Received: 9 March 2021 | Revised: 8 February 2022 | Accepted: 8 February 2022

DOI: 10.1111/puar.13483

RESEARCH ARTICLE

PAR PUBLIC ADMINISTRATION REVIEW ASPA

Explaining Why the Computer Says No: Algorithmic Transparency Affects the Perceived Trustworthiness of Automated Decision-Making

Stephan Grimmelikhuijsen

Utrecht University School of Governance

Funding information
Netherlands Organization for Scientific Research (NWO), Grant/Award Number: VENI-451-15-024

Abstract

Algorithms based on Artificial Intelligence technologies are slowly transforming street-level bureaucracies, yet a lack of algorithmic transparency may jeopardize citizen trust. Based on procedural fairness theory, this article hypothesizes that two core elements of algorithmic transparency (accessibility and explainability) are crucial to strengthening the perceived trustworthiness of street-level decision-making. This is tested in one experimental scenario with low discretion (a denied visa application) and one scenario with high discretion (a suspicion of welfare fraud). The results show that: (1) explainability has a more pronounced effect on trust than the accessibility of the algorithm; (2) the effect of algorithmic transparency not only pertains to trust in the algorithm itself but also—partially—to trust in the human decision-maker; (3) the effects of algorithmic transparency are not robust across decision context. These findings imply that transparency-as-accessibility is insufficient to foster citizen trust. Algorithmic explainability must be addressed to maintain and foster trustworthiness algorithmic decision-making.

Evidence for Practice

- Algorithmic transparency consists of accessibility and explainability.
- This study finds that accessibility, though important, is not sufficient to foster trust.
- Governments must address the explainability of algorithmic decision-making to earn citizen trust in algorithms and bureaucrats working with them.

INTRODUCTION

On June 25 and July 7, 2018, the City of Rotterdam used a system called SyRI (*Systeem Risico Indicatie*, or: "System Risk Indication") to carry out a risk analysis of welfare fraud on 12,000 addresses in a deprived neighborhood. The risk analysis used an algorithm that was fed by 17 datasets containing personal data on someone's fiscal, residential, educational, and labor situation. The city never published the algorithm's parameters and decision rules, nor were investigated residents informed they were investigated for welfare fraud. Residents and activists protested and finally, in 2020, a Dutch Court prohibited governments to use SyRI. A core reason for this, according to the verdict, was a lack of transparency of the algorithm used by this system.

The example above highlights the profound implications of automated decision-making and decision assistance in street-level bureaucracies. Where past automation replaced the need for human interference in high-volume, relatively simple, decisions with little discretion (Bovens and Zouridis 2002), a new generation of algorithmic applications under the umbrella of artificial intelligence (AI) is targeted to automate medium and high discretionary decisions, which are set to affect access to and apportioning of government resources (Young, Bullock and Lecy 2019; Zouridis, Van Eck and Bovens 2020). For individual bureaucrats, this means that their decisions are increasingly being steered and disciplined by refined computer systems, which will eventually affect how bureaucrats interact with individual citizens

(Peeters 2020; Peeters & Widlak, 2018). Other authors highlight that the introduction of algorithms in public organizations are altering organizational structures, routines, and culture (Meijer, Lorenz and Wessels 2021; Vogl et al. 2020) and requires different competencies from organizational leaders (Coulthart and Riccucci 2021).

While some emphasize the potential of such far-reaching automated decision-making to make government services more equitable and efficient (e.g. Pencheva et al. 2020), others have heavily criticized this for producing biased and even discriminatory predictions because of biased model parameters and/or biased data (Eubanks 2018; O'Neil 2016). Often, human biases are consciously and unconsciously automated and integrated into automated decision-making.

A criticism underlying these potential biases is a lack of algorithmic transparency and ultimately accountability (Busuioc 2020; Meijer and Grimmelikhuijsen 2020). First, a new generation of algorithms uses techniques to detect patterns in data using only inputs (e.g. a training dataset provided by humans). How certain patterns and outputs based on these input data are generated has been referred to as an algorithmic 'black box'. Such algorithms are not readily understandable to humans, making them *unexplainable* to citizens (Burrell 2016). Second, algorithms are sometimes deliberately made *inaccessible* as they are often developed by commercial parties and subject and protected by intellectual property. Other algorithms are not accessible because governments fear that citizens subject to those algorithms will game the system once they have figured out how it works (Mittelstadt et al. 2016).

The lack of algorithmic transparency in street-level bureaucracies specifically raises concerns about the trustworthiness of bureaucratic decision-making in which these algorithms play a role (Došilović et al. 2018; Widlak and Peeters 2018). An elaborate body of literature on procedural fairness shows that decisions that are not well-explained or not open to comment are less acceptable and decrease trust in the decision-maker (e.g. Lind, Kanfer and Early 1990; Tyler 2006). Inaccessible and inexplainable algorithms may therefore erode trust and this has led computer scientists to put algorithmic transparency central as a means towards trustworthy algorithms (Miller 2019; Rudin 2019). Trustworthy algorithms are crucial for citizens specifically as they are becoming increasingly dependent on these algorithms for the provision of crucial services, such as welfare, reporting crimes or applying for a visa extension. Unlike most algorithms in the private sector, citizens often have no choice other than to trust that an algorithm treats them fairly.

While the link between algorithmic transparency and trustworthiness seems straightforward, others argue that "seeing inside a system does not necessarily mean understanding its behavior or origins" (Annany and Crawford 2018, 980). Similarly, others argue that it is also relevant, from a citizen perspective, to provide explanations, which might help to maintain legitimacy (De Fine Licht and De Fine Licht 2020). In other words, when algorithmic transparency is merely implemented as "access to code" this is important to have accountability, yet this is unlikely to increase people's understanding or perceived trustworthiness of algorithmic decision-making.

Scholars in public administration have also looked at how citizens view algorithmic versus human decision-making. Recent studies have investigated the effect of perceived fairness of automated versus human decision-making. Especially for complex tasks where "human" skills are deemed important, such as hiring and work evaluation, algorithms are met with more suspicion by the public (Lee 2018; Nagtegaal 2021). Street-level bureaucracy research specifically highlights that AI-powered automation renders new concerns about the trustworthiness of the decision-making process (Bullock et al. 2020; Peeters 2020). For instance, decisions may become less tailored in a way that does justice to circumstances unique to an individual citizen, so-called *Einzelfallgerechtigkeit* ("justice to an individual case"). This has initiated a debate on whether automation curtails human discretion in street-level decision-making too much (Buffat 2015).

This study ventures beyond the decision on whether to (partially) automate street-level decisions or not, by focusing on how algorithmic decision-making can be designed once they are implemented in practice. Because of various waves of automation in street-level decision-making, algorithms are already part and parcel of many street-level bureaucracies (e.g. Bovens and Zouridis 2002). Furthermore, many street-level decisions are not purely "algorithm" or neither are they purely "human"; in many cases, automation supports or supplants only part of a decision-making process (Young, Bullock and Lecy 2019; Zouridis, Van Eck and Bovens 2020). Therefore, this article builds on previous work by testing various elements of algorithmic transparency in human-machine interaction, rather than exploring the effects of algorithmic versus human decision-making (e.g. Nagtegaal 2021; Schiff, Schiff and Pierson 2021).

The importance of algorithmic transparency is often highlighted as a mechanism to ensure trustworthy algorithms, yet this hypothesized effect is debated and has been little tested empirically in a public administration context. Furthermore, this study will provide a more refined test as it will conceptually distinguish—and empirically test—the effect of both accessibility and explainability on the perceived trustworthiness of automated decision-making. More systematic, and more in-depth empirical research in our field is needed. The following research question is central to this article:

What is the effect of algorithmic transparency on the perceived trustworthiness of automated decision-making?

To answer this question, I developed two related scenario-based survey experiments. Both experiments employed a between-subjects 2×2 factorial design. Each factor independently varies one particular dimension of algorithmic transparency: explainability and accessibility.

The first scenario described the automated decision of a visa application. Such automated decision-making can be considered a "classic" form of automation where discretionary power is relatively low (Bovens and Zouridis 2002). In contrast, the second scenario entailed a street-level bureaucrat who used an algorithm to predict welfare fraud, which reflects a predictive algorithm that is typical of the more controversial AI applications in government. This also reflects that relatively new ways in which AI algorithms change street-level bureaucracy: not just low discretionary decisions are automated but also medium to high discretionary decisions are being affected by automation (e.g. Young, Bullock and Lecy 2019).

First, the findings provide support for a positive effect of explainability and a limited positive effect of accessibility on the perceived trustworthiness of algorithms. Second, algorithmic transparency also affects the trustworthiness of bureaucrats who use algorithmic systems to support their decision-making. Third, the effect of algorithmic transparency is not robust across the two study contexts. In the visa application experiments, only explainability had an effect, whereas in the welfare fraud experiment both accessibility and explainability affect the perceived trustworthiness of the algorithm. I will explore possible explanations and provide recommendations for future research.

ALGORITHMIC TRANSPARENCY AND STREET-LEVEL BUREAUCRACY

An algorithm in a purely mathematical sense can be defined simply as a finite set of rules that provide a sequence to solve a certain problem (Hill 2015). Relatively simple "IF-THEN" algorithms have affected bureaucracies for decades. Since the 1990s government agencies have automated various types of bureaucratic decisions. For instance, the Dutch executive agency for education (DUO) decides on whether students receive a government-subsidized loan and the height of that loan. Millions of decisions are generated almost completely automated, transforming agencies into "system-level bureaucracies," indicating that discretion is no longer executed at the level of the individual civil servant, but at the level of developers of computer systems (Bovens and Zouridis 2002).

One key feature that distinguishes these older algorithmic applications from their more recent counterparts, however, is that they have been programmed and modeled by human beings and have—or are at least supposed to have—programmed and transparent decision rules that are traceable and explainable. Machine learning algorithms, on the other hand, have the potential to "learn" from vast amounts of data without being explicitly programmed by human beings (Meijer and Grimmelikhuijsen 2020).

A combination of dynamic big data in interplay with self-learning algorithms yields many possible applications. For instance, various countries use algorithms to detect patterns between data on financial behavior (e.g. expenses), living area, family composition and welfare benefits to indicate the risk of fraudulent behavior. High expenses, while somebody is living alone in a cheap apartment, might indicate that this person did not report income to the welfare agency (Zouridis, Van Eck and Bovens 2020). In general, these applications are fed by large amounts of data (big data) that capture our everyday behavior—online and offline—and these data are much more fine-grained and dynamic than in the past. These data are then used to derive small patterns and correlations from this data (Janssen and Van den Hoven 2015).

As these algorithms permeate street-level bureaucracies, their use is being increasingly criticized. In particular, various types of machine learning algorithms can be "black boxes," whose internal parameters or data are either unknown or uninterpretable to a human observer (Guidotti et al. 2018, 5). Black boxes can be acceptable if there are little to no consequences attached to the application of the algorithm. However, especially at the street-level decision-making is often highly consequential and algorithms are increasingly used to automate human discretion in such decisions (Young, Bullock and Lecy 2019).

In response, scholars in data ethics (e.g. Lepri et al. 2018), computer science (Miller 2019) and more recently in public administration (Busuioc 2020), have called for more transparent algorithms to render them more accountable. The literature on algorithmic transparency generally considers two elements: *accessibility* and *explainability* (Giest and Grimmelikhuijsen 2020) which are summarized in table 1.

These two elements also fit with more generic definitions of government transparency. A systematic review by Cucciniello et al. (2017) found that most definitions address the availability of information of decision-making processes, budgets, operations, or performance of governmental bodies. For example, Grimmelikhuijsen and Meijer (2014) include access to information in a much-used definition of government transparency, and De Fine

TABLE 1 Core Dimensions of Algorithmic Transparency

	Description
Accessibility	Public availability of source code, model and/or data. This study focuses on access by external experts who can inspect and analyze an algorithm for bias and functionality.
Explainability	The outcomes of an algorithm can be explained in a way a human can understand how or why an algorithmic decision was reached. This study focuses on publishing underlying reasons of a decision.

Licht (2014) highlights that explaining the rationale behind government decision-making is a crucial element of transparency. Building on this generic transparency literature, I define algorithmic transparency by explicitly including explainability and accessibility: Algorithmic transparency is achieved when *external actors can access the underlying data and code of an algorithm and the outcomes produced by it are explainable in a way a human being can understand*.

First, we consider the *accessibility* of algorithms. Accessibility is an important issue because often the source code or model is not available to outsiders as it is considered the intellectual property of the company that developed it (Mittelstadt et al. 2016). Also, the algorithm might be kept inaccessible for privacy reasons or purportedly to prevent users from gaming the system (Burrell 2016; Kitchin 2017).[1] This is risky since inaccessible algorithms are inscrutable and can more easily produce decisions or recommendations that are biased, discriminatory, or inaccurate.

Accessibility of algorithms goes beyond being accessible to the public since to most of the public it will be unclear what they are looking at when they see code (Annany and Crawford 2018). Even experts struggle to understand what software code will do in practice, as the source code only gives very limited information to predict how a computer system will behave in practice (Kroll et al. 2016). Specifically, the source code of an unsupervised machine learning algorithm does not tell ordinary citizens much because a pattern emerges automatically from the specific data under analysis, occasionally in ways that no human can explain (Burrell 2016).

Hence, transparency as merely having a publicly accessible algorithm will not be sufficient to improve an algorithm's trustworthiness. Therefore, on top of transparency-as-code-availability, scholars have argued for algorithms to be audited by an independent auditor (Tutt 2017) or to include technical tools that are incorporated in a system's design to ensure an algorithm complies with legal procedures and standards (Kroll et al. 2016). In this paper, accessibility means not just public availability, but accessibility means that external independent experts can access an algorithm for inspection and analysis to assess if it is compliant and does not violate any rules.

Finally, accessibility does not only concern the algorithm itself, but also the underlying data. Using single or combined high-volume structured and unstructured datasets ("big data") algorithms can detect new patterns. However, linking various fine-grained datasets has led to privacy concerns (Mergel, Rethemeyer and Isett 2016). Furthermore, a large body of research had pointed out that many datasets are biased, which leads to biased predictions of an algorithm (Eubanks 2018). To be able to assess an algorithm the quality of underlying data, these must be accessible to independent experts.

The second core element of algorithmic transparency is *explainability*. A lack of explainable models and outcomes is at the heart of the debate on the algorithmic black box explained earlier (Burrell 2016; Lepri et al. 2018; Miller 2019). Different techniques have been put forward to increase the explainability of AI algorithms.[2] They could be developed to be inherently transparent in their decisions (Rudin 2019), or if the algorithm is a black-box, then various explanation techniques have been developed, such as XAI (Miller 2019). In this article, we will not discuss in detail the technical ways in which explainability is achieved (but see Adadi and Berrada 2018 for an overview).

Algorithmic explainability can take different forms and the type of explanation that is needed depends on the user of the algorithm. In this article, we focus on simplified explanations that can be understood by a non-expert (citizen) audience. Such an explanation should, for instance, provide the reasons for what was the crucial variable that contributed to an algorithmic outcome (Friedrich and Zanker 2011; Kizilcec 2016). Citizens want to receive clear reasons for a (negative) decision to assess the fairness of a decision (De Fine Licht and De Fine Licht 2020; Tyler 2006) or to be able to contest an algorithmic decision (Mittelstadt et al. 2019).

TRUSTWORTHINESS OF ALGORITHMIC DECISIONS

Algorithmic transparency is quoted as a crucial mechanism to increase trust, yet the evidence so far is mixed and there are limited studies applying the effects of algorithmic transparency on trustworthiness to street-level bureaucracy. This section develops hypotheses regarding the effects of explainability and accessibility by first discussing the existing evidence about the perceived trustworthiness of algorithms in general and, second, by specifically highlighting specific studies that have tested the effect of algorithmic transparency in contexts outside of public administration.

A well-known study on the trustworthiness of algorithms was carried out by Dietvorst et al. (2015). In a series of experiments, they found that people are subject to what they term "algorithm aversion": people tend to trust algorithmic predictions less than human predictions after seeing them err, even if they saw an algorithm outperform a human prediction. Interestingly, algorithm aversion has not been found to occur in other contexts. For instance, a recent study by Araujo et al. (2020) found that algorithmic systems are generally seen as equally fair and sometimes even fairer than human decision-makers. Moreover, fully automated decision-making, especially relatively simple decisions can also increase the perceived fairness of decision-making (Miller and Keiser 2020).

In more complex decision algorithms are much harder to scrutinize which makes the issue of procedural fairness more pressing (Kroll et al. 2016; Lepri et al. 2018). For instance, Lee (2018) carried out an experimental study in which participants were asked to rate the trustworthiness of algorithmic versus human assessments. For "technical" tasks, such as planning a work schedule, there was no difference in rating, however, for tasks that required "human judgment," such as hiring decisions, algorithms were perceived as less trustworthy. In a recent study, Nagtegaal (2021) carried out a similar study in a public management context and found that for complex decisions (hiring), managerial (human) judgment led to high ratings of perceived fairness, whereas for relatively 'simple' tasks (pension calculation) the opposite was true.

This evidence highlights the importance of decision complexity for the trustworthiness of algorithmic decisions. Street-level bureaucrats deal with a range of decisions, ranging from relatively simple, such as a visa application or child benefits, to more complex, such as assessing which welfare recipients are likely committing fraud (Maynard-Moody and Portillo 2010). By introducing AI into the latter, more complex, decision-making processes in street-level bureaucracies could therefore come at the risk of eroding perceived trustworthiness.

Thus, to sustain perceived trustworthiness in high discretionary decisions, algorithmic transparency is even more important. At the same time, we know very little about the effect of algorithmic transparency on trustworthiness in such situations. A positive effect of algorithmic transparency can be hypothesized based on procedural fairness theory. Tyler and colleagues have developed this concept and throughout the past decades, many studies have shown that people's perceived fairness of decision-making procedures affects their overall trust in authority and decision outcomes (Tyler 2006). Central to procedural justice theory is the relation between how authorities use their power and how subordinates assess their claims of power. When decision-making power is exercised in a way that is well-explained and procedurally fair, it is more likely that the decision is acceptable and the decision-making authority is trusted (Sunshine and Tyler 2003; Tyler and Huo 2002).

For a person to trust decision-making when he or she is subject to a negative decision outcome—e.g. s/he has been denied benefits—it is crucial that the procedure of a decision was fair. Algorithmic transparency taps into various elements of procedural fairness. Explaining an algorithmic decision will—ideally—show that a decision was taken in an unbiased and well-considered manner. Furthermore, the accessibility of an algorithm can be an important cue signaling openness. Therefore, we expect that transparent algorithms—algorithms that are accessible and explainable—improve their trustworthiness among the general public.

There is some empirical evidence showing that algorithmic transparency can indeed increase user trust in its decisions. Kizilcec (2016) found that providing an explanation about how an algorithm calculated student grades increased students' perceived trustworthiness. At the same time, this study also found that too much detailed information backfired and led to less perceived trustworthiness. In addition, a recent experiment by Schiff, Schiff and Pierson (2021) finds that citizens have less trust in automated decisions in which there is "transparency failure," i.e., the decisions are not understood by government officials themselves.

While a positive effect on perceived trustworthiness is expected, it is not a straightforward relation. Algorithm aversion may occur when people are shown that an algorithm makes mistakes, even when people see that an algorithm outperforms human judgment (Dietvorst et al. 2015). This suggests that being transparent about how an algorithm works could even decrease its trustworthiness. Indeed, in an experiment, Schmidt et al. (2020) found that highlighting elements of the algorithm's confidence in certain predictions decrease the trust of its users in the algorithm. Furthermore, Rader et al. (2018) conclude that although explaining the algorithm of Facebook's NewsFeed increased user awareness of the algorithm, yet also caused participants to perceive the NewsFeed as less fair and more biased. Overall, these studies indicate that algorithmic transparency could also lead to *less* perceived trustworthiness.

It should be noted that most of these studies that tested the effect of algorithmic transparency on trust employed samples with limited resemblance to the overall population (e.g. Kizilcec 2016) or are about decisions that have a little direct impact on citizen lives (e.g. Schmidt et al. 2020). To what extent algorithmic transparency in the context of street-level bureaucracies affects citizen trust remains an open question. Although the review above shows evidence is mixed it does tilt towards a positive relation (Schiff *et al*. 2021). Here, we specifically look into the effect two defining elements of the following algorithmic transparency:

> **H1.** An accessible algorithm will be perceived as more trustworthy, compared to an algorithm that is not accessible.
>
> **H2.** An explainable algorithm will be perceived as more trustworthy, compared to an algorithm that is not explained.

Hypotheses H1 and H2 regard the trustworthiness of a decision-making algorithm, however, in many instances algorithmic systems are not deciding autonomously and human decision-makers are still involved. Automation is shifting the ways in which human bureaucrats decide. For low discretion decisions, human judgment may be the limited and passive and only function as an "emergency break" when things go ultimately wrong. In medium to high discretion decisions, AI will assist street-level bureaucrats in the form of, for instance, risk profiles and risk

assessments (Bullock et al. 2020; Zouridis, Van Eck and Bovens 2020). In other words, advanced algorithms operate in interaction with human decision-makers. Therefore, we expect that algorithmic transparency will not only affect the perceived trustworthiness of the algorithmic decision but that this effect extends to the human decision-maker (H3 and H4).

> **H3.** A street-level bureaucrat that decides based on an accessible algorithm will be perceived as more trustworthy, compared to a street-level bureaucrat who is basing his/her decision on an inaccessible algorithm.

> **H4.** A street-level bureaucrat that decides based on an explainable algorithm will be perceived as more trustworthy, compared to a street-level bureaucrat who is basing his/her decision on an unexplained algorithm.

Furthermore, we hypothesize that the decision context moderates the effect of algorithmic transparency. One of the crucial debates in automated street-level automation concerns the relation between automation and discretion and whether increased automation curtails or enables discretionary space of street-level bureaucrats (Buffat 2015). Young, Bullock and Lecy (2019) theorize that the introduction of more advanced AI-powered decision systems has a gradual effect on discretion. They posit that in more complex tasks such as crime control or fraud detection, automation mostly supports the human judgment. The complexity of these tasks makes them hard to fully automate, therefore humans dominate such decision-making, and discretion is maintained to a large extent (De Boer and Raaphorst 2021). In simple tasks, automation largely takes over the decision-making process and human judgment only comes in to correct errors (Zouridis, Van Eck and Bovens 2020). Prior research, not specifically on street-level decision-making, showed that people perceive complex decisions as fairer when they are made by humans instead of machines (Lee 2018; Nagtegaal 2021).

In situations with relatively high levels of discretion, algorithmic transparency is expected to play a greater role in generating trustworthiness in the human decision-maker. As mentioned before, high discretion is often needed in more complex and impactful decisions such as detecting welfare fraud, it is exactly these decisions that will require more transparency to be trusted by citizens. We, therefore, hypothesize that algorithmic transparency will have a stronger effect on the perceived trustworthiness of the street-level decision-maker in a situation of high discretion.

> **H5.** The effect of algorithmic transparency (accessibility and explainability) on the perceived trustworthiness of a street-level bureaucrat will be more pronounced in a high discretion decision.

METHOD

Research Design

To test the aforementioned hypotheses, I designed two survey experiments with a factorial between-subjects 2×2 design. Each factor consisted of one of the two core dimensions of algorithmic transparency: accessibility (high or low) and explainability (high or low). One experiment contained a low discretion decision (visa application) and the other experiment contained a high discretion decision (decision to search house based on suspicion of welfare fraud). Each participant participated in both experiments, the order of the experiments was randomized, and after the first experiment participants completed a short distraction task to reduce spill-over effects from the first to the second experiment. Each experiment consisted of a short vignette with a realistic situation in which an automated decision was described. To ensure the realism of the vignette was discussed with three experts: one expert on street-level decision-making, one AI expert, and an expert on e-government. The vignette went through a couple of iterations because of these discussions.

The first part of the vignette was generic and the same for all participants and consisted of a short explanation of a decision. Participants were asked to imagine this situation in their own lives. Participants were aware of the hypothetical nature of the vignettes but were explicitly asked to immerse in the imaginary situation. I deliberately chose two realistic situations that are already used in (Dutch) bureaucracies to make them more imaginable. The experimental procedure is outlined in figure 1. Please note that both the order of experiments and the assignment to a treatment condition were random.

Materials and Selection Experimental Scenarios

The materials consisted of two vignettes related to two negative decision outcomes. A negative decision outcome was thought to be more appropriate for this study. Based on procedural fairness theory it is expected that transparency plays a role in citizen trust when they are confronted with a violation of their expectations of a decision (i.e. a negative outcome) (Tyler 2006). Theoretically, there is of course a possibility that even positive decision outcomes have a negative effect on trust. Perhaps people find themselves gaining benefits too easily and the system can be gamed by others in similar ways (Mittelstadt et al. 2016). Still, based on procedural justice theory it is expected that individuals who are confronted with a negative decision, will respond more strongly than those who "get what they want." Procedural justice theory predicts that individuals maintain their satisfaction with negative

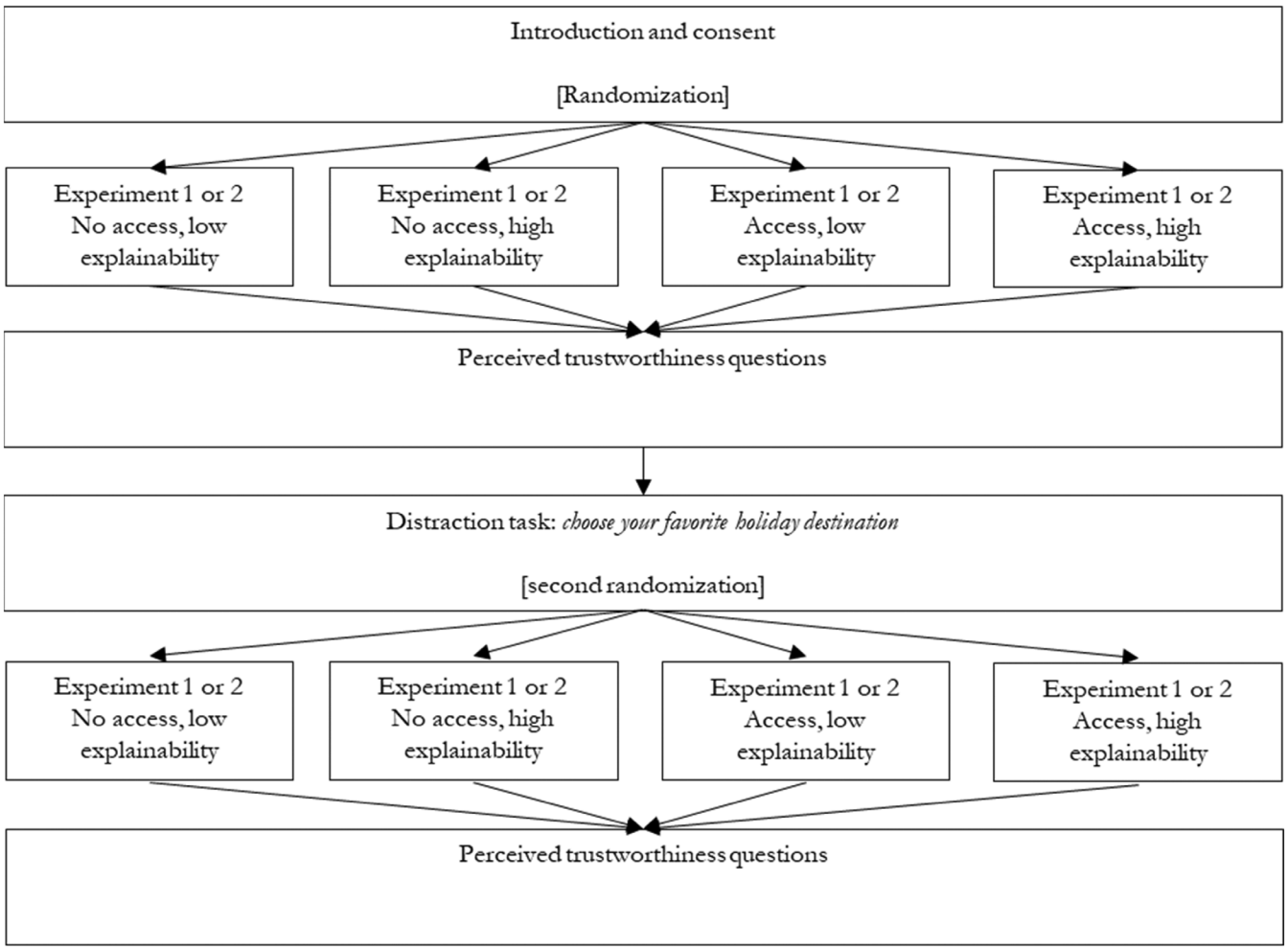

FIGURE 1 Experimental Set-up

decision outcomes if the preceding procedure leading up to it is perceived to be just and fair (Lind, Kanfer and Early 1990). In line with this theory, transparency is thought to foster perceptions of a fair procedure (e.g. Porumbescu and Grimmelikhuijsen 2018).

Two different scenarios of typical street-level decision-making were selected (see Appendix A for complete scenarios). The first scenario regarded a visa application procedure in which a visa was rejected because a person traveled to a "suspect country." The second scenario concerned a decision to conduct a house search for suspected welfare fraud.

First, these scenarios were chosen they both are easily understandable and recognizable to citizens, even if they have never had to deal with such government services in their real life. Second, the scenarios reflect different characteristics of street-level decision-making, most importantly decision discretion. In the visa application scenario decisional discretion is relatively low and street-level bureaucrats are more likely to semi-automatically follow algorithmic output. This relates to the relatively low complexity of the task that is being automated. In a case of low task complexity, the algorithm is more dominant and pushes out human discretion (Young, Bullock and Lecy 2019).

In welfare fraud, decision task complexity is higher as it concerns a much uncertain outcome (i.e. the chance that a welfare recipient committed fraud) with many more potential variables that can "predict" fraudulent behavior. In such complex tasks, the discretion of a human being remains relatively strong. In contrast, the second scenario entailed a street-level bureaucrat who used an algorithm to predict welfare fraud, which reflects the predictive algorithm that is typical of more recent (and contentious) AI applications in government. This also reflects that relatively new ways in which AI algorithms change street-level bureaucracy: not just low discretionary decisions are automated but also medium to high discretionary decisions are being affected by automation (e.g. Young, Bullock and Lecy 2019).

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

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

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

[Random assignment to one of the conditions below]

- The underlying code (algorithm) of the computer system is not accessible: the functioning of the computer system cannot be determined. [low accessibility]
- The underlying code (algorithm) of the computer system is accessible online to everyone: experts can determine how well the computer system functions and if the correct information is used. [high accessibility]

[Random assignment to one of the conditions below]

- The computer system only indicates that your visa is rejected but not what the reason is behind the rejection. [low explainability]
- The computer system indicates that your visa is rejected because in the past 5 years you have traveled at least once to a "suspect" country. [high explainability]

Data Collection

Data was collected through the high-quality Longitudinal Internet Studies for the Social Sciences (LISS) Panel. The LISS Panel is a representative sample of Dutch individuals who participate in monthly Internet surveys. The panel is based on a true probability sample of households drawn from the population register. Households that could not otherwise participate are provided with a computer and Internet connection. A longitudinal survey is fielded in the panel every year, covering a large variety of domains including work, education, income, housing, time use, political views, values, and personality (Scherpenzeel and Das 2010).

The data were collected in weeks 17 and 18 in 2019. In total 1,125 panel members were approached and 897 members completed the survey (response rate 79.7 percent). The sample was similar to the Dutch population with regard to sex, however, we observed that the average age and the average level of education (percentage who obtained bachelor degree) was somewhat higher than in the population (details in tables B3, Appendix B). Balance tests revealed no significant differences between the four treatment groups with regard to these and other variables (see tables B1, B2, Appendix B).

Measures

Perceived Trustworthiness

The core dependent variables were the perceived trustworthiness of the algorithm and the perceived trustworthiness of the human decision-maker. Trustworthiness judgments in humans or technology have different characteristics, yet they both are premised on the same underlying contextual conditions: risk, uncertainty and lack of control (McKnight et al. 2011). While McKnight et al. argue that people mainly evaluate the trustworthiness of simple technologies such as Excel in terms of effectiveness and functionality, other studies suggest that more complex technological systems, such as trust in medical technology, is assessed also in terms of their integrity and honesty (Montague et al. 2009). The type of technology in our study is complex and is more than a simple application that simply needs to "work" or "easy to use." Therefore, I include perceived honesty as a component of the perceived trustworthiness of algorithms, which is in line with existing conceptualizations of trustworthiness (Grimmelikhuijsen and Knies 2017; Mayer et al. 1995). It should be noted that the items do not literally mention the word "algorithm," but "computer system" to make the items easier to understand. The following items were used:

I trust that the computer system that was used to [recommend the house search / deny the visa application]… (1—Totally disagree; 7—Totally agree). Cronbach's alpha: 0.92 (exp 1), 0.92 (exp 2)

1. …used the correct information.
2. …did not take an incorrect decision here.
3. …assessed my situation honestly.
4. …used all relevant information.

To measure the perceived trustworthiness of the human decision-maker I used the same items. This was done to increase the comparability of both dependent variables. However, as mentioned above, it could be argued that this is a different type of trustworthiness (McKnight et al. 2011) and that applying the same measures is inappropriate. Still, there is a great deal of overlap in terms of contextual conditions and technology by which humans are significantly impacted are assessed on similar dimensions as persons, such as competence and honesty (Montague et al. 2009). The following measures were used in the experiments.

I trust that the employee of [the visa agency / welfare agency] that decided to [deny the visa / carry out the house search]… (1—Totally disagree; 7—Totally agree). Cronbach's alpha: 0.95 (exp 1), 0.94 (exp 2)

1. …used the correct information.
2. …did not take an incorrect decision here.
3. …assessed my situation honestly.
4. …used all relevant information.

Manipulation Check

Each experiment included two items to assess whether the manipulations (i.e. explainability and accessibility) worked the way it was intended. The wording and results of this check are in tables C1–C4, Appendix C. As expected, accessibility was perceived higher in the "high accessibility" condition and the "high explainability" condition also yielded the expected group differences.

RESULTS

The means and standard deviations of each experiment are shown in table 2.

Table 2 provides some interesting insights. First, in general, the human decision-maker (bureaucrat) tends to be trusted more than the algorithm. Although algorithms were initially hailed as improving the efficiency and fairness of government decision-making, the recent debate has taken a more critical turn. Perhaps this is reflected in lower trust levels in algorithmic systems, especially when they are involved in consequential and complex decisions. Second, the overall level of perceived trustworthiness of the algorithms is lower in Study 1 (visa application) than in Study 2 (welfare fraud), yet there is little difference between the trustworthiness of the bureaucrat in each study. It may indicate that people have already become more accustomed to algorithmic decision-making in government services such as visa applications, compared to welfare. Thirdly, the

TABLE 2 Means of Perceived Trustworthiness and Standard Deviations for Study 1 and Study 2

	Study 1, Visa Application		Study 2, Welfare Fraud	
Trustworthiness of…	**…Algorithm**	**…Bureaucrat**	**…Algorithm**	**…Bureaucrat**
No access, low explainable	3.23 (1.42)	3.58 (1.43)	2.90 (1.50)	3.44 (1.58)
Access, low explainable	3.46 (1.54)	3.84 (1.59)	3.33 (1.54)	3.73 (1.59)
No access, high explainable	3.72 (1.55)	3.83 (1.56)	3.50 (1.58)	3.90 (1.59)
Access, high explainable	3.69 (1.47)	3.98 (1.58)	3.61 (1.54)	3.89 (1.53)
Total	3.51 (1.51)	3.79 (1.54)	3.34 (1.56)	3.74 (1.56)

TABLE 3 Results Study 1 "Visa Application"

	Trust in Algorithm			Trust in Bureaucrat		
Predictors	**Std. Beta**	**Standardized CI**	***p***	**Std. Beta**	**Standardized CI**	***p***
(Intercept)	−0.19	−0.31 to −0.07	**<.001**	−0.14	−0.26 to −0.02	**<.001**
Accessible	0.16	−0.02 to 0.34	.090	0.17	−0.02 to 0.35	.073
Explainable	0.33	0.15 to 0.50	**<.001**	0.16	−0.02 to 0.34	.079
Access* Explainable	−0.17	−0.43 to 0.09	.200	−0.07	−0.33 to 0.20	.625
Observations		905			905	
R^2/R^2 adjusted		0.018/0.015			0.009/0.006	

Notes: The bold values represent values that are significant at $p < .05$.

TABLE 4 Result Study 2 "Welfare Fraud"

	Trust in Algorithm			Trust in Bureaucrat		
Predictors	**Std. Beta**	**Standardized CI**	***p***	**Std. Beta**	**Standardized CI**	***p***
(Intercept)	−0.28	−0.41 to −0.15	**<.001**	−0.19	−0.32 to −0.06	**<.001**
Accessible	0.27	0.09 to 0.46	**.003**	0.18	−0.00 to 0.37	.051
Explainable	0.38	0.20 to 0.56	**<.001**	0.29	0.11 to 0.47	**.002**
Access × Explainable	−0.20	−0.46 to 0.05	.123	−0.19	−0.45 to 0.07	.146
Observations		905			903	
R^2/R^2 adjusted		0.030/0.027			0.014/0.011	

Notes: The bold values represent values that are significant at $p < .05$.

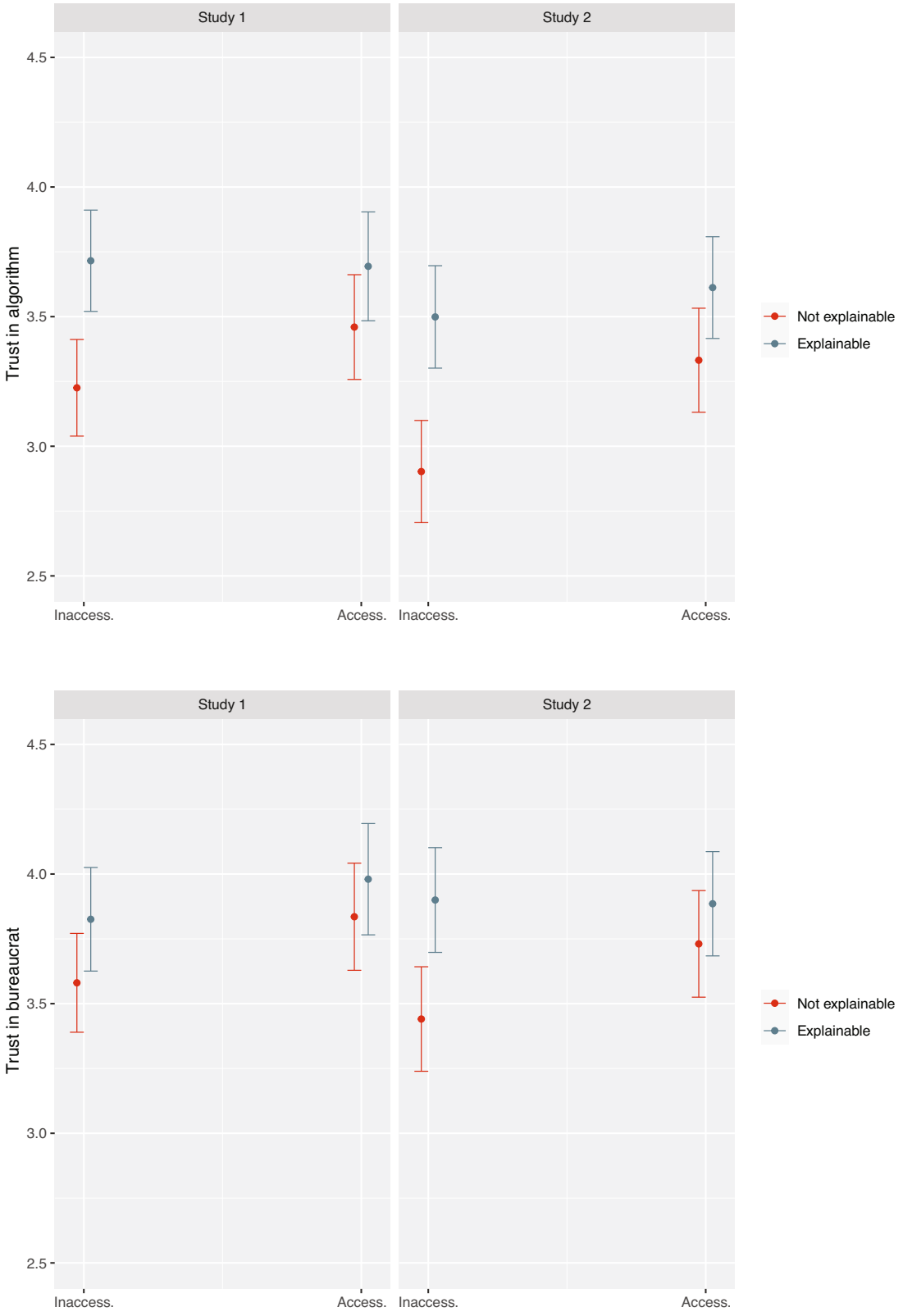

FIGURE 2 Estimated Marginal Means and Confidence Intervals (95 percent)

trustworthiness means tend to be higher in the treatment groups that include one or two dimensions of algorithmic transparency. We will further analyze this in the next section.

Study 1: Visa Application Experiment (Low Discretion)

The first experiment concerned a low discretion decision in which I assessed the effect of algorithmic transparency on trust in the human decision-maker and the algorithm making the decision (see table 3 for results).

Overall, we find no significant effect of algorithmic transparency on perceived trustworthiness in the human decision-maker. Table 4 indicates that the standardized betas are in the expected direction but small and not significant at the conventional alpha-level ($p < .05$). We do find a significant effect of explainability on trustworthiness of the algorithm itself ($B = 0.16$, $p < .001$) and again we see expected yet non-significant patters for accessibility ($B = 0.08$, $p = .09$). This indicates that algorithmic transparency in the visa application case hardly affects trust in a human decision-maker that uses an algorithmic system and that, from the two components of transparency, only explainability has a clear positive effect on trust in the algorithm itself. Figure 2 shows the mean trust values in both the human and algorithm for both studies.

Study 2: Welfare Fraud Experiment (High Discretion)

The second experiment concerned a street-level decision in the welfare domain. In the scenario, a human bureaucrat decision-maker decides to pursue a case of welfare fraud by doing a house search. The search was prompted by an algorithmically powered recommendation. As described in the first paragraph of the Introduction, this is a realistic case in the Netherlands where, after the experiment was finished, a judge decided that a welfare fraud detection system providing such recommendations is no longer allowed to be used because of its opaqueness and supposed bias. Table 4 shows how transparency affects trust in this case.

Compared to the first experiment we find more pronounced effects of algorithmic transparency: also trustworthiness of the human decision-maker is affected by algorithmic transparency. Perhaps this time, the human decision-maker (a social worker) is perceived to have more responsibility to follow-up on the system's recommendation now that a house search requires more discretion and is also a much stronger infringement than compared to the visa experiment. We find a strong significant effect of explainability ($B = .15$, $p = .002$) and a borderline significant effect of accessibility ($B = 0.09$, $p = .051$). Both elements of transparency affect trust in the algorithm itself, this time also accessibility increases trust ($B = 0.14$, $p = .003$) next to explainability ($B = 0.19$, $p < .001$).

TABLE 5 Summary of Hypotheses

Hypothesis	Study 1: visa	Study 2: welfare
H1: Accessibility → trust in algorithm	Rejected	Supported
H2: Explainability → trust in algorithm	Supported	Supported
H3: Accessibility → trust in bureaucrat	Rejected	Rejected
H4: Explainability → trust in bureaucrat	Rejected	Supported
H5: Transparency has stronger effect in high discretion scenario	Inconclusive evidence	

Table 5 shows that we find convincing support across both experiments for H2: in both studies providing an explanation for an algorithmic decision that affected trust in the algorithm. Furthermore, we find mixed evidence that algorithmic accessibility affects trust in a bureaucrat (H3). For H1 and H4 the evidence is also mixed, and effects seem to depend on the context of the decision. For instance, algorithmic accessibility only had an effect in the case of welfare fraud; it did not increase trust in the algorithm in the visa application experiment (H1).[3]

Finally, the effect of algorithmic explainability on the trustworthiness of a human, bureaucrat, decision-maker was only found in the high discretion welfare fraud experiment ($p = .002$). This suggests that explainability has a stronger effect dependent on the decision context, which is in line with H5 (algorithmic transparency has a stronger effect in a high discretion scenario). To assess H5 further, we carried out an additional three-way interaction analysis in which "Study" is entered as an additional variable, next to accessibility and explainability. If H5 is to be accepted, we would expect that the treatment variables interact with the new "Study" variable. However, we found no significant interaction effects of the Study with the treatment variables on trust in the algorithm nor trust in the bureaucrat (see Appendix E for the full analysis). This means that while we find different effects in each study, we cannot be sure that these differences are not due to chance. Statisticians have warned that interaction effects are notoriously hard to detect with sufficient statistical power (e.g. Simonsohn 2014). With both studies combined the sample size is 1,803 and we would be able to detect effect sizes of f = 0.076 (assuming $p < .05$ and power = 0.90). Such effect sizes are relatively large for interactions, especially when you consider that we predicted an ordinal interaction, which is a type of interaction in which the direction of the effect is not completely flipped, but only gradually smaller or larger, which are generally the kind of interactions social scientists predict (Lakens and Caldwell 2021). I will discuss the potential implications of the findings on H5 in more detail in the Conclusion and Discussion section.

CONCLUSION AND DISCUSSION

Recent academic and policy debates have emphasized algorithmic transparency as a means to secure the trustworthiness of bureaucratic decision-making, but empirical tests have been lacking. In response, this article tested whether two core components of algorithmic transparency, accessibility and explainability, affect citizen trust in a human decision-maker and trust in an algorithm. The two experiments in this paper yield three main conclusions. Overall, this suggests that: (1) explainability yields more importance to citizen trust than the accessibility of the algorithm; (2) the effect of algorithmic transparency not only pertains to trust in the algorithm but also—partly—to trust in the human decision-maker and (3) the effects of algorithmic transparency are not robust across study context: in the high discretion scenario we found that both accessibility and explainability affected the perceived trustworthiness of algorithmic decision-making.

First, the findings present support for the positive effect of explainable algorithms on citizen trust in those algorithms. This is in line with the assumption that algorithmic explainability is key to public trust. For instance, the General Data Protection Regulation (GDPR), enacted by the European Union to protect individual privacy, is also meant to protect individuals against opaque algorithms. Article 22 highlights explainable artificial intelligence, though it is not yet clear how this should look in practice. This study underscores the practical importance of algorithmic explainability in the context of street-level decision-making. This conclusion also underlines the point made by De Fine Licht and De Fine Licht (2020). In a conceptual piece on the legitimacy of algorithmic public decisions, they argue that providing a decision justification is needed to sustain government legitimacy.

Second, the effect of algorithmic transparency not only pertains to trust in the algorithm but also—to an extent—to trust in the human decision-maker that uses the algorithm. The effect depends on the decision context as we found no effect in the visa application experiment, but we did find an effect in the welfare fraud detection experiment. Here the role of human discretion in the decision becomes apparent. Discretion is a crucial concept in street-level decision-making and studies suggest that algorithms have started interfering in the discretion of more complex bureaucratic decisions by prompting decision recommendations to bureaucrats, especially in complex tasks risk assessment (Bullock 2019; Young, Bullock and Lecy 2019; Zouridis, Van Eck and Bovens 2020).

Third and finally, we have seen that the effect of algorithmic transparency is not robust across study context. Specifically, in the experiment on welfare fraud detection, both accessibility and explainability had significant effects on trust in the algorithm and bureaucrat, while in the visa experiment only a significant effect of explainability on trust in the algorithm was found. An interaction analysis, however, did not reveal a significant interaction between study context and either of the transparency treatments. This means we cannot state with certainty that the study context actually alters the effect transparency. Still, we may expect that in decisions with higher discretion (and/or higher stakes) both accessibility and explainability are important, whereas "only" explainability is relevant to citizens in other decision contexts.

I consider two promising directions for future research for exploring these decision contexts more systematically. First, researchers might consider the *personal relevance of a decision outcome*. In the case of welfare fraud detection, the consequences of the decision are much more pervasive than in the visa application case. This aligns with findings from the procedural justice literature, where scholars have found that perceived fairness is especially relevant for trust in authority when the decision outcome matters more to the recipient (Grootelaar and van den Bos 2018). In the event of such a negative decision outcome, people will want to seek information that helps them to interpret the situation (Fiske and Taylor 1991). Second, to disentangle the context-specificity of algorithmic transparency, future research could *systematically test decisions with varying levels of discretion and complexity*. For instance, Schiff, Schiff and Pierson (2021) find that a transparency failure in automated court decisions has a negative impact on trust, while they find no such effect in child welfare decisions. Future research may design a series of vignettes where potentially relevant context characteristics are systematically varied, such as the degree of personal relevance, the degree of complexity of a decision and the degree of discretion.

This study is subject to some limitations. First, this article presents two survey experiments with realistic, yet hypothetical scenarios. Survey experiments have been criticized for their lack of external validity. Although survey experimental effects sometimes have been found to replicate in real-world effects (e.g. Barabas and Jerit 2010), survey experimental treatments are often presented in a "clean" survey environment and thus the effects found in a survey experiment may not be as strong as they occur in the real world (e.g. Gaines, Kuklinski and Quirk 2007). While I underscore this criticism, the current manipulation (algorithmic transparency) is hard to test in the real world in an ethical manner. For this reason, it is important to first establish the potential effect of the high internal validity of a survey experiment. Future research could employ different (experimental) methods to further probe the external validity of our findings. To test the real-life effect of algorithmic transparency public administration scholars should work closely with computer scientists to carry out field experimental tests in which real or mock algorithms are tested, changed, and retested.

A second limitation is the type of explanation that was included in the experiment. There is still discussion among AI scholars on what kind of explanations an algorithm could or should produce, depending on the skills and competence of the user (e.g. Ahmad et al. 2018;

Došilović et al. 2018). From a perspective of public administration, it is also interesting to question what kind of justifications should be addressed to which public. Here we only focused on explanations that are interpretable for the "end user," i.e. a non-expert citizen requesting a government decision. Other types of arguments and explanations need to be systematically developed and tested and future research could also look into the type of explanations that street-level bureaucrats need to be able to trust and understand an algorithm.

A third limitation is that transparency as operationalized in the two experiments—as a first step in empirically testing algorithmic transparency—are rather crude as they, for instance, provide either accessibility or no accessibility at all. There can be multiple ways of making data and code accessible, and there can be multiple ways to justify a decision and different degrees of transparency in the decision (Miller 2019; Mittelstadt et al. 2019). Even more so, there can be different ways to present (complex) government information. Subsequent research may test more subtle and gradual transparency treatments.

A final limitation is that in this experiment transparency did not reveal any algorithmic bias. Instead, I chose to test a "positive" scenario for algorithmic transparency in which transparency reveals a clear and apparently non-biased explanation for a negative decision outcome. At the same time, many journalists and some scholars report critically about algorithmic bias and discrimination (e.g. Eubanks 2018). Future research could compare biased and unbiased explanations in an attempt to better capture the current criticism on algorithmic decision-making in government.

The conclusions of this study have important implications for theory on digitization in public organizations and street-level bureaucracy in particular. Several scholars have pointed out that automation is affecting managerial decision-making (Meijer, Lorenz and Wessels 2021), organizational accountability and power constellations (Zouridis, Van Eck and Bovens 2020) and discretion of street-level bureaucrats (Young, Bullock and Lecy 2019). With street-level bureaucracies transforming so profoundly we must stay keen on safeguarding public values such as transparency. Scholars have repeatedly argued for transparent and/or accountable algorithms (e.g. Busuioc 2020), this is one of the first studies carried out in a public administration context to provide empirical evidence for this.

In particular, by disentangling the effects accessibility and explainability we can now also state that transparency as accessibility is not a sufficient condition for public trust. This implies that to maintain trust in algorithmic decision-making, governments must invest in the explainability of algorithms in addition to making source code and data accessible to independent experts. The negative consequences of non-transparent algorithms in public administration for citizen trust—and fairness for that matter—are simply too dire to ignore.

ACKNOWLEDGMENTS

I thank the following people for their comments on a previous version of paper and/or feedback on the experimental design: Mark Bovens, Albert Meijer, Floris Bex, Lars Tummers, Robin Bouwman and Marij Swinkels. I also thank three anonymous reviewers for their constructive feedback on the manuscript. Data collection was funded by a research grant from the Netherlands Organization for Scientific Research (NWO) under grant number: VENI-451-15-024.

ENDNOTES

1. Whether algorithmic transparency indeed leads to gaming is subject to debate. When the goals of a system and a user are aligned, gaming might actually help to improve a system (Rudin 2019).
2. There is some debate in the computer science literature on explainability and whether and how it can be discerned from interpretability. There is no clear consensus in the literature on what constitutes an "explainable" versus and "interpretable" model. Sometimes the terms are used interchangeably (e.g. Mittelstadt et al. 2019). A useful distinction is made in a much-cited review piece on explainability in AI by Miller. Miller (2019) argues that explainability and interpretability are two partially overlapping concepts. He defines, based on work by Biran and Cotton (2017) interpretability of a model as "the degree to which an observer can understand the cause of a decision. Explanation is thus one mode in which an observer may obtain understanding" (Miller 2019, 8). In this article, I adopt this distinction between explainability and interpretability.
3. Appendix D1 and D2 provides an analysis of order effects of Study 1 and Study 2. Patterns here are similar as the analysis with all participants included.

REFERENCES

Adadi, Amina, and Mohammed Berrada. 2018. Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6: 52138–60.

Ahmad, Muhammad Aurangzeb, Eckert, Carly & Teredesai, Ankur. 2018, August. Interpretable Machine Learning in Healthcare. In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (pp. 559–560).

Annany, Mike, and Kate Crawford. 2018. Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability. *New Media & Society* 20(3): 973–89.

Araujo, Theo, Natali Helberger, Sanne Kruikemeier, and Claes De Vreese. 2020. In AI We Trust? Perceptions About Automated Decision-Making by Artificial Intelligence. *AI & Society* 35: 1–13.

Barabas, Jason, and Jennifer Jerit. 2010. Are Survey Experiments Externally Valid? *American Political Science Review* 104(2): 226–42.

Biran, Or and Courtenay Cotton. 2017. Explanation and Justification in Machine Learning: A Survey. IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI), pp. 8–13

de Boer, Noortje, and Nadine Raaphorst. 2021. Automation and Discretion: Explaining the Effect of Automation on how Street-Level Bureaucrats Enforce. *Public Management Review* 1-21: 1–21. https://doi.org/10.1080/14719037.2021.1937684.

Bovens, Mark, and Stavros Zouridis. 2002. From Street-Level to System-Level Bureaucracies: How Information and Communication Technology Is Transforming Administrative Discretion and Constitutional Control. *Public Administration Review* 62(2): 174–84.

Buffat, Aurélien 2015. Street-Level Bureaucracy and e-Government. *Public Management Review* 17(1): 149–61.

Bullock, Justin 2019. Artificial Intelligence, Discretion, and Bureaucracy. *The American Review of Public Administration* 49(7): 751–61.

Bullock, Justin, Matthew Young, and Yi-Fan Fang. 2020. Artificial Intelligence, Bureaucratic Form, and Discretion in Public Service. *Information Polity* 25(4): 491–506.

Burrell, Jenna. 2016. How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms. *Big Data & Society* 3(1): 1–12.

Busuioc, Madalina. 2020. Accountable Artificial Intelligence: Holding Algorithms to Account. *Public Administration Review* 81: 825–36. https://doi.org/10.1111/puar.13293.

Coulthart, Stephen, and Ryan Riccucci. 2022. Putting Big Data to Work in Government: The Case of the United States Border Patrol. *Public Administration Review* 82: 280–89. https://doi.org/10.1111/puar.13431.

Cucciniello, Maria, Gregory Porumbescu, and Stephan Grimmelikhuijsen. 2017. 25Years of Transparency Research: Evidence and Future Directions. *Public Administration Review* 77(1): 32–44.

Dietvorst, Berkeley, Joseph Simmons, and Cate Massey. 2015. Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing them Err. *Journal of Experimental Psychology: General* 144(1): 114–26.

Došilović, Filip, Brčić Mario & Hlupić Nikica (2018) Explainable Artificial Intelligence: A Survey. In: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), IEEE, 0210-0215.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

de Fine Licht, Jenny. 2014. Policy Area as a Potential Moderator of Transparency Effects: An Experiment. *Public Administration Review* 74(3): 361–71.

de Fine Licht, Karl, and Jenny de Fine Licht. 2020. Artificial Intelligence, Transparency, and Public Decision-Making. *AI & Society* 35: 1–10.

Fiske, Susan, and Shelley Taylor. 1991. *Social Cognition*, 2nd ed. New York: McGraw-Hill.

Friedrich, Gerhard, and Markus Zanker. 2011. A Taxonomy for Generating Explanations in Recommender Systems. *AI Magazine* 32(3): 90–8.

Gaines, Brian, James Kuklinski, and Paul Quirk. 2007. The Logic of the Survey Experiment Reexamined. *Political Analysis* 15: 1–20.

Giest, Sarah, and Stephan Grimmelikhuijsen. 2020. Introduction to Special Issue Algorithmic Transparency in Government: Towards a Multi-Level Perspective. *Information Polity* 25(4): 409–17.

Grimmelikhuijsen, Stephan, and Eva Knies. 2017. Validating a Scale for Citizen Trust in Government Organizations. *International Review of Administrative Sciences* 83(3): 583–601.

Grimmelikhuijsen, Stephan, and Albert Meijer. 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): 137–57.

Grootelaar, Hilke, and Kees van den Bos. 2018. How Litigants in Dutch Courtrooms Come to Trust Judges: The Role of Perceived Procedural Justice, Outcome Favorability, and Other Sociolegal Moderators. *Law & Society Review* 52(1): 234–68.

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franca Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM computing surveys (CSUR)* 51(5): 93.

Hill, Robin 2015. What an Algorithm Is. *Philosophy & Technology* 29(1): 35–59.

Janssen, Marijn, and Jeroen van den Hoven. 2015. Big and Open Linked Data (BOLD) in Government: A Challenge to Transparency and Privacy? *Government Information Quarterly* 32(4): 363–8.

Kitchin, Rob. 2017. Thinking Critically about and Researching Algorithms. *Information, Communication & Society* 20(1): 14–29.

Kizilcec, René. 2016. How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2390–5. ACM.

Kroll, Joshua, Joanna Huey, Solon Barocas, Edward Felten, Joel Reidenberg, David Robinson, and Yu, Harlan. 2016. Accountable Algorithms. *University of Pennsylvania Law Review* 165: 633.

Lakens, Daniël, and Aaron Caldwell. 2021. Simulation-Based Power Analysis for Factorial Analysis of Variance Designs. *Advances in Methods and Practices in Psychological Science* 4(1): 2515245920951503.

Lee, Min Kyung. 2018. Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management. *Big Data & Society* 5(1): 2053951718756684.

Lepri, Bruno, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. Fair, Transparent, and Accountable Algorithmic Decision-Making Processes. *Philosophy & Technology* 31(4): 611–27.

Lind, Allen, Ruth Kanfer, and Paul Earley. 1990. Voice, Control, and Procedural Justice: Instrumental and Noninstrumental Concerns in Fairness Judgments. *Journal of Personality and Social Psychology* 59(5): 952–9.

Mayer, Roger, James Davis, and David Schoorman. 1995. An Integrative Model of Organizational Trust. *Academy of Management Review* 20(3): 709–34.

Maynard-Moody, Steven, and Shannon Portillo. 2010. Street-Level Bureaucracy Theory. In *Oxford Handbook of American Bureaucracy*, edited by R. Durant, 252–77. Oxford: Oxford University Press.

McKnight, Harrison, Michelle Carter, Jason Thatcher, and Paul Clay. 2011. Trust in a Specific Technology: An Investigation of its Components and Measures. *ACM Transactions on Management Information Systems* 2(2): 12.

Meijer, Albert, and Stephan Grimmelikhuijsen. 2020. Responsible and Accountable Algorithmization. How to Generate Citizen Trust in Governmental Usage of Algorithms. In *The Algorithmic Society: Technology, Power, and Knowledge*, edited by M. Schuilenburg and R. Peeters. New York: Routledge.

Meijer, Albert, Lukas Lorenz, and Martijn Wessels. 2021. Algorithmization of Bureaucratic Organizations: Using a Practice Lens to Study how Context Shapes Predictive Policing Systems. *Public Administration Review* 81: 837–46. https://doi.org/10.1111/puar.13391.

Mergel, Ines, Karl Rethemeyer, and Kimberly Isett. 2016. Big Data in Public Affairs. *Public Administration Review* 76(6): 928–37.

Miller, Tim 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267: 1–38.

Miller, Susan, and Lael Keiser. 2020. Representative Bureaucracy and Attitudes toward Automated Decision Making. *Journal of Public Administration Research and Theory* 31: 150–65. https://doi.org/10.1093/jopart/muaa019.

Mittelstadt, Brent, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The Ethics of Algorithms: Mapping the Debate. *Big Data & Society* 3(2): 205395171667967.

Mittelstadt, Brent, Chris Russell, and Sandra Wachter. 2019, January. Explaining Explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 279–88.

Montague, Enid, Brian Kleiner, and Woodrow Winchester III. 2009. Empirically Understanding Trust in Medical Technology. *International Journal of Industrial Ergonomics* 39(4): 628–34.

Nagtegaal, Rosanna. 2021. The Impact of Using Algorithms for Managerial Decisions on Public employees' Procedural Justice. *Government Information Quarterly* 38(1): 101536.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Penguin Publishers.

Peeters, R., and Widlak, A. 2018. The Digital Cage: Administrative Exclusion through Information Architecture–The Case of the Dutch Civil Registry's Master Data Management System. *Government Information Quarterly* 35(2): 175–83.

Peeters, Rik. 2020. The Agency of Algorithms: Understanding Human-Algorithm Interaction in Administrative Decision-Making. *Information Polity* 25(4): 507–22.

Pencheva, Irina, Marc Esteve, and Slava Mikhaylov. 2020. Big Data and AI—A Transformational Shift for Government: So, What Next for Research? *Public Policy and Administration* 35(1): 24–44.

Porumbescu, Greg, and Stephan Grimmelikhuijsen. 2018. Linking Decision-Making Procedures to Decision Acceptance and Citizen Voice: Evidence from Two Studies. *The American Review of Public Administration* 48(8): 902–14.

Rader, E., Cotter, K., and Cho, J. 2018. Explanations as mechanisms for supporting algorithmic transparency. *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

Rudin, Cynthia. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* 1(5): 206–15.

Scherpenzeel, Annette, and Marcel Das. 2010. True Longitudinal and Probability-Based Internet Panels: Evidence from The Netherlands. In *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies*, edited by Marcel Das, Peter Ester, and Lars Kaczmirek, 77–104.

Schiff, Daniel, Kaylyn Schiff, and Patrick Pierson. 2021. Assessing Public Value Failure in Government Adoption of Artificial Intelligence. *Public Administration*. https://doi.org/10.1111/padm.12742.

Schmidt, Phillip, Felix Biessmann, and Timm Teubner. 2020. Transparency and Trust in Artificial Intelligence Systems. *Journal of Decision Systems* 29(4): 260–78.

Simonsohn, Uri. 2014. No-Way Interactions. *Datacolada* (blog). https://doi.org/10.15200/winn.142559.90552.

Sunshine, Jason, and Tom Tyler. 2003. The Role of Procedural Justice and Legitimacy in Shaping Public Support for Policing. *Law & society Review* 37(3): 513–48.

Tutt, Andrew. 2017. An FDA for Algorithms. *Administrative Law Review*: 83–123.

Tyler, Tom. 2006. *Why People Obey the Law*. New Haven: Yale University Press.

Tyler, Tom, and Yuen Huo. 2002. *Trust in the Law: Encouraging Public Cooperation with the Police and Courts*. Russell Sage Foundation.

Vogl, Thomas, Cathrine Seidelin, Bharath Ganesh, and Jonathan Bright. 2020. Smart Technology and the Emergence of Algorithmic Bureaucracy: Artificial Intelligence in UKLocal Authorities. *Public Administration Review* 80(6): 946–61.

Wennekers, A., Boelhouwer, J., van Campen, C., and Bijl, R. 2018. *De sociale staat van Nederland 2018*, Netherlands Institute for Social Research (SCP).

Young, Matthew, Justin Bullock, and Jesse Lecy. 2019. Artificial Discretion as a Tool of Governance: A Framework for Understanding the Impact of Artificial Intelligence on Public Administration. *Perspectives on Public Management and Governance* 2(4): 301–13.

Zouridis, Stavros, Marlies van Eck, and Mark Bovens. 2020. Automated Discretion. In *Discretion and the Quest for Controlled Freedom*, edited by Tony Evans and Peter Hupe, 313–29. Cham: Palgrave Macmillan.

AUTHOR BIOGRAPHY

Stephan Grimmelikhuijsen is an associate professor at Utrecht University, The Netherlands. His research centers theme concerning technology in government, citizen-state interactions and behavioral public administration.
Email: s.g.grimmelikhuijsen@uu.nl

How to cite this article: Grimmelikhuijsen, Stephan. 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

APPENDIX A: Full Experimental Scenarios (Translated Versions From Dutch)

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

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

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

1A. The underlying code (algorithm) of the computer system is not accessible: the functioning of the computer system cannot be determined. [low accessibility].

1B. The underlying code (algorithm) of the computer system is accessible online to everyone: experts can determine how well the computer system functions and if the correct information is used. [high accessibility].

2A. The computer system only indicates that your visa is rejected but not what the reason is behind the rejection. [low explainability].

2B. The computer system indicates that your visa is rejected because in the past 5 years you have traveled at least once to a "suspect" country. [high explainability].

The second experiment contained a vignette that was similar in structure, yet it concerned a situation in which an algorithm was used to detect welfare fraud.

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

Imagine: You are a single person and dependent on government assistance. One day an investigating officers comes to your home to search the house: you are being suspected of welfare fraud. The officer does not find evidence of fraud in your house. You remain a suspect until the investigation is finished.

You experience the accusation and house search as a serious breach of your privacy and you want to know why you are a suspect. The investigating officer indicates that the suspicion has been based on a computer system:

1A. The underlying code (algorithm) of the computer system is not accessible: the functioning of the computer system cannot be determined. [low accessibility].

1B. The underlying code (algorithm) of the computer system is accessible online to everyone: experts can determine how well the computer system functions and if the correct information is used. [high accessibility].

2A. The computer system only indicates that you are a suspect, but not what the reason is behind the rejection. [low explainability].

2B. The computer system indicates that you are a suspect, because information from your energy supplier shows that you have been using much more gas and electricity. This indicates that you may be living together. Since you have not filed this you may have unfairly received assistance. [high explainability].

APPENDIX B: Balance Tests

TABLE B1 Study 1 (Visa Application)

	Sex (% M)	Age	Education[a]	Socio-economic Status (SES)[b]
Low access, low interpret	51.7	51.8	4.0	2.4
High access, low interpret	50.0	49.7	3.9	2.4
Low access, high interpret	48.5	50.8	4.0	2.5
High access, high interpret	48.5	52.6	3.8	2.4
Total	49.8	51.2	3.9	2.4

Notes: Sex ($\chi^2(3) = .68$, $p = .878$); Age ($F[3, 914] = 1.37$, $p = .251$); Education ($F[3,913] = .92$, $p = .433$); SES ($F[3,912] = .18$, $p = .913$).
[a]Range 1–6: 1 = finished primary education, 6 = university (master) degree. Median education was 4 ("secondary vocational education").
[b]Range 1–5: 1 = high, 5 = low.

TABLE B2 Study 2 (Welfare Fraud Detection)

	Sex (% M)	Age	Education	Socio-economic Status (SES)
Low access, low interpret	49.4	51.6	3.9	2.4
High access, low interpret	48.4	50.3	4.1	2.4
Low access, high interpret	52.6	51.9	3.9	2.4
High access, high interpret	48.7	51.0	3.9	2.4
Total	49.8	51.2	3.9	2.4

Notes: Sex ($\chi^2(3) = 1.02$, $p = .796$); Age ($F[3, 914] = 0.49$, $p = .688$); Education ($F[3,913] = .1.21$, $p = .3.06$); SES ($F[3,912] = .36$, $p = .783$).

TABLE B3 Sample Compared to Dutch Population

	Sample	Dutch Population (From: Wennekers et al., 2018)
Age	51.2	41.8
% female	50.2	50.4
% primary vocational education (bachelor degree) or higher	43.2	30.0

APPENDIX C: Manipulation Checks

C.1 | Manipulation Checks Study 1 (Visa Application)

TABLE C1 Accessibility

	I feel that everyone who is willing and able could get access to the underlying code (the algorithm) of the computer system to check how it works (1 = strong disagree to 7 = strong agree)		
Predictors	**Estimates**	**CI**	***p***
(Intercept)	3.45	3.24 to 3.66	**<.001**
Accessibility	0.64	0.33 to 0.94	**<.001**
Explainability	0.12	−0.18 to 0.42	.429
Access × explain	−0.19	−0.63 to 0.25	.389
Observations		901	
R^2/R^2 adjusted		0.026/0.023	

Note: The bold values represent values that are significant at $p < .05$.

TABLE C2 Explainability

	The explanation provided about why my visa was rejected was sufficient to me. (1 = strong disagree to 7 = strong agree)		
Predictors	**Estimates**	**CI**	***p***
(Intercept)	2.65	2.45 to 2.86	**<.001**
Accessible	−0.07	−0.37 to 0.23	.639
Explainability	0.67	0.38 to 0.97	**<.001**
Access*explain	0.02	−0.41 to 0.45	.921
Observations		901	
R^2/R^2 adjusted		0.042/0.038	

Note: The bold values represent values that are significant at $p < .05$.

C.2 | Manipulation Check Study 2 (Welfare Fraud)

TABLE C3 Accessibility

	I feel that everyone who is willing and able could get access to the underlying code (the algorithm) of the computer system to check how it works. (1 = strong disagree to 7 = strong agree)		
Predictors	**Estimates**	**CI**	***p***
(Intercept)	3.52	3.30 to 3.73	**<.001**
Accessible	0.66	0.35 to 0.97	**<.001**
Explainability	0.07	−0.24 to 0.38	.653
Access*explain	−0.14	−0.58 to 0.30	.531
Observations		900	
R^2/R^2 adjusted		0.031/0.027	

Note: The bold values represent values that are significant at $p < .05$.

TABLE C4 Explainability

	The explanation provided about why my visa was rejected was sufficient to me. (1 = strong disagree to 7 = strong agree)		
Predictors	**Estimates**	**CI**	***p***
(Intercept)	2.37	2.15 to 2.60	**<.001**
Accessible	0.22	−0.10 to 0.53	.177
Explainable	0.81	0.50 to 1.12	**<.001**
Access × explain	0.02	−0.42 to 0.47	.919
Observations		900	
R^2/R^2 adjusted		0.060/0.057	

Note: The bold values represent values that are significant at $p < .05$.

APPENDIX D: Assessing Order Effects

This appendix provides an additional analysis which only test the study 1 that came first in order. This assessment indicates whether order effects were present. In general, the results of this subset analyses are very similar to the main results. The main difference is that the subset analysis is underpowered and as a result the error margins are inflated. Still, the means and mean differences are high comparable (see figures D1 and D2).

TABLE D1 Study 1, First Order

	Trust in Algorithm			Trust in Bureaucrat		
Predictors	**Std. Beta**	**Standardized CI**	***p***	**Std. Beta**	**Standardized CI**	***p***
(Intercept)	−0.11	−0.27 to 0.06	**<.001**	−0.10	−0.27 to 0.07	**<.001**
Accessible	0.06	−0.19 to 0.31	.625	0.11	−0.14 to 0.36	.388
Explainability	0.20	−0.05 to 0.46	.112	0.13	−0.12 to 0.38	.320
Access × explain	−0.06	−0.43 to 0.31	.733	−0.04	−0.41 to 0.33	.837
Observations		457			457	
R^2		0.008			0.006	

Note: The bold values represent values that are significant at $p < .05$.

TABLE D2 Study 2, First Order

	Trust in Algorithm			Trust in Bureaucrat		
Predictors	**Std. Beta**	**Standardized CI**	***p***	**Std. Beta**	**Standardized CI**	***p***
(Intercept)	−0.23	−0.42 to −0.05	**<.001**	−0.21	−0.40 to −0.03	**<.001**
Accessible	0.19	−0.08 to 0.46	.160	0.18	−0.09 to 0.45	.184
Explainability	0.35	0.09 to 0.60	**.007**	0.30	0.05 to 0.56	**.018**
Access × explain	−0.16	−0.53 to 0.21	.394	−0.14	−0.51 to 0.23	.466
Observations		453			452	
R^2		0.022			0.018	

Note: The bold values represent values that are significant at $p < .05$.

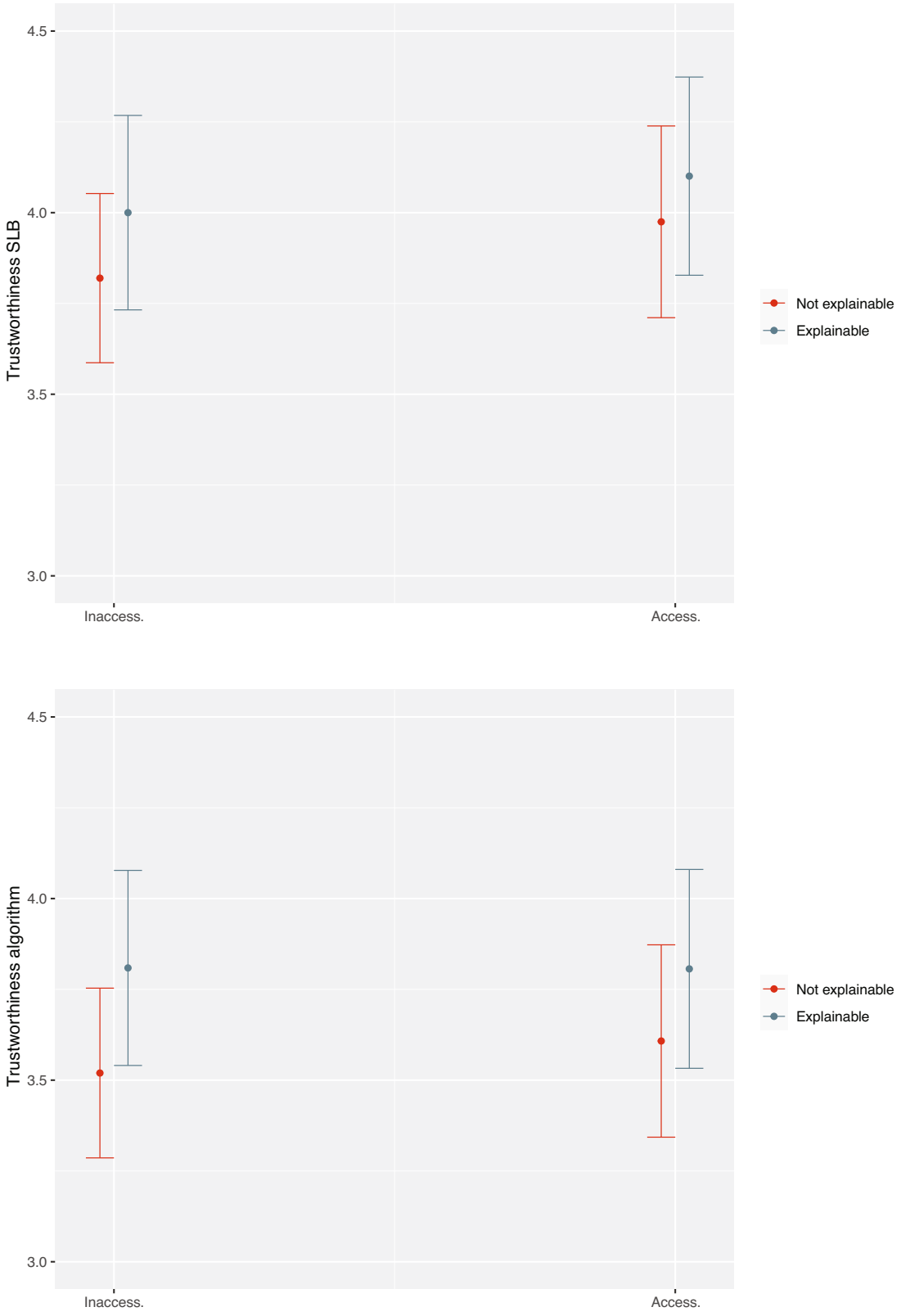

FIGURE D1 Study 1, First Order

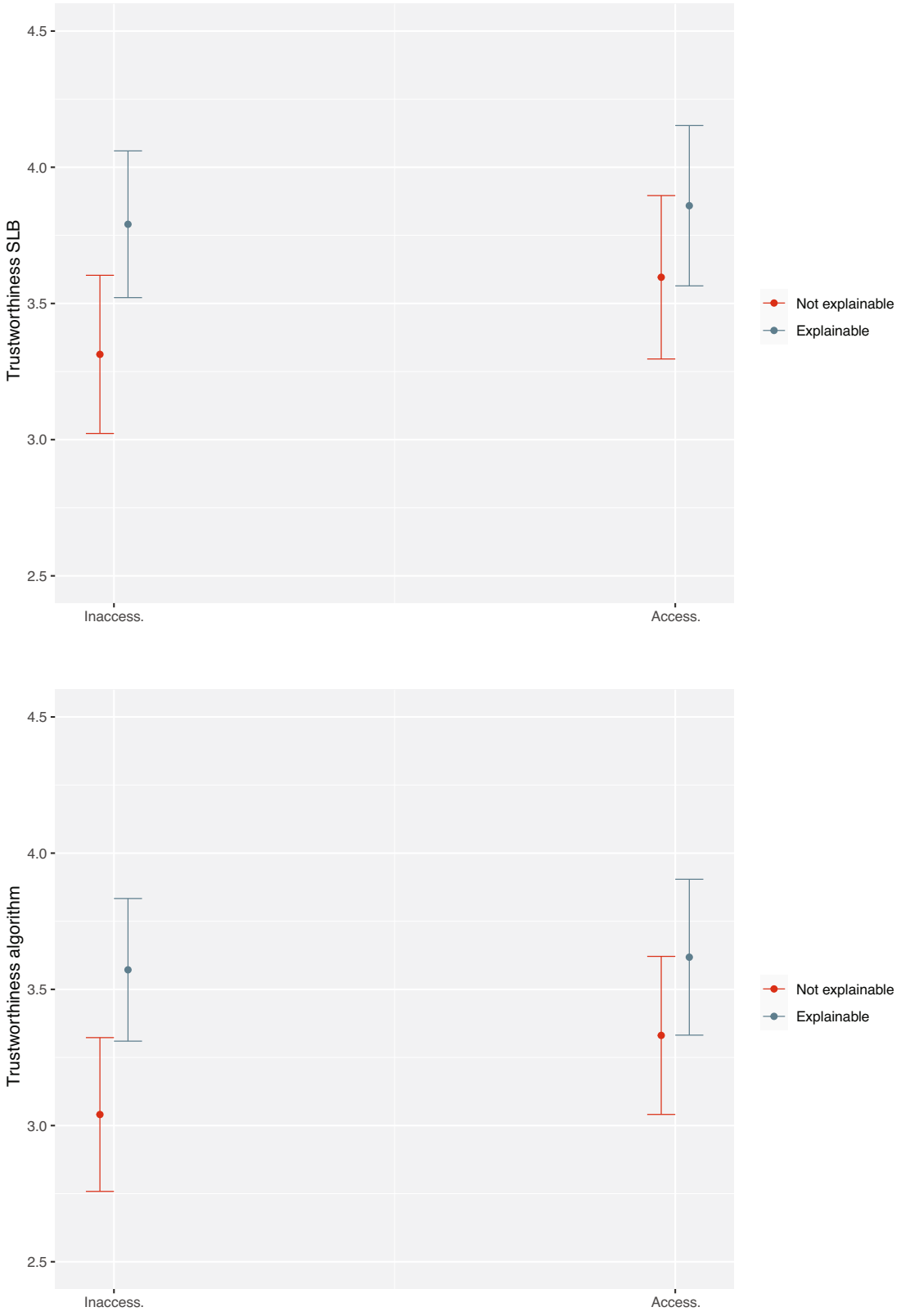

FIGURE D2 Study 2, First Order

APPENDIX E: Three-Way Interaction Analysis

	Trust in Algorithm		Trust in Bureaucrat	
Predictors	**Std. Beta**	***p***	**Std. Beta**	***p***
(Intercept)	−0.13	**<.001**	−0.12	**<.001**
Accessibility	0.15	.095	0.16	.076
Explainability	0.32	**<.001**	0.16	.082
Study1 [Study 2]	−0.21	**.020**	−0.09	.324
Access × explain	−0.17	.207	−0.06	.629
Access × study	0.13	.331	0.02	.865
Explain × study	0.07	.591	0.14	.291
Access × explain × study 2	−0.04	.832	−0.13	.486
Observations	1,810		1,808	
R^2/R^2 adjusted	0.028/0.024		0.012/0.008	

Note: The bold values represent values that are significant at $p < .05$.

APPENDIX F: Distraction Task Between Study 1 and Study 2

Here you view two pictures with holiday destinations. Of these two, what is your favorite destination?