Received: 11 February 2022 | Revised: 9 December 2022 | Accepted: 13 December 2022

DOI: 10.1111/puar.13602

RESEARCH ARTICLE

PAR PUBLIC ADMINISTRATION REVIEW ASPA

'Just like I thought': Street-level bureaucrats trust AI recommendations if they confirm their professional judgment

Friso Selten [1] | **Marcel Robeer** [2] | **Stephan Grimmelikhuijsen** [3]

[1]Institute of Public Administration, Leiden University, The Hague, The Netherlands

[2]National Police Lab AI, Utrecht University, Utrecht, The Netherlands

[3]Utrecht University School of Governance, Utrecht, The Netherlands

Correspondence
Friso Selten, Institute of Public Administration, Leiden University, Turfmark 99, 2511 DP, The Hague, The Netherlands.
Email: f.j.selten@fgga.leidenuniv.nl

Funding information
Dutch National Science Foundation, Grant/Award Number: 406.DI.19.011

[Correction added on 6 February 2023, after first online publication: The copyright line was changed.]

Abstract

Artificial Intelligence is increasingly used to support and improve street-level decision-making, but empirical evidence on how street-level bureaucrats' work is affected by AI technologies is scarce. We investigate how AI recommendations affect street-level bureaucrats' decision-making and if explainable AI increases trust in such recommendations. We experimentally tested a realistic mock predictive policing system in a sample of Dutch police officers using a 2 × 2 factorial design. We found that police officers trust and follow AI recommendations that are congruent with their intuitive professional judgment. We found no effect of explanations on trust in AI recommendations. We conclude that police officers do not blindly trust AI technologies, but follow AI recommendations that confirm what they already thought. This highlights the potential of street-level discretion in correcting faulty AI recommendations on the one hand, but, on the other hand, poses serious limits to the hope that fair AI systems can correct human biases.

Evidence for practice

- Artificial Intelligence-based recommendations play an increasingly important role in supporting decision-making by street-level bureaucrats, such as police officers.
- Street-level bureaucrats trust and follow AI recommendations that are congruent with their intuitive professional judgment.
- AI systems do not overturn intuitive professional judgments, even if they are well-explained.

Artificial Intelligence (AI) is rapidly changing public organizations across the globe (Young et al., 2019). Specifically, machine learning approaches not only automate routine administrative tasks, but are used to design AI systems that improve the quality of discretionary decision-making of street-level bureaucrats by steering their judgment (Bullock, 2019; Zouridis et al., 2020). However, how street-level bureaucrats interact with AI systems can be complex. For instance, a predictive policing system might recommend a police officer to surveil in a certain area, while the police officer thinks that other neighborhoods have much higher crime risks. Similarly, an AI system might recommend that a defendant should be released on parole, while the judge believes the defendant should remain in custody (Brayne & Christin, 2021). Street-level bureaucrats, confronted with such a dilemma, have to decide: do they follow the AI recommendation or their own intuitive professional judgment?

Scholars have noted that the empirical knowledge of the impact of AI on street-level bureaucrats' behavior is limited (Giest & Grimmelikhuijsen, 2020; Peeters, 2020). Therefore, the first aim of this article is to investigate what happens when AI recommendations are congruent or incongruent with a street-level bureaucrats' intuitive professional knowledge, that is, their expertise based on training activities and on-the-ground experience (Maynard-Moody & Musheno, 2000). We test two prominent and competing theories from psychology to better understand how professional knowledge and AI recommendations interact: automation bias and confirmation bias.

On the one hand, the use of AI can restrict the exercise of frontline discretion because decision makers are overconfident in the rationality of AI (Skitka et al., 1999; Young et al., 2019). Such *automation bias* leads users to wrongfully neglect evidence that originates from outside a computer system. Automation bias has been found in highly automated environments such as aviation and health care (Lyell & Coiera, 2017), and would indeed suggest that street-level decision-making is strongly affected by computer prompts. On the other hand, from the literature on motivated reasoning and *confirmation bias*, we know that individuals tend to stick with a preferred conclusion and that this leads to selective and biased information processing (Kunda, 1990; Taber & Lodge, 2006). Such confirmation bias also occurs amongst those who are more knowledgeable about a topic (Mendel et al., 2011). This would suggest that street-level bureaucrats will not follow all AI recommendations but ignore AI-generated output in case it contradicts their professional knowledge.

The second aim of this paper is to investigate how explainable AI (XAI) affects the trustworthiness and acceptance of AI recommendations. There is a growing demand for AI that not only performs well, but that is also transparent, explainable, and trustworthy (Giest & Grimmelikhuijsen, 2020). This is the goal of a specific area of AI research called explainable AI (XAI) (Adadi & Berrada, 2018). Research into XAI demonstrated that explanations enabled spotting algorithmic mistakes (Ribeiro et al., 2016). At the same time, XAI can have negative effects. Van der Waa et al. (2021) demonstrated that explanations can persuade users to follow incorrect recommendations. Overall, the empirical knowledge on the impact of XAI is limited, specifically within complex public decision-making processes (Giest & Grimmelikhuijsen, 2020; Peeters, 2020).

This article investigates the effects of AI recommendations on a typical street-level bureaucrat: the police officer (Lipsky, 2010; Maynard-Moody & Musheno, 2003). Investigating the effects of AI recommendations on police officers is especially relevant as the police force is one of the largest public-sector areas in which AI systems are being implemented that can heavily infringe on people's lives (e.g. Meijer et al., 2021). In addition, in many countries the police are at the forefront of AI adoption. Police organizations use AI systems to, for instance, forecast high crime risk areas, pre-identify young offenders, analyze vehicle movement patterns, and assist citizens with crime reporting (Dechesne et al., 2019; Meijer & Wessels, 2019).

At the same time, police work cannot be completely automated given the high degree of uncertainty and political sensitivity associated with the tasks they perform (Bullock et al., 2020). Police officers and AI systems, therefore, have to interact and collaborate. In the present study, we research this interaction by investigating how police officers utilize AI recommendations that are congruent and incongruent with their professional judgment and how explainable AI affects how they perceive these recommendations. We investigate the following research question:

What is the effect of AI recommendations and explainable AI on decision-making of street-level police officers?

To answer this question, we designed a 2 × 2 repeated measures factorial vignette experiment in which we tested how police officers interact with a realistic mock AI system that assists police officers with fencing off the area of a crime. This application is based on an AI system currently being developed by the Dutch police. A population-based sample of 124 street-level police was recruited for the experiment. Participants completed three similar vignettes with high mundane realism, resulting in 294 observations in total. Participants were exposed to a combination of the following two factors: an AI recommendation that was congruent or incongruent with their intuitive professional knowledge (first factor), and an AI recommendation that was explained or unexplained (second factor).

The results of this study indicate that police officers only trust AI recommendations that confirm what they already thought; police officers have more trust in AI recommendations that are congruent with their professional knowledge than AI recommendations that are incongruent with their professional knowledge. This implies that rather than being subject to automation bias, street-level bureaucrats are prone to confirmation bias when interacting with AI systems. Moreover, we found that police officers' trust in AI recommendations is not affected by AI-generated explanations (XAI), meaning that it will be hard to overturn intuitive professional judgments, even if AI recommendations are well-explained. In the next section, we will first discuss the role of AI in street-level decision-making and then turn to formulating and testing three hypotheses on how (explained) AI recommendations are expected to impact street-level decision makers.

ARTIFICIAL INTELLIGENCE IN STREET-LEVEL BUREAUCRACY

Street-level decision-making is characterized by the exercise of administrative discretion (Maynard-Moody & Musheno, 2003, 9). Exercising administrative discretion is necessary because of a mismatch between general rules and their application in specific local situations. Public officials are expected to base their decisions on pre-defined laws, procedures, and standards but these rules hardly ever fully correspond to the complex local realities of frontline work. Street-level bureaucrats translate general rules and competing values into client-level decisions (Lipsky, 2010). This constitutes administrative discretion: "the freedom that street-level bureaucrats have in determining the sort, quantity, and quality of sanctions and rewards during policy implementation" (Tummers & Bekkers, 2014, p. 529).

Administrative discretion has positive and negative consequences. The advantage of administrative discretion is that it allows for experience, local knowledge, sympathy, empathy, insight, and flexibility in frontline work

(Maynard-Moody & Musheno, 2003). Administrative discretion allows for targeting decisions to the specifics of the local situation. However, these discretionary practices do not only have desirable consequences. The translation of general rules into local decision-making is grounded in imperfect information and the street-level bureaucrat's conceptions of justice and appropriate action (Tummers & Bekkers, 2014). As a result, human decision-making is subject to cognitive limitations and bounded rationality (Kahneman, 2013; Simon, 1957). Consequently, administrative discretion has been linked to reduced policy-making effectiveness and efficiency, biased and discriminatory decision-making processes, and unlawful and corruptive behavior (Binns, 2020; Young et al., 2019). These adverse outcomes of administrative discretion imply that it should be controlled (Davis, 1970).

Artificial Intelligence (AI) is a set of technologies that can be used to exercise control over administrative discretion. AI is an umbrella term for systems that display intelligent behavior by, with some level of autonomy, reacting to their environment to achieve specific goals (Zuiderwijk et al., 2021). AI systems can be rule-based, but modern AI systems, especially those used to improve front-line decision-making, often employ machine learning techniques (Grimmelikhuijsen & Meijer, 2022). Machine-learned AI systems, then, are different from traditional statistical modeling as there is no formalization or a priori theorization of relationships between variables (Athey & Imbens, 2019). AI systems that use machine learning can make complex decisions in individual cases by analyzing available information and by making inferences based on the extent to which this case shares characteristics with a group of other cases. From a technical-rational perspective, AI systems can use this to information to reduce human arbitrariness in street-level decision-making, thereby enhancing public decision-making accuracy, consistency, objectivity, and efficiency (Binns, 2020; Young et al., 2019).

However, AI systems also have their limitations. Street-level decision-making requires individual judgment, but AI systems are not capable of providing true case-by-case decisions because it is impossible to capture all relevant local aspects in a mathematical model (Binns, 2020). The use of AI, therefore, limits an organization's ability to make appropriate judgments about individual cases (Bannister & Connolly, 2020). Moreover, AI systems often produce disproportionately adverse outcomes for disadvantaged groups as demonstrated by, for example, O'Neill (2016) in *Weapons of Math Destruction*. When not implemented properly, AI systems, being trained on historical data collected by humans, reproduce existing human decision-making biases and errors. Because of these potential negative effects of automation, AI systems are often implemented as 'decision-support systems' rather than as autonomous agents (Veale & Brass, 2019). In decision-support arrangements, AI systems inform and augment decision-making, but a human decision maker is kept 'in-the-loop' and makes the final decision (Bullock, 2019; Busuioc, 2021).

Having a human-in-the-loop is, however, no guarantee to correct errors made by an AI system when decision makers become subject to automation bias (Peeters, 2020). Automation bias is a prominent decision-making error that refers to "the use of automated cues as a heuristic replacement for vigilant information seeking and processing" (Mosier et al., 1998, p. 48). Multiple types of automation bias exist but AI-based recommendations are especially prone to inducing errors of commission in decision-making processes. Errors of commission occur when a decision maker trusts and follows an AI recommendation even in the face of other indicators showing that this recommendation is illogical (Skitka et al., 1999). Automation bias and commission errors have been extensively documented in other industries that have been automated, such as aviation and health care, where it resulted in new decision-making errors (Lyell & Coiera, 2017).

Concerns have been raised that street-level decision makers will also be susceptible to automation bias (Peeters, 2020; Zerilli et al., 2019) but a first experimental test by Alon-Barkat and Busuioc (2022) suggests otherwise. In their experimental study, they found that decision makers may not be susceptible to automation bias because AI recommendations did not over-ride the decision makers' existing stereotypes and discriminatory biases. While such biases are important to investigate, our study takes a different perspective and focuses on whether street-level bureaucrats' professional—and often intuitive—knowledge affects the use of AI recommendations. With professional knowledge, we refer to the factors that guide the behavior of street level-bureaucrats, that is, their expertise based on training activities and on-the-ground experience (Maynard-Moody & Musheno, 2000).

Indeed, explorations in case studies by for example Meijer et al. (2021) and Snow (2021) suggest that street-level bureaucrats place a great deal of trust in these professional intuitions and it is unlikely that AI recommendations will be trusted if they go against their intuitive professional knowledge. These qualitative findings speak to the literature on motivated reasoning and confirmation bias, which describes that individuals only selectively use information if it confirms their initial or preferred conclusions (Kunda, 1990; Taber & Lodge, 2006). Moreover, Mendel et al. (2011) demonstrated that not only laymen are subject to confirmation bias, but this bias also occurs amongst those who are more knowledgeable about a topic. These findings suggest that police officers might not be subject to automation bias, but are more likely to be prone to confirmation bias when interpreting AI recommendations and therefore only trust AI recommendations that are congruent with their professional judgment. To test these two prominent and competing theories from psychological science, we propose the following hypothesis:

> **H1.** *Street-level bureaucrats perceive AI recommendations that are congruent with their professional judgment as more trustworthy than AI*

recommendations that are incongruent with their professional judgment.

EXPLAINABLE AI AND TRUSTWORTHINESS

Explaining AI recommendations using Explainable AI (XAI) techniques has been argued to be fundamental to increase the trustworthiness of AI by preventing automation-induced decision-making biases (Ahmad et al., 2018). The second step in this research is, therefore, investigating how XAI affects the perceived trustworthiness of AI recommendations.

XAI is expected to increase users' understanding of AI systems. It gives decision makers insight into how the recommendation is constructed and therefore enables informed decision-making. According to Doran et al. (2017), XAI should provide understanding to non-technical audiences by answering the *why-question*. Truly explainable AI that supports frontline work should not only display which data were used but also provide a line of reasoning about why and in what way that data were used.

Two approaches are distinguished in the XAI literature to answer this why-question: global explanations that explain the AI systems' procedures in general, or local explanations that explain outcomes in specific situations (Adadi & Berrada, 2018; Ahmad et al., 2018). Global explanations explain the functioning of the AI system; they explain the general procedures the AI uses. Global explanations do not show how a specific recommendation was constructed. Explanations that do provide recommendation-specific insights are referred to as local explanations (Ribeiro et al., 2016). Because the discretionary tasks of street-level bureaucrats involve case-by-case judgment, local explanations are most suited to support frontline work. This research therefore explicitly focuses on the effect of local explanations.

Specifically, in our experiment we provide the participants with explanations that are every day, contrastive, and simple; as explanations with these characteristics have been found to be most effective and persuasive for a non-technical audience (Miller, 2019). Everyday explanations provide explicit cues about why an AI system constructed a specific recommendation (Lipton, 1990). Contrastive explanations demonstrate why this specific recommendation was advised in comparison to a recommendation that was not advised (Mercado et al., 2016). Formulating simple explanations has been demonstrated to be more effective in communicating information to users than long and complex explanations (Thagard, 1989).

The idea that XAI increases user trust is grounded in the reasoning that people will not trust systems they do not understand (Giest & Grimmelikhuijsen, 2020). Unexplained AI systems are opaque and not understandable to decision makers (Burrell, 2016). Explaining the AI systems' functioning, which is expected to increase understandability, is therefore related to stimulating the use and trust of AI recommendations (Grimmelikhuijsen, 2023; Schiff et al., 2022). So far most studies that have investigated the use of AI in public decision-making have focused on citizens and not street-level bureaucrats. Still, based on this empirical evidence we hypothesize that similar effects may be found in this group:

> **H2.** *Street-level bureaucrats perceive explained AI recommendations as more trustworthy than unexplained AI recommendations*.

The two hypotheses formulated above relate professional knowledge and XAI to the perceived trustworthiness of AI recommendations. While it is crucial to understand what factors affect the perception of AI recommendations by street-level bureaucrats, it is important to consider how this perception actually affects their behavior. How does the perceived trustworthiness of AI recommendations affect the decision of street-level bureaucrats to follow or override an AI recommendation?

Information systems research highlights that the perceived trustworthiness of a technological system affects whether people will use this system (Kim et al., 2008; McKnight et al., 2011). Therefore, as a final step in this study, in H3 we relate an increase in perceived trustworthiness of an AI recommendation to an increase in the likelihood that street-level bureaucrats will follow this recommendation. The hypotheses tested in this study are summarized in Figure 1.

> **H3.** *Higher perceived trustworthiness of AI recommendations by street-level bureaucrats is related to an increased likelihood of following the recommendation*.

EXPERIMENTAL METHODS AND MEASUREMENTS

The first aim of this article is to study how police officers perceive AI recommendations and the extent to which AI recommendations induces automation bias or confirmation bias in street-level decision-making. The second aim is to research the effects of XAI. These two aims were tested in a repeated measures factorial vignette experiment in a sample of street-level police officers.

More specifically, participating police officers were presented with three scenarios (vignettes) in an online survey: a *burglary*, an *ATM robbery*, and a *stabbing incident*. Each of these scenarios involved a description of a crime that had been committed. In a dispatch report, officers were asked to help fence off the area of this crime. In this task, they were assisted by a mock AI system. This system predicted the flight routes of offenders. Participants were presented with two locations alongside predicted escape routes. The experimental task was choosing one of these by the AI-assigned locations.

This experiment has high ecological validity. The experimental task was selected to appeal to a wide range

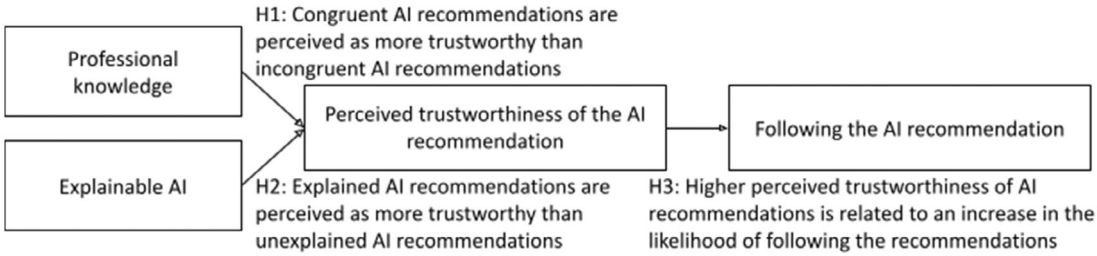


FIGURE 1 Hypotheses and expected relationships

of police officers. All street-level police officers have experience—in training and mostly also in practice—with this type of flight situation. The mock AI system was based on a system that is developed by the Dutch police organization. Additionally, the ecological validity of the design was strengthened by incorporating feedback from two police officers and three academic experts on the (socio-)technical sides of AI.

The initial experimental design was then presented to a group of Ph.D. candidates who research the use of AI within the Dutch National Police and qualitatively tested in collaboration with two street-level police officers. Based on the feedback of these groups, textual changes were made so that the experimental design better reflected the everyday reality of police work. Lastly, a small pilot study was conducted amongst a lay audience ($n = 10$). This is to ensure that the design of the experiment was clear and understandable.

Data collection

This study obtained a population-based sample. Data were gathered in collaboration with the Dutch National Police. The survey was distributed to four contact persons within four local police departments. These forwarded the invitations within their team to police officers that were or had been active in street-level work. The initial invite was sent in the second week of May 2021, and a reminder in the third or fourth week. Data collection was stopped on June 3. Investigating the influence of AI on frontline work amongst a population-based sample is important, but rare. Conducting this study with actual police officers is especially relevant in light of its focus on decision-making biases and XAI. Both of which are influenced by prior knowledge of the decision maker.

The only inclusion criterium used in this study was that participants completed all questions of at least one of the three scenarios, meaning no attention checks were used to remove participants. Including data from participants that did not fully concentrate on the experiment means that we measured the overall effect of the experimental conditions, that is, the intent to treat effect. This effect is more consistent with the effects the AI recommendations will have when implemented in real decision-making procedures, which strengthens the external validity of the results (Hansen & Tummers, 2020).

The survey, as distributed by local police departments, was sent to approximately 400 street-level police officers. A total of 152 police officers responded to the survey. 28 participants did not satisfy the inclusion criterium. Consequently, the final sample consists of 124 police officers. Together, they completed 294 vignettes, as some participants completed only one or two vignettes. Data and materials can be found in the Open Science framework via: doi.org/10.17605/OSF.IO/R6QAE.

Sample description

A representative sample of street-level police officers was obtained. The average age in our sample is approximately 47.9, and 25% of the participants identified as female. This is comparable to the averages in the Dutch police, where the average age is 45.2, and in which 34.7% of employees identify as female (Politie, 2020). 93% of the participants indicated having an executive status. This means that almost all participants were qualified to conduct street-level police work. Given that we specifically asked distributors of the experiment to only forward the invitations to police officers that are or have been active in street-level work we assume the other seven percent of officers also have experience with street-level tasks. Hence, we did not remove participants based on this criterium. Furthermore, most participants had post-secondary vocational education. This is the education level required for street-level police work in the Netherlands. Accordingly, the obtained sample is representative of the target population. Table A1 in Appendix A presents an overview of these descriptive statistics per experimental group and balance tests revealed no significant differences between the four treatment groups with regard to these variables.

Additionally, we measured three characteristics that could potentially influence the trust of decision makers in automated advice: knowledge about algorithms, knowledge about the mock-algorithm, and general trust in technology. As can be seen in Table A1 in Appendix A, balance tests reveal no statistically significant difference between the four treatment groups on the first two variables. However, the test demonstrate a small but statistically significant difference between the experimental groups' general trust in technology variable. Post-hoc analysis reveal the group that received the congruent

FIGURE 2 The map used in the ATM robbery scenario

unexplained advice had statistically significantly more trust in technology than the participants in the incongruent unexplained advice group. However, given the small difference between the groups (less than 0.6 point on a scale of 1–7) and the possibility this variation is a product of the experimental manipulations (the mock AI's congruent or incongruent advice may have increased participants' trust in technology), we do not believe that this difference had a substantial effect on the study's findings.

Experimental design

This study presented participants with three vignettes. These vignettes were described in short text fragments. The most important information was also displayed in a figure that represented a map of the area surrounding the crime (Figure 2 presents an example).

The first factor in the vignettes was used to assess how street-level bureaucrats' intuitive professional knowledge affects their perceived trustworthiness of AI recommendations. In each scenario, the mock AI system advised one location that was expected to be congruent with the professional judgment of police officers and one location that was incongruent with their professional judgment. The congruence of a location was determined based on the multiple rounds of qualitative interviews with police officers. For example, in the ATM robbery vignette presented in Figure 2, police officers indicated offenders of this type of crime almost always escape using fast cars via a highway. Location A in this vignette, therefore, is congruent with the professional judgment of the police officers because this location is close to the highway, while location B is incongruent with their professional judgment.

The second factor manipulated whether the AI recommendation was explained to the police officers. The explanations provided by the AI were, as highlighted in the theoretical section, simple, explicit, and contrastive—explanations that provide understandability for the non-technical audience of this study. The specific content of the explanations was grounded in the qualitative feedback of police officers. For example, in the ATM robbery vignette, the explanation provided was: "Location A was recommended by the AI because suspects of robberies often flee via the highway. Location B was recommended because the suspects cannot then flee into the neighboring town via this route". Table 1 summarizes the four experimental groups created in each of the three vignettes and Table B1 in Appendix B presents a detailed overview of the three vignettes and the images that were presented in the burglary and stabbing incident vignettes.

TABLE 1 Type of recommendation per experimental group

	Unexplained	Explained
Incongruent	Group 1	Group 3
Congruent	Group 2	Group 4

Randomization gives a repeated measures factorial survey the robustness of an experimental method (Taylor, 2006). Hence, scenarios were displayed to participants in random order and in each scenario participants were randomly assigned to one of four experimental groups. A schematic overview of the full experimental procedure is presented in Appendix C.

Dependent variables

This study measured the effect of AI recommendations on two variables. The first dimension investigated was the perceived trustworthiness of the recommendations by police officers. This was measured using a scale developed by Grimmelikhuijsen (2023) to measure the perceived

TABLE 2 Manipulation checks

				Significance test			
	N	Mean	SD	*t*	Df	*d*	*p*
Congruency of the recommendation	294	4.65	1.63	5.65	288.97	0.64	<.001
Effect of explanation	294	4.49	1.48	3.34	262.47	0.40	<.001

trustworthiness of AI systems. This scale is based on scales that have been developed to measure trust more generally (e.g. Grimmelikhuijsen & Knies, 2017; Mayer et al., 1995), and scales designed to evaluate the trustworthiness of technological systems specifically (McKnight et al., 2011).

This perceived trustworthiness of the AI recommendations scale includes the following four items: 'I trust that the AI…': (1) '…used the correct information', (2) '…gave a correct recommendation, (3) '…assessed my situation honestly', (4) '…used all relevant information' [translated from Dutch]. These items all have been measured on a 1 (no trust at all) to 7 (complete trust) scale. To validate this scale's measurements, we included a fifth item that directly asked participants whether they trusted the AI recommendation. These five items combined form a reliable scale to measure the perceived trustworthiness of the AI recommendation (Cronbach's Alpha above .92 and a principal component analysis demonstrated all variables load on the first component with factor loadings above .74).

Secondly, we assess whether or not the perceived trustworthiness of the AI recommendation had an effect on the likelihood of police officers following the AI. To conduct this analysis, a dummy variable was created in which 0 indicates a choice against the AI (i.e., the police officer chose a different location than the AI recommended), and 1 indicates going along with the AI (i.e., the police officer chose the location that was recommended by the AI).

Manipulation checks

Two manipulation checks were used to assess whether the experimental treatments—the congruence of the advice and the effect of XAI—were successful. Table 2 presents the results of these checks.

The first manipulation check assessed if one of the two AI recommendations was congruent with the professional knowledge of the police officers. This was measured by asking, on a 1 (strongly disagree) to 7 (strongly agree) scale, if respondents found that the AI recommendation they received aligned with their professional judgment. As indicated in Table 2 this manipulation was successful: police officers perceived the locations that were designed to be congruent with their own knowledge to be more in line with their professional judgment.

The second manipulation check assessed how participants experienced the provided explanations about the AI recommendation. This was investigated by asking respondents on a 1 (strongly disagree) to 7 (strongly agree) scale how detailed respondents perceived the recommendation to be. As indicated in Table 2, this manipulation was also successful: police officers perceived the explained AI recommendation to be more detailed than the unexplained AI recommendation.

Lastly, to obtain insight into the mundane reality of the experiment, participating police officers were asked if they could relate to the experimental scenarios. On average, police officers indicated that they could somewhat agree or agree with this statement (mean score = 5.24, SD = 1.25). These are good scores given that survey experiments always contain a degree of artificialness (Jilke et al., 2017). This experiment can thus be considered to have a high mundane reality, which strengthens the ecological validity of the research findings.

Deviations from pre-analysis plan

Research ethics were taken into consideration when conducting the experiment. Most importantly, the research question, hypotheses, exclusion criteria, and measurements were registered in the Open Science Framework prior to the execution of the experiment. The pre-registry can be viewed at this link: https://doi.org/10.17605/OSF.IO/TWY9V but during the execution of the study some changes were made to the hypotheses, experimental design, and analytical approach. First, the hypotheses in the article were rephrased to be more precise than those that were preregistered but they identify the same relationships. Secondly, one experimental group, a group that did not receive an AI recommendation, was removed from the analyses because this group was not suitable to test any of our preregistered hypotheses. Thirdly, instead of the pre-registered traditional regressions, Generalized Estimation Equations were used to analyze the data, as this analytical approach accounts for possible within-subject effects introduced by the repeated measures factorial design when deriving the variability estimates of the coefficients (see Hubbard et al., 2010 for a detailed description of Generalized Estimation Equations). A full description of the changes from the preregistration is provided in Appendix D.

RESULTS

The aim of this study was to investigate how professional knowledge and XAI affect the perceived trustworthiness

TABLE 3 Number of respondents and mean trust in the AI per experimental group

	Unexplained			Explained		
	N	Mean	SD	N	Mean	SD
Incongruent	72	4.73	1.26	90	4.98	1.17
Congruent	56	5.26	1.10	76	5.25	1.18

of AI recommendations by street-level bureaucrats. Additionally, it was assessed how an increase in the perceived trustworthiness of AI recommendations was related to the decisions of street-level bureaucrats to follow the AI recommendations. Table 3 describes the number of participants and the mean perceived trustworthiness of the AI recommendation per experimental group.

Since respondents completed multiple vignettes, the observations in our sample are not completely independent. In these circumstances, Generalized Estimating Equations (GEEs) are found to provide closer approximations of population averages than basic regression approaches which assume independence of observations (Hubbard et al., 2010; Liang & Zeger, 1986). GEEs provide a semiparametric way to analyze clustered data and are comparable to a repeated measures ANOVA, but achieve higher power with a smaller sample size and lower number of repeated measurements, and can be used to analyze categorical responses as well as continuous outcome variables (Ma et al., 2012). Because these aspects apply to our study design, GEE analyses were performed to analyze our data using the *Geepack* package for R (Halekoh et al., 2006). More specifically, GEEs were used to compute the Wald Chi-Square coefficient, the 95% confidence intervals, and associated p values. To increase the robustness of the results, we also tested the three hypotheses with traditional statistical approaches that assume independence of observations. Tables E1 and E2 in Appendix E present the outcomes of these analyses and reveals that the results of the GEEs and the traditional statistical approaches are equal.

First, a GEE was used to investigate how police officers utilize AI recommendations that are congruent and incongruent with their professional judgment (H1) and how explainable AI affects their perception of these recommendations (H2). We also explored an interaction between the congruence and explanation manipulations but found no significant effect ($\beta = -.259$ [−.888, .371], $p = .42$). Presented in the article are therefore the outcomes of the GEE without interaction effect. The results of this analysis are highlighted in Figure 3.

These results indicate police officers statistically significantly perceive AI recommendations that are congruent with their professional knowledge as more trustworthy than AI recommendations that are incongruent with their judgment ($\beta = -.377$ [−.622, −.132], $p = .003$). This finding is in line with our expectations outlined in H1.

In contrast, we found no support for H2. While Figure 3 seems to show the existence of some effect of

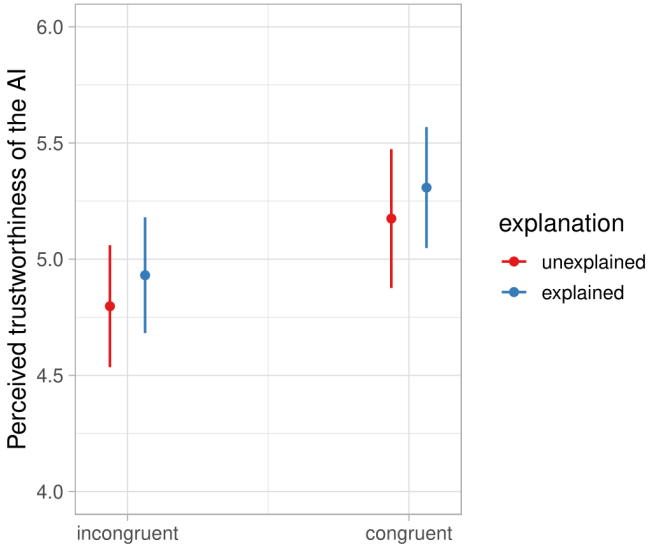


FIGURE 3 The effect of professional judgment and explanations on police officers' perceived trustworthiness of AI recommendations

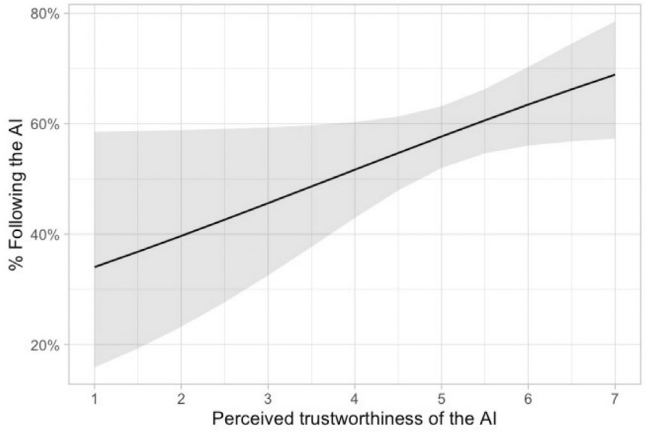


FIGURE 4 The relation between police officers' perceived trustworthiness of AI recommendations and their tendency to follow the recommendation

explaining the AI recommendation on how trustworthy the police officers perceived the AI recommendation to be, the GEE demonstrates that this effect was not statistically significant ($\beta = -.133$ [−.400, .134], $p = .328$).

As a final step in our research, a GEE with a logit link function was used to investigate how changes in perceived trustworthiness of AI recommendations are related to a police officer's choice to follow the recommendation (the police officer agrees with the AI and goes to the location that was recommended by the AI) or to oppose it (the police officer chooses to go to the other location). The results of this analysis, as presented in Figure 4, are in line with H3. An increase in perceived trustworthiness is statistically significantly related to an increase in the likelihood of police officers following the AI recommendation ($\beta = .243$ [.006, .480], $p = .045$).

DISCUSSION AND CONCLUSION

Our findings provide support for the effect of professional knowledge on the perceived trustworthiness of AI recommendations (H1), but not for the effect of providing explanations (H2). Additionally, an increase in the perceived trustworthiness of AI recommendations was found to be related to an increase in the likelihood of police officers following the AI recommendation (H3). These results hold three core conclusions.

The first conclusion is that the risk of automation bias appears to be less prominent in frontline decision-making than in other domains that are automated, instead street-level bureaucrats appear to be prone to confirmation bias when interpreting AI recommendations. This study reveals that police officers perceive AI recommendations that are congruent with their professional knowledge as more trustworthy than AI recommendations that are incongruent with their professional knowledge. This finding is in line with qualitative studies that indicated that decision makers weigh information provided by an AI system to their own knowledge (e.g., Meijer et al., 2021; Snow, 2021). Additionally, this study resonates with the findings by Alon Barkat and Busuioc (2021), who showed that decision makers are more likely to trust AI recommendations when they fit with existing stereotypes and biases. The use of AI for frontline decision-making is, however, relatively new and automation bias mainly occurs in fields with a long history of using highly reliable AI systems (Peeters, 2020). Future research should therefore investigate how repeated use of reliable AI systems by street-level bureaucrats affects the occurrence of automation bias and confirmation bias.

The second conclusion is that the positive effects associated with XAI might be less prominent than currently assumed in the literature (e.g. Miller, 2019). Our results suggest that a small positive effect of explaining both congruent and incongruent AI recommendations on the perceived reliability of these recommendations might exist, but this effect was not statistically significant. This demonstrates that the effect of prior knowledge is far more important for how street-level bureaucrats interpret and use AI recommendations than the effect of explaining the rationale behind these recommendations. Even though a study with a larger sample size might be able to detect a small effect of XAI, this raises questions about the meaningfulness of such a small effect in practice and provides a sobering message to the high expectations of explaining how AI systems function in the literature (e.g. Weller, 2019; Zerilli et al., 2019). We recommend future research to assess if other types of XAI, such as global explanations, have different effects.

Our third conclusion is that an increase in perceived trustworthiness is related to a change in the behavior of police officers. We found that police officers who perceived AI recommendations to be more trustworthy were also more likely to follow these recommendations. This is important for the implementation of AI in frontline decision-making tasks. AI can enhance the work of street-level bureaucrats, but also produce adverse outcomes when it is unfair, biased, or faulty (Veale & Binns, 2017). The results of this study, especially in combination with the first conclusion, show that street-level bureaucrats are likely to—at least to a certain extent—be able to mitigate these adverse outcomes. Street-level bureaucrats do not blindly trust and follow all AI recommendations but weigh such recommendations against their professional knowledge. Our study thereby provides micro-level evidence of the importance of maintaining human discretion to overturn AI recommendations when producing incongruent recommendations.

The persistence of human discretion, however, introduces new problems in human-computer interaction. On the one hand, it means a human-in-the-loop will not be able to correct all biases in an AI system. Our findings indicate that police officers will not correct errors when they align with police officers' personal biased judgment. This is especially worrisome as AI systems are prone to reproducing existing human decision-making biases and errors (O'Neill, 2016). On the other hand, the persistence of human discretion means unbiased AI systems—AI systems that have been de-biased through careful data gathering efforts, model building, and testing and validation—may not be able to persuade police officers to make another and likely fairer decision since recommendations from such AI systems are by design incongruent with the professional judgment of the human-in-the-loop.

This means that we cannot (only) rely on individual bureaucrats to assess the quality of AI systems. We need to design proper institutional and organizational safeguards when implementing AI technologies such as predictive policing system in frontline decision-making (e.g. Grimmelikhuijsen & Meijer, 2022). Case study research provides a valuable methodology to investigate these safeguards. For instance, Brayne and Christin (2021) showed that not only technical specifications, but also organizational policies and rules determine how police officers interact with AI systems. Meijer et al. (2021) found that administrative cultures and existing administrative patterns determined how similar predictive policing systems in the Netherlands and Germany were used. In the Netherlands, the system was used in a less restrictive manner and seen as a "helping hand", whereas in Germany police officers could less easily divert from recommendations by the predictive policing system.

Future experimental studies can also be directed at these institutional and organizational arrangements that influence how AI systems are used in practice. Firstly, while we devoted ample attention to creating an ecologically valid design and even though this experiment was perceived as realistic by participants, survey experiments always remain an artificial setting. Testing the occurrence of automation bias, confirmation bias, and the effects of XAI, therefore not only has to be done in the 'clean' context of the lab, but also in 'messy' and complex field settings. For instance, police officers in the field often work in small teams or experience pressure to make quick decisions, which may lead to different decision-making

dynamics. Here, we especially see value in the use of field experiments (e.g. Hansen & Tummers, 2020). Secondly, this study sought to understand the effect of AI recommendations on police officers. While frontline workers share similar characteristics (Maynard-Moody & Musheno, 2003), have previous studies highlighted that the use of AI systems is influenced by the specific type of AI system at hand and the organizational context in which this system is used (Bullock et al., 2020; Meijer et al., 2021). It would, therefore, be valuable to test the effects of a variety of AI systems on other types of street-level bureaucrats, such as teachers and welfare workers, and add situations in which intuitive professional knowledge is less clear.

In sum, the present study examined street-level bureaucrats' trust in and use of AI recommendations. Also, we investigated the role of explainable AI in this relationship. We find that street-level bureaucrats might be prone to confirmation bias when interpreting AI recommendations. In this study, police officers had more trust in AI recommendations when they were congruent with their professional judgment; they trusted the AI recommendations that confirmed what they already thought. Even recommendations that were supported by explanations were trusted more if they aligned with this professional judgment. This implies that street-level bureaucrats are able to correct biased and faulty AI systems, but it also means that even if AI recommendations are well-explained and accurate, it will be hard to persuade street-level bureaucrats to trust and follow-up counterintuitive AI recommendations.

ACKNOWLEDGMENTS

We thank Albert Meijer, Floris Bex, and Michel van Slobbe for their valuable comments on the research design and/or earlier versions of the article. We also thank three anonymous reviewers for their constructive feedback on the manuscript. We acknowledge funding by the Dutch National Science Foundation (NWO), under grant number 406.DI.19.011 (ALGOPOL).

REFERENCES

Adadi, Amina, and Mohammed Berrada. 2018. "Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6: 52138–60. https://doi.org/10.1109/ACCESS.2018.2870052.

Ahmad, Muhammad Aurangzeb, Carly Eckert, and Ankur Teredesai. 2018. "Interpretable Machine Learning in Healthcare." In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* 559–60. Washington DC USA: ACM. https://doi.org/10.1145/3233547.3233667.

Alon-Barkat, Saar, and Madalina Busuioc. 2022. "Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice." *Journal of Public Administration Research and Theory* 33(1): muac007. https://doi.org/10.1093/jopart/muac007.

Athey, Susan, and Guido W. Imbens. 2019. "Machine Learning Methods that Economists Should Know about." *Annual Review of Economics* 11(1): 685–725. https://doi.org/10.1146/annurev-economics-080217-053433.

Bannister, Frank, and Regina Connolly. 2020. "Administration by Algorithm: A Risk Management Framework." *Information Polity* 25: 471–90. https://doi.org/10.3233/IP-200249.

Binns, Reuben. 2020. "Human Judgment in Algorithmic Loops: Individual Justice and Automated Decision-Making." *Regulation & Governance* 16: rego.12358. https://doi.org/10.1111/rego.12358.

Brayne, Sarah, and Angèle Christin. 2021. "Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts." *Social Problems* 68(3): 608–24. https://doi.org/10.1093/socpro/spaa004.

Bullock, Justin B. 2019. "Artificial Intelligence, Discretion, and Bureaucracy." *The American Review of Public Administration* 49(7): 751–61. https://doi.org/10.1177/0275074019856123.

Bullock, Justin B., Matthew M. Young, and Yi-Fan Wang. 2020. "Artificial Intelligence, Bureaucratic Form, and Discretion in Public Service." *Information Polity* 25: 491–506. https://doi.org/10.3233/IP-200223.

Burrell, Jenna. 2016. "How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3(1): 205395171562251. https://doi.org/10.1177/2053951715622512.

Busuioc, Madalina. 2021. "Accountable Artificial Intelligence: Holding Algorithms to Account." *Public Administration Review* 81(5): 825–36. https://doi.org/10.1111/puar.13293.

Davis, Kenneth Culp. 1970. *Discretionary Justice: A Preliminary Inquiry*. Baton Rouge: Louisiana State University Press.

Dechesne, Francien, Virginia Dignum, Lexo Zardiashvili, and Jordi Bieger. 2019. *AI & Ethics at the Police* (Leiden/Delft) https://www.universiteitleiden.nl/binaries/content/assets/rechtsgeleerdheid/instituut-voor-metajuridica/artificiele-intelligentie-en-ethiek-bij-de-politie/ai-and-ethics-at-the-police-towards-responsible-use-of-artificial-intelligence-at-the-dutch-police-2019.pdf.

Doran, Derek, Sarah Schulz, and Tarek R. Besold. 2017. http://arxiv.org/abs/1710.00794. "What Does Explainable AI Really Mean? A New Conceptualization of Perspectives". *arXiv:1710.00794 [cs]*.

Giest, Sarah, and Stephan Grimmelikhuijsen. 2020. "'Introduction to Special Issue Algorithmic Transparency in Government: Towards a Multi-Level Perspective'. Edited by Sarah Giest and Stephan Grimmelikhuijsen." *Information Polity* 25(4): 409–17. https://doi.org/10.3233/IP-200010.

Grimmelikhuijsen, Stephan. 2023. "Explaining why the Computer Says No: Algorithmic Transparency Affects the Perceived Trustworthiness of Automated Decision-Making." *Public Administration Review* 83: 241–62. https://doi.org/10.1111/puar.13483.

Grimmelikhuijsen, Stephan, and Eva Knies. 2017. "Validating a Scale for Citizen Trust in Government Organizations." *International Review of Administrative Sciences* 83(3): 583–601. https://doi.org/10.1177/0020852315585950.

Grimmelikhuijsen, Stephan, and Albert Meijer. 2022. "Legitimacy of Algorithmic Decision-Making: Six Threats and the Need for a Calibrated Institutional Response." *Perspectives on Public Management and Governance* 28: gvac008. https://doi.org/10.1093/ppmgov/gvac008.

Halekoh, Ulrich, Søren Højsgaard, and Jun Yan. 2006. "The *R* Package Geepack for Generalized Estimating Equations." *Journal of Statistical Software* 15(2): 1–11. https://doi.org/10.18637/jss.v015.i02.

Hansen, Jesper Asring, and Lars Tummers. 2020. "A Systematic Review of Field Experiments in Public Administration." *Public Administration Review* 80(6): 921–31. https://doi.org/10.1111/puar.13181.

Hubbard, Alan E., Jennifer Ahern, Nancy L. Fleischer, Mark Van der Laan, Sheri A. Lippman, Nicholas Jewell, Tim Bruckner, and William A. Satariano. 2010. "To GEE or Not to GEE: Comparing Population Average and Mixed Models for Estimating the Associations between Neighborhood Risk Factors and Health." *Epidemiology* 21(4): 467–74. https://doi.org/10.1097/EDE.0b013e3181caeb90.

Jilke, Sebastian R., and Gregg G. Van Ryzin. 2017. "Survey Experiments for Public Management Research." In *Experiments in Public Management Research*, edited by Oliver James, Sebastian R. Jilke, and Gregg G. Van Ryzin, 117–38. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316676912.007.

Kahneman, Daniel. 2013. *Thinking, Fast and Slow*. 1st pbk. ed. New York: Farrar, Straus and Giroux.

Kim, Dan J., Donald L. Ferrin, and H. Raghav Rao. 2008. "A Trust-Based Consumer Decision-Making Model in Electronic Commerce: The Role of Trust, Perceived Risk, and their Antecedents." *Decision Support Systems* 44(2): 544–64. https://doi.org/10.1016/j.dss.2007.07.001.

Kunda, Ziva. 1990. "The Case for Motivated Reasoning." *Psychological Bulletin* 108(3): 480–98. https://doi.org/10.1037/0033-2909.108.3.480.

Liang, Kung-Yee, and en Scott L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73(1): 13–22. https://doi.org/10.1093/biomet/73.1.13.

Lipsky, Michael. 2010. *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services*. 30th anniversary expanded ed. New York: Russell Sage Foundation.

Lipton, Peter. 1990. "Contrastive Explanation." *Royal Institute of Philosophy Supplement* 27: 247–66. https://doi.org/10.1017/S1358246100005130.

Lyell, David, and Enrico Coiera. 2017. "Automation Bias and Verification Complexity: A Systematic Review." *Journal of the American Medical Informatics Association* 24(2): 423–31. https://doi.org/10.1093/jamia/ocw105.

Ma, Yan, Madhu Mazumdar, and Stavros G. Memtsoudis. 2012. "Beyond Repeated-Measures Analysis of Variance: Advanced Statistical Methods for the Analysis of Longitudinal Data in Anesthesia Research." *Regional Anesthesia and Pain Medicine* 37(1): 99–105. https://doi.org/10.1097/AAP.0b013e31823ebc74.

Mayer, Roger C., James H. Davis, and F. David Schoorman. 1995. "An Integrative Model of Organizational Trust." *The Academy of Management Review* 20(3): 709. https://doi.org/10.2307/258792.

Maynard-Moody, S., and M. Musheno. 2000. "State Agent or Citizen Agent: Two Narratives of Discretion." *Journal of Public Administration Research and Theory* 10(2): 329–58. https://doi.org/10.1093/oxfordjournals.jpart.a024272.

Maynard-Moody, Steven, and Michael Musheno. 2003. *Cops, Teachers, Counselors: Stories from the Front Lines of Public Service*. Ann Arbor, MI: University of Michigan Press. https://doi.org/10.3998/mpub.11924.

McKnight, D. Harrison, Michelle Carter, Jason Bennett Thatcher, and Paul F. Clay. 2011. "'Trust in a Specific Technology: An Investigation of its Components and Measures'. *ACM Transactions on*." *Management Information Systems* 2(2): 1–25. https://doi.org/10.1145/1985347.1985353.

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(5): 837–46. https://doi.org/10.1111/puar.13391.

Meijer, Albert, and Martijn Wessels. 2019. "Predictive Policing: Review of Benefits and Drawbacks." *International Journal of Public Administration* 42(12): 1031–9. https://doi.org/10.1080/01900692.2019.1575664.

Mendel, R., E. Traut-Mattausch, E. Jonas, S. Leucht, J. M. Kane, K. Maino, W. Kissling, and J. Hamann. 2011. "Confirmation Bias: Why Psychiatrists Stick to Wrong Preliminary Diagnoses." *Psychological Medicine* 41(12): 2651–9. https://doi.org/10.1017/S0033291711000808.

Mercado, Joseph E., Michael A. Rupp, Jessie Y. C. Chen, Michael J. Barnes, Daniel Barber, and Katelyn Procci. 2016. "Intelligent Agent Transparency in Human–Agent Teaming for Multi-UxV Management." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 58(3): 401–15. https://doi.org/10.1177/0018720815621206.

Miller, Tim. 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.

Mosier, Kathleen L., Linda J. Skitka, Susan Heers, and Mark Burdick. 1998. "Automation Bias: Decision Making and Performance in High-Tech Cockpits." *The International Journal of Aviation Psychology* 8(1): 47–63. https://doi.org/10.1207/s15327108ijap0801_3.

O'Neill, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishing Group.

Peeters, Rik. 2020. "The Agency of Algorithms: Understanding Human-Algorithm Interaction in Administrative Decision-Making." *Information Polity* 25: 507–22. https://doi.org/10.3233/IP-200253.

Politie. 2020 https://www.rijksoverheid.nl/documenten/jaarverslagen/2021/05/19/nationale-politie-2020. *Jaarverantwoording Politie*.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "'Why Should I Trust you?': Explaining the Recommendations of any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–44. San Francisco California USA: ACM. https://doi.org/10.1145/2939672.2939778.

Schiff, Daniel S., Kaylyn Jackson Schiff, and Patrick Pierson. 2022. "Assessing Public Value Failure in Government Adoption of Artificial Intelligence." *Public Administration* 100(3): 653–73. https://doi.org/10.1111/padm.12742.

Skitka, Linda J., Kathleen L. Mosier, and Mark Burdick. 1999. "Does Automation Bias Decision-Making?" *International Journal of Human-Computer Studies* 51(5): 991–1006. https://doi.org/10.1006/ijhc.1999.0252.

Snow, Thea. 2021. "From Satisficing to Artificing: The Evolution of Administrative Decision-Making in the Age of the Algorithm." *Data & Policy* 3: e3. https://doi.org/10.1017/dap.2020.25.

Simon, Herbert A. 1957. *Models of Man; Social and Rational*. New York: Wiley.

Taber, Charles S., and en Milton Lodge. 2006. "Motivated Skepticism in the Evaluation of Political Beliefs." *American Journal of Political Science* 50(3): 755–69. https://doi.org/10.1111/j.1540-5907.2006.00214.x.

Taylor, B. J. 2006. "Factorial Surveys: Using Vignettes to Study Professional Judgement." *British Journal of Social Work* 36(7): 1187–207.

Thagard, Paul. 1989. "Explanatory Coherence." *Behavioral and Brain Sciences* 12(3): 435–67. https://doi.org/10.1017/S0140525X00057046.

Tummers, Lars, and Victor Bekkers. 2014. "Policy Implementation, Street-Level Bureaucracy, and the Importance of Discretion." *Public Management Review* 16(4): 527–47. https://doi.org/10.1080/14719037.2013.841978.

van der Waa, Jasper, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. "Evaluating XAI: A Comparison of Rule-Based and Example-Based Explanations." *Artificial Intelligence* 291: 103404. https://doi.org/10.1016/j.artint.2020.103404.

Veale, Michael, and Reuben Binns. 2017. "Fairer Machine Learning in the Real World: Mitigating Discrimination without Collecting Sensitive Data." *Big Data & Society* 4(2): 205395171774353. https://doi.org/10.1177/2053951717743530.

Veale, Michael, and Irina Brass. 2019. "Administration by Algorithm?: Public Management Meets Public Sector Machine Learning." In *Algorithmic Regulation*, edited by Michael Veale and Irina Brass, 121–49. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198838494.003.0006.

Weller, Adrian. 2019. http://arxiv.org/abs/1708.01870. "Transparency: Motivations and Challenges". *arXiv:1708.01870 [cs]*.

Young, Matthew M., Justin B. Bullock, and Jesse D. 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): gvz014. https://doi.org/10.1093/ppmgov/gvz014.

Zerilli, John, Alistair Knott, James Maclaurin, and Colin Gavaghan. 2019. "Algorithmic Decision-Making and the Control Problem." *Minds and Machines* 29(4): 555–78. https://doi.org/10.1007/s11023-019-09513-7.

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: Springer International Publishing. https://doi.org/10.1007/978-3-030-19566-3_20.

Zuiderwijk, Anneke, Yu-Che Chen, and Fadi Salem. 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.

AUTHOR BIOGRAPHIES

Friso Selten is a PhD candidate at Leiden University, The Netherlands. His research focuses on the social and technical capabilities public organizations need to adopt AI technologies.
f.j.selten@fgga.leidenuniv.nl

Marcel Robeer is a PhD candidate at the Netherlands Police Lab AI. He has a joint appointment at the Netherlands National Police and Utrecht University, The Netherlands. His research focuses on technical aspects of explainable AI, and operationalizing AI transparency and ethics in law enforcement.
m.j.robeer@uu.nl

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

How to cite this article: Selten, Friso, Marcel Robeer, and Stephan Grimmelikhuijsen. 2023. "'Just like I Thought': Street-Level Bureaucrats Trust AI Recommendations if they Confirm Their Professional Judgment." *Public Administration Review* 83(2): 263–278. https://doi.org/10.1111/puar.13602

APPENDIX A

See Table A1.

TABLE A1 Balance tests

	Sex (% F)	Executive (% yes)	Age[a]	Education[b]	Knowledge AI[c]	Trust Technology[c]
Incongruent: Explained	16	94	3.66	3.10	2.84	5.10
Incongruent: Unexplained	45	92	4	3.12	2.96	5.03
Congruent: Explained	42	93	3.88	3.24	3.04	5.16
Congruent: Unexplained	36	90	3.62	3.23	3.19	5.60

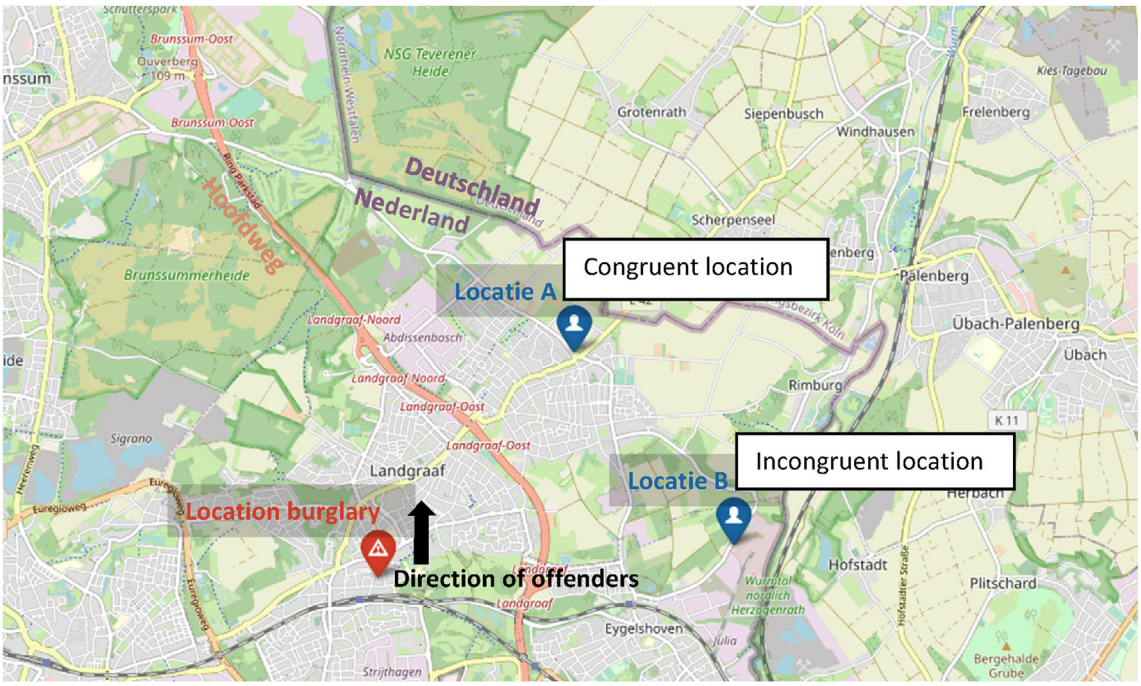
Note: Sex ($\chi^2(3) = 7.00$, $p = .06$); Executive ($\chi^2(3) = 0.70$, $p = .90$); Age ($F(3, 280) = 0.182$, $p = .14$); Education ($F(3, 280) = 0.81$, $p = .49$); Knowledge AI Education ($F(3, 280) = 0.53$, $p = .66$); Trust Technology; ($F(3, 280) = 2.89$, $p = .04$).
[a]Range 1–7: 1 = < 18, 7 = > 70.
[b]Range 1–5: 1 = Finished primary education, 5 = University degree.
[c]Range 1–7: 1 = completely disagree, 7 = completely agree.

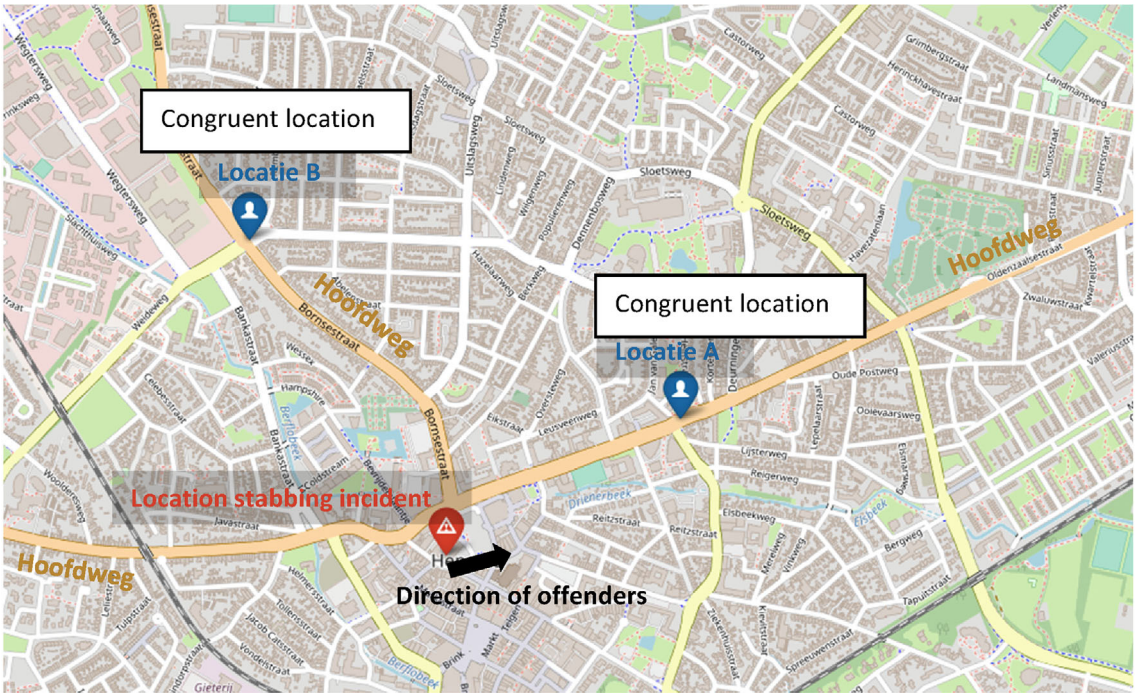
APPENDIX B: SUMMARY OF EXPERIMENTAL SCENARIOS (TRANSLATED FROM DUTCH)

See Table B1.

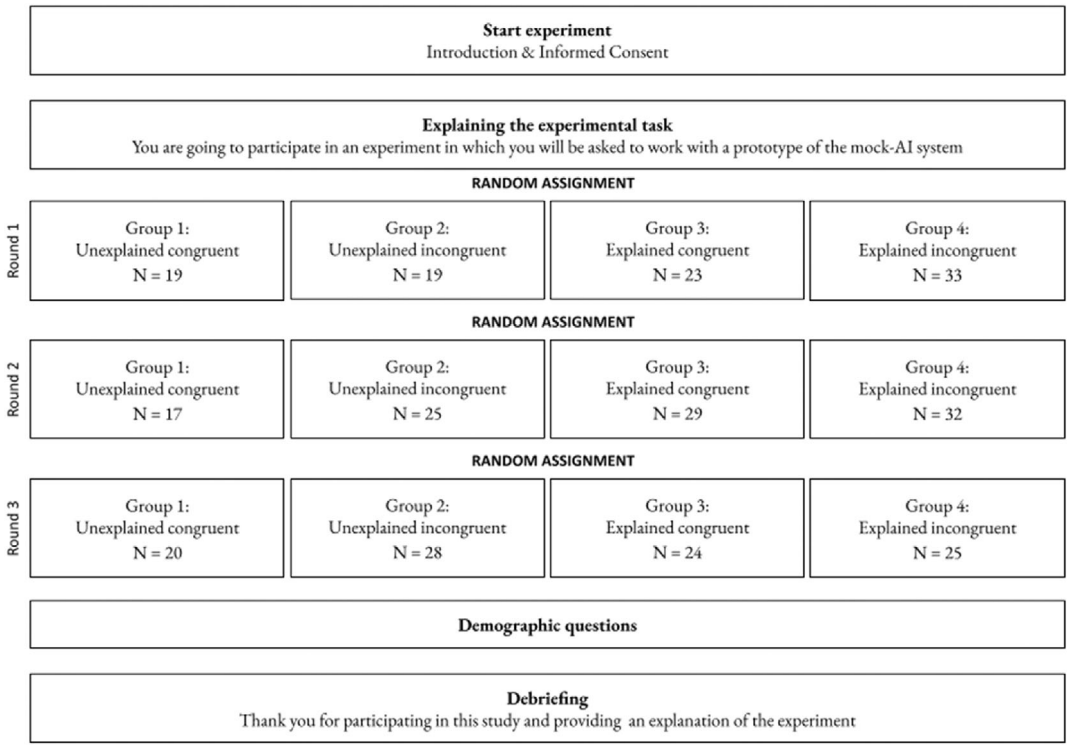
TABLE B1 The three experimental scenarios

Scenario	Description	Location rationale	Explanation presented
Burglary	Two offenders fled north on a scooter with a German license plate	Offenders of crimes near the border often flee abroad. The congruent location is alongside the fastest route to the German border.	Congruent location because this it is the quickest route to Germany Incongruent location because burglary suspects often flee via quiet roads
ATM robbery	Two offenders fled west in a fast car	Offenders of ATM-robberies often flee via the highway. The congruent location is at the nearby highway entrance.	Congruent location because suspects of robberies often flee via the highway Incongruent location because the suspects cannot then flee into the neighboring town via this route
Stabbing incident	The offender drove off east in a red car	The offender has fled in a car and was last seen heading east. The congruent location is situated east of the location where the crime has been committed crime	Congruent location because it is a central point along a major road in the direction of escape (east) from the suspect Incongruent location because it is a central location where many possible escape routes converge





APPENDIX C: EXPERIMENTAL SET-UP



APPENDIX D: DEVIATIONS FROM PRE-ANALYSIS PLAN

Three deviations between the pre-registered and the reported study have to be reported; we simplified the hypotheses, excluded one experimental group, and revised the statistical approach.

First, the revisions of the hypotheses aim to make the hypothesized relationship more clear but do not alter the proposed relationship:

- The first hypothesis in the pre-registry (*In general, police officers are likely to trust algorithmic advice*) does not propose a testable relationship but is a descriptive question which is answered in Table 3 in the article.
- H2 The second hypothesis in the pre-registry (*In general, police officers are more likely to trust algorithmic advice that complies with their tacit knowledge than algorithmic advice that conflicts with their tacit knowledge*) is rephrased in the article to H2 (*Street-level bureaucrats perceive AI recommendations that are congruent with their professional judgment as more trustworthy than AI recommendations that are incongruent with their professional judgment*).
- H3 The third hypothesis in the pre-registry (*In general, police officers are more likely to trust explained algorithmic advice than unexplained algorithmic advice*) is rephrased in the article to H2 (*Street-level bureaucrats perceive explained AI recommendations as more trustworthy than unexplained AI recommendations*).
- The fourth and fifth hypothesis in the pre-registry are about testing for an interaction between congruence and explanation manipulations. We registered these hypotheses based on initial research into the subject of XAI. However, when doing more in-depth research into XAI literature, we found little evidence to support hypothesizing the existence of an interaction. Previous studies predict a positive effect of XAI on trustworthiness of the AI recommendation or a negative effect—but we found no solid evidence to hypothesize an interaction with the congruency manipulation. We, therefore, decided that we could not hypothesize the existence of this interaction effect in the manuscript. In the results section, we do test for this interaction and find that it is not statistically significant.
- In the pre-registry, we were not clear about how we would relate trustworthiness and behavior. In the article, we added H3 to test this relationship: Higher perceived trustworthiness of AI recommendations by street-level bureaucrats is related to an increased likelihood of following the recommendation.

Secondly, one experimental group was excluded from the analyses because this group could not be used to test the hypotheses. In the experiment, some participants were assigned to a fifth experimental group in which they were asked to choose between location A and B without receiving an AI recommendation. This group could not be

used to test any of the pre-registered hypotheses and was, therefore, excluded from the final study.

Thirdly, the statistical approach was changed because we did not accurately account for the within-subject effects in our pre-registered statistical procedure. While we in the pre-registry did indicate that we wanted to combine the data of the different scenarios to increase the power of the experiment, we did not accurately account for the repeated measures design in the analysis plan. In the final manuscript, we have revised the statistical approach to account for the potential dependencies between observations.

APPENDIX E: ROBUSTNESS CHECKS

See Table E1 and E2.

TABLE E1 Confirmation of hypotheses 1 and 2 using a two-way ANOVA

	Df	Sum of Sq	Mean Sq	*F*-value	*p*-value
Explanation	1	1.4	1.44	1.03	.311
Congruency	1	10.3	10.35	7.39	.007
Explanation × Congruency	1	1.2	1.19	0.85	.357
Residuals	290	406.2	1.401		

TABLE E2 Confirmation of H3 using a logistic regression

	Est.	S.E.	*Z*-value	*p*-value
Intercept	−0.905	0.518	−1.748	.080
Trustworthiness	0.243	0.100	2.420	.016
	$\chi^2 = 6.001$, $p = .014$ AIC = 398.343, BIC = 405.710			