. *Intelligent Systems with Applications*, 16, 200110. <https://doi.org/10.1016/j.iswa.2022.200110>

- Ojo, A., Mellouli, S., & Ahmadi Zeleti, F. (2019). *A realist perspective on AI-era public management** [Paper presentation]. 20th Annual International Conference on Digital Government Research (pp. 159–170). <https://doi.org/10.1145/3325112.3325261>
- Petty, R. E., & Briñol, P. (2011). *The elaboration likelihood model in: Handbook of theories of social psychology*. SAGE Publications Ltd.
- Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In *Communication and persuasion* (pp. 1–24) Springer.
- Quijano-Sánchez, L., Cantador, I., Cortés-Cediel, M. E., & Gil, O. (2020). Recommender systems for smart cities. *Information Systems*, 92, 101545. <https://doi.org/10.1016/j.is.2020.101545>
- Rader, E., Cotter, K., & Cho, J. (2018). *Explanations as mechanisms for supporting algorithmic transparency* [Paper presentation]. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1–13). <https://doi.org/10.1145/3173574.3173677>
- Rothstein, B. O., & Teorell, J. A. (2008). What is quality of government? A theory of impartial government institutions. *Governance*, 21(2), 165–190. <https://doi.org/10.1111/j.1468-0491.2008.00391.x>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Rudin, C., Wang, C., & Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2(1). <https://assets.pubpub.org/0nx9n90c/61c6ac06-6f37-4f64-ba96-a3b8df8cf9b6.pdf>
- Schiff, D. S., Schiff, K. J., & Pierson, P. (2022). Assessing public value failure in government adoption of artificial intelligence. *Public Administration*, 100(3), 653–673. <https://doi.org/10.1111/padm.12742>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*.
- Singh, R., Dourish, P., Howe, P., Miller, T., Sonenberg, L., Velloso, E., Vetere, F. (2021a). Directive explanations for actionable explainability in machine learning applications. *ArXiv:2102.02671* [Cs]. <http://arxiv.org/abs/2102.02671>
- Singh, R., Ehsan, U., Cheong, M., Riedl, M. O., Miller, T. (2021b). LEx: A framework for operationalising layers of machine learning explanations. *ArXiv:2104.09612* [Cs]. <http://arxiv.org/abs/2104.09612>
- Sniderman, P. M. (2018). Some advances in the design of survey experiments. *Annual Review of Political Science*, 21(1), 259–275. <https://doi.org/10.1146/annurev-polisci-042716-115726>
- Tintarev, N., & Masthoff, J. (2012). Evaluating the effectiveness of explanations for recommender systems: Methodological issues and empirical studies on the impact of personalization. *User Modeling and User-Adapted Interaction*, 22(4-5), 399–439. <https://doi.org/10.1007/s11257-011-9117-5>
- Tyler, T. R. (1990). *Why people obey the law: Procedural justice, legitimacy, and compliance*. Yale University Press.
- Vidotto, G., Massidda, D., Noventa, S., & Vicentini, M. (2012). Trusting beliefs: A functional measurement study. *Psicologica: International Journal of Methodology and Experimental Psychology*, 33(3), 575–590.
- Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. *Science Robotics*, 2(6), eaan6080. <https://doi.org/10.1126/scirobotics.aan6080>

- Wang, W., & Benbasat, I. (2007). Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems*, 23(4), 217–246. <https://doi.org/10.2753/MIS0742-1222230410>
- WebshopKeurmerk (n.d). *Waarom velen kiezen voor Webshop Keurmerk?*. Retrieved November 15, 2022, from <https://www.keurmerk.info/nl/home>
- Welch, E. W., Hinnant, C. C., & Moon, M. J. (2005). Linking citizen satisfaction with e-government and trust in government. *Journal of Public Administration Research and Theory*, 15(3), 371–391. <https://academic.oup.com/jpart/article-abstract/15/3/371/941130?login=false>
- Wieringa, M. (2020). *What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability* [Paper presentation]. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 1–18). <https://doi.org/10.1145/3351095.3372833>
- Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, 41(1), 118–132. <https://doi.org/10.1177/0162243915605575>
- Zuiderwijk, A., Chen, Y.-C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly*, 38(3), 101577. <https://doi.org/10.1016/j.giq.2021.101577>