The Politics of Using AI in Policy Implementation Evidence from a Field Experiment[*]

Yotam Margalit[†]and Shir Raviv[‡]

March 2025

Draft: Comments Welcome

Abstract

The use of AI by government agencies in guiding important decisions (e.g., on policing, welfare, education) has generated backlash and led to calls for greater public input in AI regulation. But what exactly would such public input reflect? Does personal experience with the technology or learning about its implications shape people's views on using AI in government? We study these questions experimentally. We track the attitudes of over 1,500 workers, where the boss who allocates them to task (human vs. AI), the tasks' content and valence are all randomly assigned. Over a three-wave panel, we find that personal experience with AI-as-boss affected workers' performance, but not their policy attitudes. In contrast, exposure to information about the technology generated significant attitudinal change, even when it went against their experience or prior views. Our findings highlight the promise and potential challenges of involving public input in shaping AI governance.

[*] This study was pre-registered on 02/14/2023 (#121928).

[†] Department of Political Science, Tel-Aviv University and Department of Political Economy, King's College London.

[‡] Data Science Institute, Columbia University.

1 Introduction

In 2017, the city of Toronto introduced an ambitious plan to leverage AI and data-driven tools to create a "smart city" in part of its jurisdiction. The initiative, a collaboration with a subsidiary of Google's parent company, Alphabet Inc, promised to improve public services and promote urban development. By analyzing reams of data recorded from a network of sensors it sought to deploy, the aim was to use AI to optimize decisions on policy challenges ranging from efficient energy use to parking and waste disposal. Yet as the project advanced toward implementation, it faced strong opposition from a diverse coalition of stakeholders, including activists, academics, journalists, and residents, who raised concerns about the project's implications with regard to privacy, surveillance, and social justice. After three years of contentious public debate, the project was abandoned (Lorinc, 2022).[1]

Toronto's initiative is just one among a series of high-profile cases in which vocal public opposition hindered the implementation of AI-based initiatives in public policy. In the UK, for example, the Department of Education scrapped its use of an AI algorithm to predict and replace students' qualifying exam grades during the pandemic, in the face of fierce public criticism (Walsh, 2020). For similar reasons, the New Orleans Police Department gave up the use of algorithms to predict crime hot spots and to guide its allocation of policing units (Winston, 2018). Major tech companies, such as Amazon, Microsoft, and IBM, have all had to pull out of projects worth billions of dollars providing facial recognition technology to police departments across the country, in response to public outcry against questionable use of such systems (Heilweil, 2020). Notably, in some of these cases, the AI-based technology was deemed to offer a significant improvement over prior methods, but it was nonetheless withdrawn in response to public opposition. These backlashes can have far-reaching political

[1] The project was ostensibly dropped due to the economic implications of Covid, but as a series of in-depth accounts indicate, it would have been terminated even if the pandemic had not broken out (O'Kane, 2022).

repercussions, as demonstrated in the Netherlands, where a biased algorithm used to detect child benefit fraud wrongly accused thousands of families, sparking widespread public outcry and ultimately resulting in the resignation of the Dutch government in 2021 (Guardian, 2021).

Such cases raise concerns that rapid deployment of AI technology could undermine public trust and hinder the future adoption of innovative technologies, even if those prove beneficial (Evgeniou, Hardoon, and Ovchinnikov, n.d.).[2] These concerns may yet prove warranted. So far, however, not much is known about how the use of AI technology affects people's attitudes toward the issue, despite growing familiarity with it (e.g. ChatGPT). How do people view the use of AI-based algorithms in determining high-stakes decisions in public policy? How do these views evolve in response to personal experience with AI and to growing information about the technology's implications?

Answers to these questions are particularly pertinent given the widening use of AI algorithms in making decisions across an array of policy domains. From decisions regarding the allocation of food stamps and the granting of parole, to selection of tax audit targets and the deployment pattern of police patrols, many functions that were once performed solely by human officials are increasingly delegated to AI-based systems (e.g., Bansak et al., 2018; Toros and Flaming, 2018; Yeung, 2020). As this phenomenon expands, there is also a growing recognition among both government and business leaders of the need for the public's input to ensure that AI development is aligned with citizens' values and preferences (Mays et al., 2021; Management and Budget, 2020). For example, the Biden Administration recently put forth a "Blueprint for an AI Bill of Rights," stressing the importance of engaging the public on all stages of developing automated systems, especially before their implementation (White-House, 2022). Elsewhere, a recent study finds that U.S. state legislators view the

[2] A similar sentiment was recently expressed in a public letter signed by thousands of AI experts and industry leaders, including Elon Musk, who called for a pause on the development of AI systems that are more advanced than GPT4 (News, 2023).

public's input on ethical and social issues related to AI as crucial (Schiff and O'Shaughnessy, 2023).

Despite such calls for public input, it is unclear what the public's input about AI would reflect, since in other politically salient issues involving scientific knowledge and domain expertise (e.g., climate change or Covid vaccinations), partisanship and ideological leanings appear to shape much of the public debate. It is therefore not obvious that people are willing or able to form educated views about AI's potential benefits and risks. In this paper, we develop a theoretical and empirical account of the evolving public debate regarding the use of the technology in various policy domains. We focus on the way different levels of engagement with AI affect people's views, especially the influence of personal experience with the technology and exposure to information about its potential impact.

Of course, the challenge of addressing this question is that individuals' level of engagement with the technology is not random and people who choose to engage with the technology may differ substantially in their policy views. We therefore designed and conducted a field experiment in which we randomly assigned the exposure to AI-based decision making. Specifically, we hired more than 1,500 American workers to perform paid tasks on an online labor market platform, and then using a three-wave panel survey, we track their views on AI-based decision making in various policy domains.

The experiment consisted of a factorial design of three treatments. The first varied the decision-maker who hired and assigned workers to tasks: a computer algorithm or a human employer; the second treatment varied the nature of the experience (i.e., whether it was in line with or against the worker's preferences); the third factor varied the content of the tasks that the workers performed, exposing them to either positive, negative, or placebo information about AI and its implications.

Our analysis finds no evidence that personal exposure to the algorithm-as-boss had an impact on workers' support for AI policy. This result, which remains consistent across a wide array of tests, is particularly notable given that exposure to the algorithm's decisions did influence workers' behavior on the job (such as performance, time spent on the task, and willingness to work). However, our results indicate that AI-related attitudes are not solely determined by prior dispositions or beliefs. Rather, we find that workers significantly updated their attitudes after being exposed to information about AI and its societal implications, an effect that held days after exposure to the information. Notably, people were particularly prone to revise their views on the use of AI in making decisions regarding resource allocation, i.e., cases in which people had less clear preferences for a particular decision maker. Furthermore, we find that people update their views even when the information does not conform with their general predispositions. The results indicate that, at this stage of the public debate, attitudes are sufficiently malleable and can be influenced by exposure to relevant information.

By and large, the findings suggest that people make little connection between their personal interactions with AI decision-making systems and the broader question of the appropriate use of the technology in guiding public policy. The reasons for this disconnect require further research, but it appears that people think about this policy question more generally, and perhaps take into consideration the broader social impact they perceive the technology is offering.

Our findings contribute to the growing literature on the determinants of public opinion regarding the use of algorithmic decision systems (ADM) in public policy (Bansak and Paulson, 2023). Prior studies identified several factors associated with initial attitudes on this issue, such as trust in technology, personality traits, and social norms (e.g. Zhang, 2021; Schiff, Schiff, and Pierson, 2022). More recently, studies have shown that these attitudes depend on the specific design features of the technology (Kennedy, Waggoner, and Ward, 2022) and the context in which it is implemented (Horowitz, 2016; Wenzelburger and Achtziger, 2023; Raviv, 2023). Yet importantly, all prior work has focused on a snapshot of attitudes when individuals have limited knowledge or experience with AI to inform their judgments. This study adds to that work by systematically examining the evolution of people's attitudes in response to acquiring information about the technology or to experiencing AI firsthand.

Finally, the findings contribute to the growing literature on the political ramifications of the recent advancements in AI and digitization, focusing specifically on the way the current wave of automation in the labor market affects voters' preferences and behavior (e.g., Anelli, Colantone, and Stanig, 2019; Gallego et al., 2022; Kurer and Hausermann, 2022; Bicchi, Gallego, and Kuo, 2023; Schöll and Kurer, 2023). While this body of work examines the risk of workers being replaced by AI technology, we study the political implications of working under machine-guided decisions, an increasingly common experience in recent years that is largely unexplored in the extant literature.

2 Drivers of public opinion on the use of AI in Policy

How do views evolve as a result of more engagement with the technology and its implications? Answers to these questions are not obvious ex-ante. The literature on the public adoption of emerging technology debates the extent to which people can change their judgment about new technology.

One strand of research emphasizes a cognitive process of learning and holds that people's attitudes often evolve and change as they acquire more knowledge about new technology (e.g., Yeomans et al., 2019). For example, studies suggest technological literacy is key to the way people weigh the costs, risks, and benefits of energy technologies or biotechnology (Cobb and Macoubrie, 2004; Stoutenborough and Vedlitz, 2016).

This conjecture seems particularly relevant at this early stage of the public debate over AI regulation, when most people still know little about AI and there are no widely accepted elite positions that can cue public opinion on the matter. As various actors have a growing interest in informing the public about certain benefits or potential risks of AI, more people are likely to encounter new information about the technology and revise their views accordingly.[3]

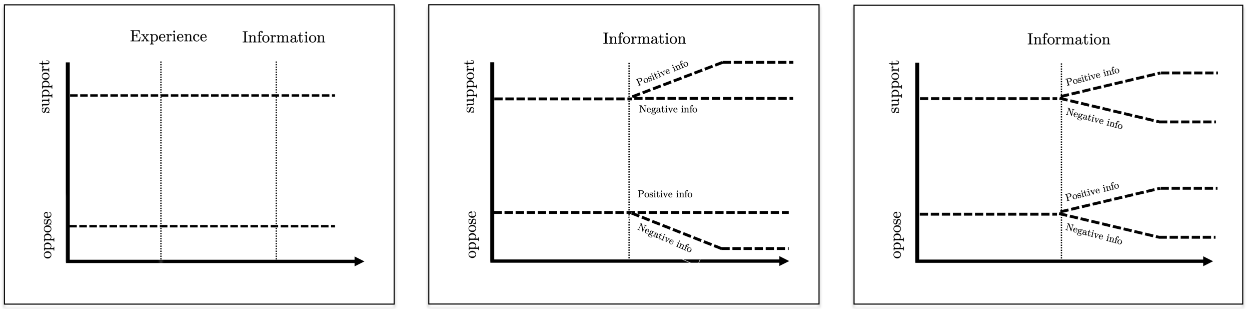
Another strand of research underscores the affective dimension, and contends that information alone is rarely sufficient to lead to attitude change. Instead, people need to have also a motivation to process the information (Scheufele and Lewenstein, 2005; Boudet, 2019). Specifically, if people cannot grasp how AI could affect their well-being, they may have little motivation to learn about the technology. Moreover, the complexity of the technology may render it difficult to comprehend and further limit the effectiveness of information on attitudes.

However, the fact that people are less informed about technological issues does not necessarily mean that they have only weak opinions on the matter (Lee, Scheufele, and Lewenstein, 2005). Studies have shown that individuals form opinions about new technologies based on predispositions, such as their general trust in technology (Araujo et al., 2020; Mays et al., 2021) or in human decision-makers (Miller and Keiser, 2021). These predispositions are often difficult to overcome and likely influence the extent to which people update their views in response to new information (e.g., Taber and Lodge, 2006). In other words, individuals are often motivated reasoners, and their response to new information largely depends on whether it is congruent with their prior beliefs (Druckman and Bolsen, 2011). If this is the case in the context of AI, biased processing of new evidence will likely cause preferences to change only slightly when the information contradicts prior views.

To illustrate the differences between these theoretical approaches, suppose two people watch a news segment of experts discussing the ProPublica report that revealed racial bias in COMPAS, the risk assessment algorithm discussed earlier. One person is initially more

[3] One such example is ProPublica's report on the risk assessment algorithm COMPAS used to assess the risk of recidivism for defendants in some US states. The report, which showed that the algorithm exhibited racial bias in predicting recidivism rates, sparked a heated public debate about the implications of AI usage in the criminal justice domain (Angwin et al., 2016).

Figure (1) Exposure to New Information - Trajectories of Preferences



(a) No updating (b) Motivated updating (c) Directional updating

Notes: Three possible patterns of attitudinal change that may result from exposure to new information about AI and its implications. The vertical axis indicates the probability of favoring AI algorithms in the policy implementation.

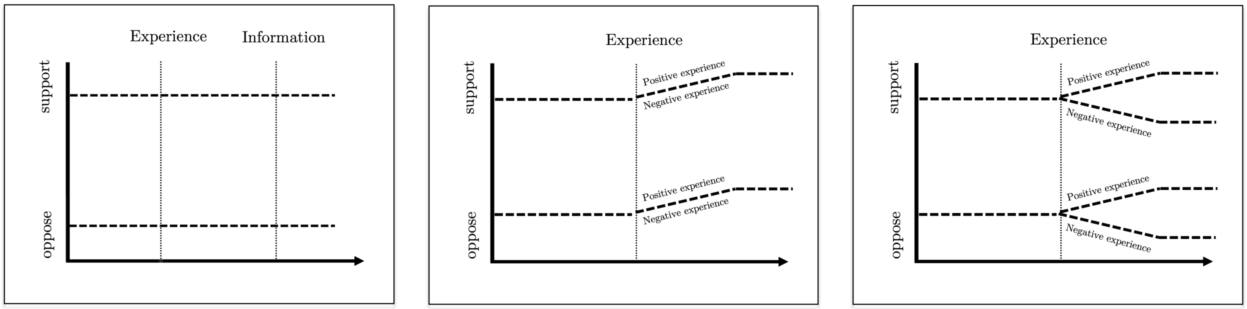
favorable towards AI, while the other is more skeptical. How does exposure to this new information affect their opinions? Figure 1 depicts three possible trajectories of attitude change that individuals may follow. Figure 1 (a) shows the attitude of both individuals remaining stable, irrespective of the information they encounter. In contrast, panel (b) shows support for the use of AI changing only in the direction of the individuals' prior beliefs; they are paying attention only to the evidence that confirms their priors while ignoring the rest. Finally, panel (c) shows support for AI change in the direction implied by the new information, irrespective of people's initial stance. For instance, learning about the biased outcomes of COMPAS algorithm would make them both more skeptical about using AI in public policy.

Having little motivation or ability to process information in an unbiased manner, a key shortcut individuals may rely on is their prior experience with the technology. In recent years, people are increasingly exposed to AI in their daily lives– from automated hiring decisions and loan approvals to insurance pricing and credit determinations. These interactions may affect how people think about AI. Specifically, direct experience with ADM may foster a sense of familiarity with and trust in AI algorithms, leading to greater acceptance of their use in public policy (Mahmud et al., 2022). This notion is often expressed by technology experts who argue that effective and accurate technologies will eventually overcome initial resistance and gain legitimacy among the public simply by people getting used to them (Haring et al., 2016; Ullman and Malle, 2017). For example, widespread exposure to ChatGPT and other large language models may enhance people's familiarity with AI, subsequently increasing their support for the use of AI in other domains. Indeed, earlier experimental studies have shown a similar response to interaction with other advanced technologies (e.g., semi-autonomous cars) (Lapinsky et al., 2008; Austin, Stevenson, and Wei-Skillern, 2006).

Alternatively, the nature of the experience plays a more prominent role in shaping opinions. Repeatedly receiving inaccurate information from ChatGPT or being denied a loan request by a bank's ADM creates a vivid heuristic that is more accessible than other sources of information. If this is the case, people's attitudes toward incorporating AI in policy could be a function of their satisfaction with the ADM they confront in their daily lives.

This expectation is consistent with research on economic voting, which suggests that less informed voters rely on their own economic experiences as heuristics to assess broader questions, such as the effectiveness of the government's economic policy and its competence (see Healy and Malhotra, 2013, for an extensive discussion). Furthermore, previous studies have shown that individuals' policy preferences are influenced by their personal experiences in an array of domains, be it in financial markets (Margalit and Shayo, 2021), the experience of extreme weather (Egan and Mullin, 2012) or in receiving government assistance (Anzia, Jares, and Malhotra, 2022). The implication of this argument is that the attitudinal impact of personal experience with AI should depend on the nature of the interaction with the algorithm: positive experiences will increase support for using AI in public policy, whereas negative experiences will have the opposite effect. Indeed, research on human-computer interaction shows that users of algorithmic systems tend to update their level of trust in algorithmic advice based on their prior interactions with these systems (e.g., Dietvorst, Simmons, and Massey, 2015).

Figure (2) Experiencing ADM - Trajectories of Preferences


(a) No Updating (b) Updating by exposure (c) Updating by experience

Notes: Three possible patterns of opinion change that could result from interacting directly with AI. The vertical axis indicates the probability of favoring AI in policy decisions.

While theoretically intuitive, we know little about the way personal experience with AI influences preferences toward the broader question of the desirability of employing AI in public policy. It is far from obvious that citizens generalize from their own encounters with algorithms to the broader question of using them in policy implementation contexts. Indeed, studies have suggested that personal experiences often remain "morselized" – disconnected from broader political contexts – unless explicitly linked through media coverage or elite discourse (Mutz, 1994). This is particularly relevant at this early stage of public debate over AI governance, where questions related to the governance and regulation of AI are not yet politicized. Without partisan cues, even consequential encounters with ADM may have limited impact on attitudes toward AI in governance.

Returning to the example of the two individuals who have different initial opinions on the use of AI in public policy. How would a personal experience with ADM affect their views? Figure 2 illustrates possible paths of attitude change. The left panel suggests that the two individuals' views remain unchanged – they view the encounter as irrelevant to the broader policy question. The middle panel indicates that their views change in a more positive direction irrespective of the encounter: simply by engaging with the technology, they develop more confidence and trust in its use in a policy setting. Finally, the right panel of the Figure implies that the two individuals' views change in accordance with the nature of the encounter. For example, if they apply for a job and an algorithm is responsible for determining their eligibility, being found suitable for the job (i.e., having a positive experience) would increase their support for AI use in policy, while being rejected (i.e., a negative experience) would decrease their level of support.

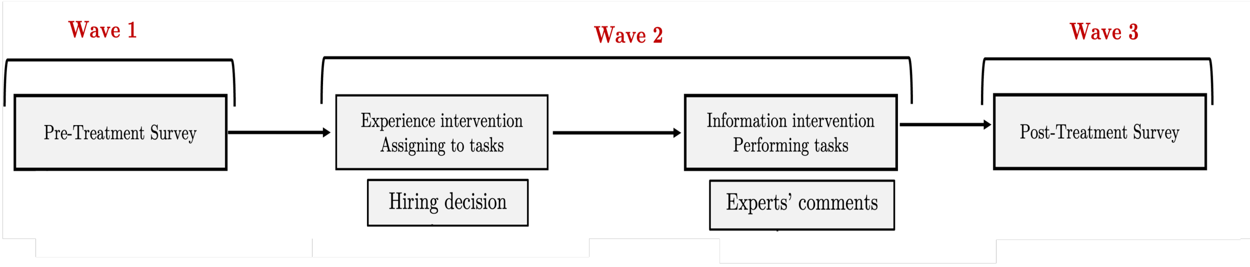
Taken together, the literature is quite ambiguous about the likely attitudinal impact that personal experience with AI and information about the technology are likely to have on our question of interest. One can find arguments why these forces would have a significant impact or none at all. In what follows, we describe an experimental approach designed to provide empirical insight regarding the impact of these different potential sources of influence.

3 Experimental Design

Our experiment focuses on the labor market which represents one of the main domains in which algorithmic decision-making is increasingly deployed (e.g., employee recruitment, task allocation, or quality assessment), making it a firsthand experience that a growing share of the population encounters. We therefore chose as our experimental setting Amazon's Mechanical Turk (MTurk) platform – the world's largest online labor market. MTurk provides employers with access to a large base of potential employees to perform a range of discrete on-demand tasks. Notably, prior research has validated MTurk as a useful and reliable setting for assessing key labor market outcomes (e.g., Burbano, 2016; McConnell et al., 2018). Furthermore, evidence suggests that findings from MTurk are comparable to those from more traditional (offline) employment settings (Horton, Rand, and Zeckhauser, 2011).

We invited 2375 American workers to perform paid tasks, and tracked their views on the use of AI in various policy domains using a three-wave panel survey, one of which was administered before the workers completed the task, and the other two were fielded after the task's completion. To evaluate the impact of personal experience with ADM on

Figure (3) Experimental Design



support for AI in public policy, the main intervention varied the decision-maker who hires and assigns workers to tasks: a computer algorithm or a member of the HR team. To assess the importance of a positive versus a negative experience with ADM, the second intervention veried whether workers were selected (by the decision-maker) to their preferred task or not. Finally, to assess the impact of exposure to new information on the updating of attitudes, the third intervention varied the *content* of the tasks that the workers performed. Specifically, we varied whether the task entailed exposure to information about positive implications of AI, negative implications of AI, or to placebo information about the fashion industry.

3.1 Sequence of the Experiment

The basic sequence of the experiment is presented in Figure SI-1. In this section, we describe in detail the rationale for, and procedure of, each stage (see Appendix A for a detailed graphical representation of the experiment's sequence).

3.1.1 Pre-Treatment Survey

In February 2023, we asked Mturk workers to complete a survey on social issues for a payment of $0.80. The survey included several pre-treatment outcomes that asked respondents how much they support or oppose using a predictive algorithm instead of a human to make determinations in various policy contexts. To minimize the possibility of demand effects, we added a host of unrelated items with the aim of blurring the focus of the study. We also collected information on pre-treatment covariates, such as age, race, ideology, education, technological literacy, and trust in institutions. For the survey questionnaire, see Appendix 3.

At the end of the survey, participants received an invitation to continue work with the same employer on an additional project involving one of two possible 8-minute tasks: (1) cataloging short texts according to their content for $1.00; or (2) rating comments by their tone for $3.00, a task that we described as "particularly suitable for people who are competent and good at seeing the bigger picture." We intentionally designed the descriptions of the tasks and the proposed wages so as to provide both material and psychological incentives for participants to have clear preferences for the latter task over the former.[4]

And since the vast majority of participants preferred the more lucrative option, we were able to clearly distinguish between participants in terms of whether they had a positive or negative experience with the employer's decision regarding the allocation of work.[5]

Figure SI-3 shows a screen capture of the invitation. We designed a distinct interface with the Analytics logo for both Waves 1 and 2. Our intention was to enhance the sense of an employer-worker setting and to help differentiate the first two waves from the third wave, which used a different user interface and requester. This way, we reduced the likelihood of participants connecting between the surveys and being strategic in their answers.

3.1.2 Experience with ADM

Three days after the initial survey, participants received an invitation in their personal MTurk inbox containing a link to the task to which they were randomly assigned. Importantly, all participants received the exact same generic invitation, informing them that they would

[4] To account for cases where participants had no meaningful preference between the tasks, we offered "don't care" option. Indifferent participants were then forced to choose between the tasks. We used this indication as a control in our analysis.

[5] As we pre-registered, we excluded the small number of participants who requested the lower-status task.

either be cataloging short texts or rating comments and that those who were found most suitable for the rating task would receive a bonus of $2 beyond the $1 base-rate. This means that only participants who clicked on the link to participate in the second wave actually received the treatment: information about their assigned task (either the desired rating task or the undesired cataloging task) and the decision maker who assigned them to the task (a computer algorithm or a member of the HR team). By tracking the clicks on the invitation link, we could monitor potentially nonrandom attrition; we elaborate on this point below.

After participants chose to proceed, they were told that not all workers are equally suitable for the higher-paid rating task, as it requires being good at "seeing the bigger picture", and then informed about their assignment. Workers randomly assigned to the negative experience treatment were told that the decision-maker evaluated their performance in the previous task and deemed them unsuited for the rating task.[6] In the positive experience condition, participants were informed that the decision-maker evaluated their performance in the previous task and had found them suitable for the rating task, as they had requested.[7]

However, and this is key, irrespective of the tasks' label and unbeknownst to the participants, the eventual task they were assigned to carry out was exactly the same one. We wrote the description of the two tasks in a way that ultimately described well the actual work the participants were asked to perform in both cases.

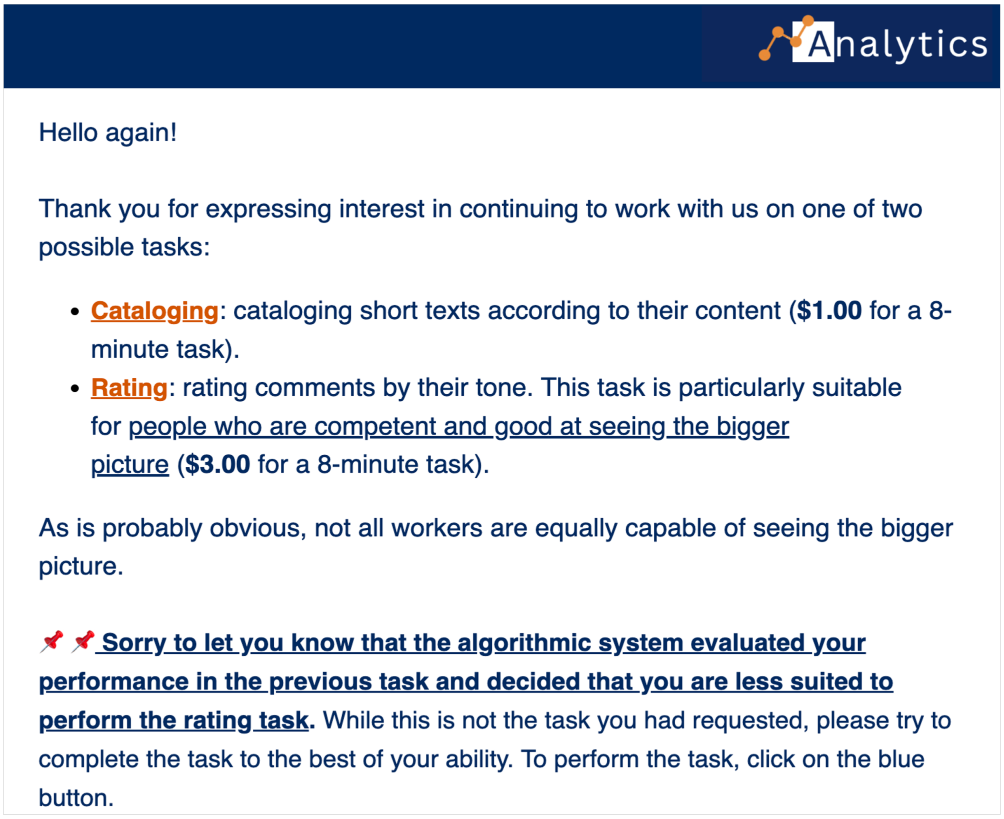
To assess the effects of working under human or an algorithm, the message to the workers explicitly mentioned the identity of the decision-maker (DM). In the human DM treatment, respondents were informed in multiple instances that the task assignment was decided by a member of the team.[8] Additionally, we included in the pre-treatment survey questions

[6] The 'previous task' in the experience treatment refers to a Rorschach test-like exercise included in the pre-treatment survey.

[7] All participants were debriefed about the experiment after the third survey wave. The exact wording of the debrief letter is provided in Appendix E, along with IRB approval.

[8] To avoid gender-based bias we alternated across respondents the member's name between "Danielle" and "Daniel".

Figure (4) Screen capture of the reply message: a negative experience with ADM



Hello again!

Thank you for expressing interest in continuing to work with us on one of two possible tasks:

- **Cataloging**: cataloging short texts according to their content (**$1.00** for a 8-minute task).
- **Rating**: rating comments by their tone. This task is particularly suitable for people who are competent and good at seeing the bigger picture (**$3.00** for a 8-minute task).

As is probably obvious, not all workers are equally capable of seeing the bigger picture.

📌 📌 **Sorry to let you know that the algorithmic system evaluated your performance in the previous task and decided that you are less suited to perform the rating task.** While this is not the task you had requested, please try to complete the task to the best of your ability. To perform the task, click on the blue button.

regarding a Rorschach image, to provide additional material on which the DM's evaluation of the participant's suitability for a task that requires "big picture thinking" could ostensibly be based.[9]

To drive home the type of experience—positive vs. negative—we asked participants to rate their satisfaction with the task to which they were assigned. This also served as a manipulation check, confirming that participants with the negative experience (i.e., assigned to their less preferred task) were indeed less satisfied with the decision, while those with positive experiences were more content.

Finally, participants had the opportunity to provide feedback on the decision made by either the human or the ADM, allowing the participants to express dissatisfaction with the

[9] To help ensure that participants read the message and received the treatment; the survey was programmed to allow participants to proceed to the next page only after 15 seconds.

decision. 36% of the participants opted to share their feelings. Unsurprisingly, most of them (76%) had a negative experience and expressed disappointment or frustration at being denied the higher-paying option.[10]

The workers' comments reveal that they were aware of who assigned them to the task, reassuringly confirming that the decision-maker treatment was noticeable. For instance, 44% of workers assigned to a task by an algorithm specifically mentioned the algorithm when making their appeal. They wrote, for example, "An algorithm doesn't know me personally and can't determine how I will perform." Others wrote: "Algorithms have bugs sometimes. It's not my fault." and "I don't believe the algorithm. I am very good at seeing the big picture." Similarly, workers in the human condition explicitly mentioned the name of the team member who assigned them to the task: "How did Danielle reach that decision?"; "Danielle has no idea who I am or what I can do." Similarly, "What did Daniel base his decision on?" or "Daniel is clueless about me."

The randomized assignment of the participants into treatments was used to generate groups that have similar characteristics on average. To further increase comparability across treatments, we used block randomization and grouped the sample according to their perceptions of suitability for performing the rating task based on their answers to the question in the pre-treatment survey. All these sampling decisions followed the preregistered design.

3.1.3 Exposure to Information about AI

Next, to assess the impact of new information on attitudes, we randomly manipulated the content of the tasks that participants performed. Specifically, they were asked to read eight expert comments and place them on a scale ranging from very negative to very positive. The treatment group received comments about the potential impact of AI, while the control

[10] In their answers, one participant wrote, "I feel that I put 100% effort into all these HITs; I should at least be given a chance." Another noted that "I always look at the big picture and feel like I would've done a great job as compared to other candidates on this platform."

group received comments about future fashion trends.

By integrating the information within the task itself, our aim was to increase participant engagement with the substance of the information. To further enhance this engagement, participants were also asked at the end of the task to indicate which comment was most persuasive and to explain in their own words why.

To examine whether people update their views in response to new information or instead rely on information that aligns with their prior dispositions, we also randomly manipulated the valence of the comments into either positive or negative tones. The comments were based on a Pew Research survey that asked over 900 experts in 2018 about AI and its consequences for human society (Anderson, Rainie, and Luchsinger, 2018). A negative comment about AI, for example, noted that: "AI may purposely exclude all references to race and ethnicity, but these systems still consider factors that correlate with race, such as low-income neighborhoods or employment history. As a result, their outputs can be racially discriminatory." In contrast, treatment with a positive tone included comments such as, "AI might lead to more consistent judgments than those made by humans, who may be influenced by emotional considerations or by fatigue." See the Appendix for detailed instructions of the task, the wording of the comments, and a screen capture of the user interface.

Of the eight comments each participant was asked to rate, seven had a positive (or negative) tone, depending on the treatment assignment, while one additional comment had the opposite tone. The inclusion of this contrasting comment was done to allow us to assess participants' engagement with the task by identifying potential errors in the classification of the comments.

In the final stage of the study, we conducted a follow-up survey that took place 4-7 days after carrying out the task (and 7-10 days after the original survey). To minimize potential Hawthorne effects, participants were invited by a *different employer* (requester) to complete a seemingly unrelated survey. This third wave did not include any details or information that indicated that the survey was connected to the cataloging/rating task that the workers had performed.

4 Data and Measures

4.1 Sample

Among the participants who were invited to perform further tasks, completed the post treatment survey. We did not invite to the study any of the workers classified among the most active workers on MTurk (accounting for about a fifth of the daily tasks on the platform). Our concern was that this group may possess an overly familiar understanding of AI technology, potentially skewing the study's conclusions. In addition, we stratified our sample based on two related criteria: 1) their experience on the platform, i.e., the number of prior tasks (HITs) completed and approved by the requester, and 2) the level of recent activity on the platform.

Table SI-2 presents descriptive statistics on pretreatment demographic and attitudinal variables, including all outcome variables used in subsequent analyses. As the table shows, the level of technological literacy varies substantially across the sample. Only about a quarter of the participants had a high degree of technological literacy, as measured using a principle component of four questions asking about familiarity with technology-related items. Notably, only 16% of participants were familiar with ChatGPT. Moreover, as we will show below, participants' initial views about using AI in public policy decisions closely mirror those found in nationally representative surveys using similar questions, increasing our confidence that our findings capture broader patterns rather than sensitive peculiarities of online workers.

4.2 Attrition

Table SI-1 reports attrition and completion rates for waves 2 and 3 by treatment assignment. As expected, the table shows significant differences in the completion rates of the post-treatment survey based on the type of experience but not by the identity of the decision maker. Participants assigned to the less attractive task had a slightly higher dropout rate compared to those with a positive experience ($p = 0.012$). When examining the differential attrition across waves 2 and 3, we found no significant differences between the groups (p>0.05). Among participants who completed wave 2, 83-82% also completed wave 3, a high rate compared to previous research that utilized in MTurk panels (Christenson and Glick, 2013).

4.3 Outcome Variables

Our primary dependent variable examined individuals' attitudes toward reliance on ADM in the implementation of public policy. Specifically, we asked respondents to indicate their support or opposition to using predictive algorithms instead of human decision-makers in a set of policy areas. The decisions covered a range of issues, including determination regarding the location of police patrols, the granting of parole to defendants, allocation of food stamps, where to place street lighting, approval of immigrant visa applications, increasing enforcement of illegal construction, and construction of homeless shelters. Decisions were chosen based on two relevant theoretical dimensions: the objective of the decision (assistance or sanctioning) and the population directly affected by the decision (individuals or collectives) (Raviv, 2023). In focusing on a set of policy domains, our aim was to ensure that the results are not sensitive to a specific decision context and that the attitudes we capture are generalizable across different types of decisions.

Using the questions about decisions in those different domains, we constructed an index based on a factor analysis score comprising eight items asked in wave 3.[11] By utilizing multiple items, we minimize measurement error. This approach addresses the issue of single-item measures potentially exhibiting low correlations between survey waves, even when the underlying attitude remains stable (Ansolabehere, Rodden, and Snyder, 2008; Broockman, Kalla, and Sekhon, 2017).[12]

5 Results

5.1 Attitudes toward AI in public policy

We begin by analyzing baseline preferences for using AI in public policy decision-making. Figure 5 presents the preference distribution for each policy decision. The results indicate that people are generally opposed to relying on AI algorithms in making the decisions. Consistent with prior research, we find that workers are particularly apprehensive about such use of the technology in decisions that involve sanctioning (Raviv, 2023). In cases where AI is used to assist, and particularly when required to make inferences regarding collectives (and not individuals), the public appears more open to the use of the technology, albeit still with a small proportion expressing strong support.

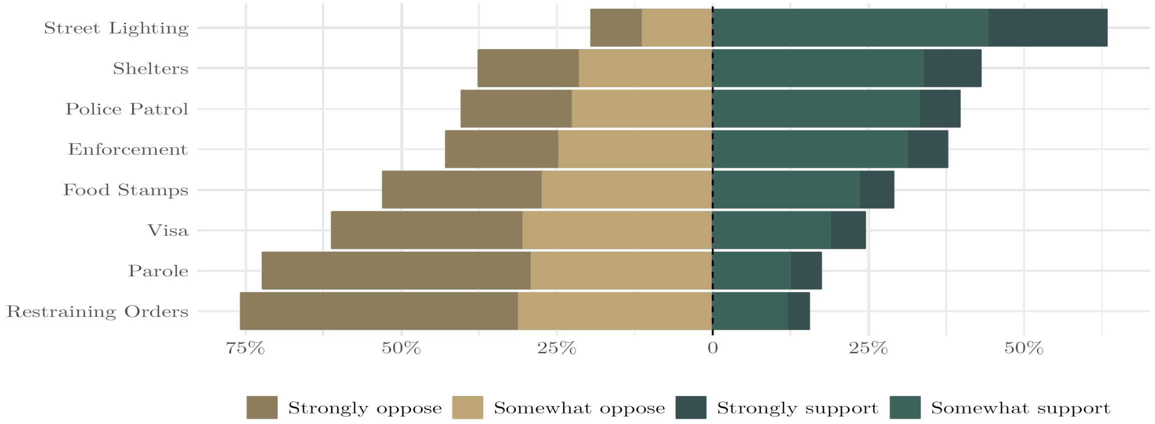
5.2 Effects of Experience on Attitudes

We begin with the causal effect of personal experience on attitudes and estimate the average treatment effect of exposure to ADM on attitudes, as measured by the post-treatment survey

[11] Our results remain similar when using instead PCA (See Appendix C).

[12] We divided the items into two separate matrices. The first matrix contains the same decisions asked in Wave 2, while the second matrix includes the remaining four decisions also asked in Wave 3 (e.g., issuing restraining order, approving immigrant visa; locating police enforcement; building shelters) By organizing the questions in this manner, we strive to eliminate potential bias in participant evaluations and ensure consistency in the outcomes measured across waves. Furthermore, to verify that participants were attentive and carefully evaluated the decisions, we incorporated an attention check within the second matrix.

Figure (5) Attitudes toward AI in public policy, pre-treatment



We measured the responses on a seven-point scale and then classified them into five categories: Strongly Oppose (1), Oppose (2-3), Indifferent (4), Support (5-6), and Strongly Support (7). The distribution calculation takes into account the indifferent category. Figure SI-2 shows the full distribution.

conducted several days after the assignment. Table 1 presents the results of linear regression models, all of which control for the pre-treatment outcome. To increase precision, some of the models also include a set of preregistered covariates (demographic and attitudinal), as measured in the pre-treatment survey.

In columns 1-3, we estimate the attitudinal impact of the employer's identity while controlling for the nature of the interaction with the employer, i.e., whether positive or negative. As pre-registered, and to ensure a clean comparison between treatment groups, column 1 includes in the sample only workers who received the placebo information, meaning that they were not exposed in the task to information about the merits or demerits of AI technology. To enhance statistical power, columns 2-3 report results for the full sample, controlling for respondents' informational treatment.

The table clearly shows that personal exposure to ADM did not affect workers' attitudes toward the use of AI in public policy. Across all specifications, the coefficient of ADM is consistently very small and below statistical significance, ranging from 0.001 (t=0.065) to 0.010 (t=1.351).

In light of these findings, a possible conjecture could be that changing attitudes is not a

Table (1) Effects of Experience on Attitudes

	DV:								
	Factor Analysis Score - Wave 3								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	(ITT)	(ITT)	(ITT)	(TOT)	(ITT)	(ITT)	(ITT)	(TOT)	(TOT)
Algorithmic DM	0.0003	0.009	0.010	0.017	−0.002	0.009	0.011	0.021	0.025
	(0.011)	(0.008)	(0.008)	(0.013)	(0.015)	(0.011)	(0.011)	(0.018)	(0.018)
Algorithmic X Negative Exp					0.005	0.0005	−0.002	−0.008	−0.013
					(0.021)	(0.016)	(0.015)	(0.027)	(0.027)
Negative experience	−0.003	0.006	0.007	0.006	−0.005	0.006	0.008	0.011	0.014
	(0.011)	(0.008)	(0.008)	(0.008)	(0.015)	(0.011)	(0.011)	(0.016)	(0.016)
Pretreatment outcome	0.798**	0.749**	0.711**	0.762**	0.798**	0.749**	0.711**	0.763**	0.723**
	(0.024)	(0.017)	(0.019)	(0.017)	(0.024)	(0.017)	(0.019)	(0.017)	(0.019)
Constant	0.095**	0.118**	0.135**	0.108**	0.096**	0.118**	0.135**	0.106**	0.141**
	(0.014)	(0.011)	(0.022)	(0.012)	(0.015)	(0.012)	(0.022)	(0.014)	(0.023)
Observations	760	1,500	1,497	1,433	760	1,500	1,497	1,433	1,430
R^2	0.603	0.573	0.584	0.584	0.603	0.573	0.584	0.583	0.594
Demographic Controls	No	No	Yes	No	No	No	Yes	No	Yes
Sample	Fashion	Full	Full	Full	Fashion	Full	Full	Full	Full
F-test (first stage)	–	–	–	201 ***	–	–	–	247 ***	77***

Notes: Linear regression models with standard errors in parentheses. The DV is the FA score of 8 items in Wave 3. The independent variables are indicators for the treatments: ADM, negative experience, and their interaction (in columns 5-8). Columns 1-3 and 5-7 show ITT estimates. Columns 4 and 8-9 show TOT estimates, using treatment assignment as an instrument for compliance: those who indicated in the manipulation checks that: (1) they completed the high-status task; and (2) the decision-maker who assigned them to the task was the requester's algorithm. Columns 1 and 5 limit the sample to placebo information, while others control for information treatments. All models control for the pretreatment outcome. Pretreatment covariates include gender, age, race, education, ideology, trust in technology, MTurk HIT record, attentiveness, self-reported suitability for the cataloging task, and indifference between tasks. Table SI-5 reports the full results. *p<0.05; **p<0.01

function of exposure to algorithms per se, but rather of the nature of exposure (i.e., positive or negative). We test this conjecture by adding to the models estimated in columns 5-7 an interaction term between the decision-maker treatment and the type of experience (positive or negative). The interactions yield a null effect among participants who had a negative experience with algorithms, namely where the algorithmic employer deemed them unfit to perform the high-status task.

Controlling for demographic characteristics, such as age, gender, education, and race, and other pre-treatment covariates, such as technological literacy and trust, does not alter these results. Furthermore, the null result holds when we examine the post-treatment outcomes collected in wave 2, right after completing the task. In sum, then, neither a positive nor a negative experience with the ADM altered subjects' attitudes, not even in the immediate term (See Appendix C.1 for additional results).

One might question whether these null results reflect the impact of real-life experiences with AI systems or whether, instead, the experimental treatment was not sufficiently strong and hence not noticed by the participants. To address this concern, we measure compliance with the treatments using two manipulation checks asked at the end of the post-treatment survey. The manipulation checks successfully distinguish between workers by their assigned decision maker (DM), as over 73% of the workers in the algorithmic DM condition reported that it was the specific algorithm used by the requester that assigned them to the task, compared to only 10% in the human DM condition ($p < .001$). 79% of the workers assigned to the human DM condition correctly identified the team member as the decision-maker, compared to only 5% in the algorithmic DM condition ($p < .001$).

One possibility is that this group of participants who complied with the treatment and correctly identified the decision-maker was a self-selected group (e.g., more attentive to the study or with less experience performing MTurk tasks). These characteristics may also have influenced their answers to the outcome questions. Hence, we cannot simply compare the treatment groups as they were randomly assigned. To address this issue, we estimate treatment-on-the-treated (TOT) effects with an instrumental variable (IV) regression, using the random assignment as an instrument for compliance. The results of the second stage and F statistics from the first stage are reported in columns 4 and 7-8.

Again, the results indicate that workers who interacted with an AI algorithm as their employer did not significantly differ from other workers in their support for employing AI in public policy. The estimated effect on the treated is, as expected, larger compared to the effect on all participants assigned to the treatment, but it is well below statistical significance.

Yet before accepting this interpretation, one must question whether the treatment, even if it was noticed by the participants, was simply too weak or inconsequential to have any meaningful impact. We assess this possibility by examining the effect of the treatment on various behavioral outcomes that are perhaps less prone to change than attitudes, such as level of performance and work commitment. If those behavioral outcomes had changed, this would indicate that the treatment was in fact effective, but not in changing subjects' views on the desired role of AI in policy implementation decisions.

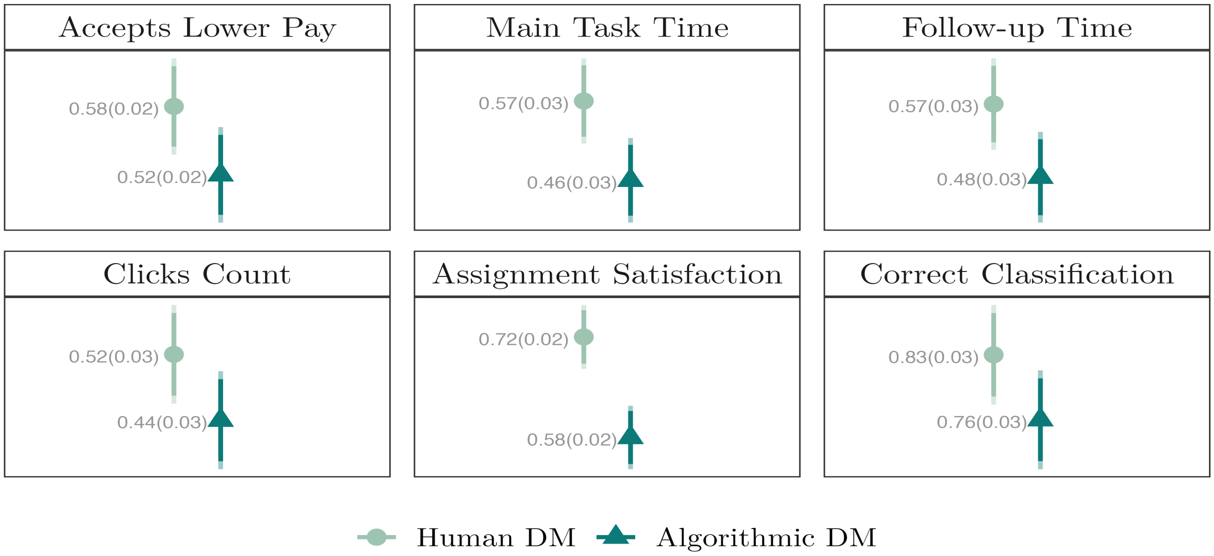
We focus on several indicators that measure performance and effort: accuracy in classifying comments with the opposite tone; time spent on the main task and the follow-up task, and thoroughness in carrying out the task, measured by the number of clicks. In addition, we asked workers to suggest a wage for completing an additional task of similar scope and length. If a worker suggests a wage lower than the amount received for the current task, we use this as an indication of high willingness to continue working with the employer. Finally, we measure job satisfaction using an item that asks workers to rate their satisfaction with their task assignment. See C.3 for a detailed description of the measures.

We re-estimate the main analysis but use these behavioral measures instead of attitudinal outcomes. Results are reported in table SI-7. Figure 6 shows the predicted values using this regression model.[13]

Our analysis reveals that workers' personal experiences with ADM in the workplace had a significant impact on a range of behavioral outcomes. For instance, workers who were assigned the task by an algorithm rather than a human were less satisfied with their assignment (p<0.001), put less effort in performing the task (p<0.05), and were significantly less likely to correctly classify the comments (p<0.1).

[13] To make the interpretation easier, we converted all outcomes to indicator variables. As the table shows, the results hold when using continuous measures.

Figure (6) Effects of Experience on Behaviors



The figure shows the predicted score of each behavioral outcome based on ITT analyses that regress them on a binary indicator for ADM, an indicator for the type of experience with the decision maker, and their interaction. Models also control for informational treatments. The thin (90%) and thick (95%) error bars represent the confidence interval around the estimate, respectively. The estimate and SE are reported as well. The full results are reported in Table SI-7.

Taken together, the results indicate that the null effects of personal experience with ADM on attitudes are not due to a weakness of the treatment. Rather, the treatment assignment appears to have been strong enough to affect behavior but not attitudes on our policy question of interest.

What do these results imply for the potential trajectory of preferences toward AI? One possibility is that preferences for using AI in public policy are based on strong predispositions about technology in general, in which case people are unlikely to change their views. Alternatively, it could be that attitudes are less sensitive to personal experiences with the technology because individuals do not link these types of daily interactions with AI and the broader question of the appropriate use of this technology in public policy decisions. In the next section, we further delve into this question by focusing on the attitudinal impact of exposure to information about AI.

5.3 Effects of Information on Attitudes

Next, we examine to what extent people update their views about the use of AI in public policy decisions in response to learning more about the technology. Our experimental design allows us to explore this question by randomly exposing workers to different types of relevant information.

In Table 2, we report results of estimating the effects of exposure to the different types of information (AI vs. fashion, positive vs. negative) as measured several days after encountering it. As pre-registered, columns 1-3 show the results on a subset of the sample, which includes only participants who were assigned to the human decision-maker.[14] Columns 4-6 include the full sample, controlling for the decision-maker.

The results show that exposure to positive information about the implications of AI has a significant effect on workers' attitudes towards the use of this technology (p<0.01). Specifically, when asked about their views in an unrelated survey several days after exposure to the information, workers who were randomly exposed to positive information about AI, as opposed to information about fashion, moved 0.043 to 0.045 points along the standardized scale toward supporting the use of the technology in public policy decision-making.

To put this effect size in context, Figure 7 plots the estimated effect of the information treatments, adjusting for key socio-demographic factors identified in the literature as determinants of attitudes toward AI.[15] Notably, the figure shows that the treatment effect of negative information on AI, for example, is larger than the differences observed between workers with higher and lower levels of education (D=-0.06, se= 0.01) The full results are reported in Table SI-9.

One concern might be that the valence of the informational content itself, regardless

[14] This is the "cleanest" comparison, as it is not contaminated by variation in experience with AI.

[15] We are particularly interested in comparing the treatment effect of information to other factors influencing attitudes toward the use of AI in public policy. Therefore, the analysis does not control for pre-treatment outcomes.

Table (2) Effects of Information on Attitudes

	DV:					
	Factor Analysis Score - Wave 3					
	Human DM Only			Full Sample		
	(1)	(2)	(3)	(4)	(5)	(6)
AI X Positive info	0.079^{***}	0.083^{***}	0.082^{***}	0.049^{**}	0.050^{**}	0.050^{**}
	(0.023)	(0.023)	(0.023)	(0.015)	(0.015)	(0.015)
Info about AI (ref: Fashion)	-0.036^{*}	-0.039^{*}	-0.037^{*}	-0.012	-0.013	-0.012
	(0.016)	(0.016)	(0.016)	(0.011)	(0.011)	(0.011)
Positive info (ref: Negative)	-0.019	-0.020	-0.020	-0.004	-0.003	-0.003
	(0.016)	(0.016)	(0.016)	(0.011)	(0.011)	(0.011)
Pre-dispositions (wave 1)	0.736^{***}	0.714^{***}	0.711^{***}	0.749^{***}	0.713^{***}	0.711^{***}
	(0.024)	(0.027)	(0.027)	(0.017)	(0.018)	(0.018)
Constant	0.125^{***}	0.151^{***}	0.170^{***}	0.108^{***}	0.111^{***}	0.125^{***}
	(0.016)	(0.026)	(0.032)	(0.012)	(0.019)	(0.023)
Model	Minimal	Socio-demog	Mturk HITs	Minimal	Socio-demog	Mturk HITs
Observations	741	741	741	1,500	1,500	1,497
R^2	0.561	0.573	0.575	0.577	0.586	0.584

Notes: The DV is the factor analysis score of 8 items asked in Wave 3. The independent variables are indicators for the treatments: information on AI (fashion as reference), positive tone (negative tone as reference), and their interaction. The models are estimated for the human decision-maker condition (columns 1-3) and the full sample (columns 4-6). Models control for the decision-maker treatment (human as reference) and for experience (positive experience as reference). Pre-treatment covariates include gender, age, race, education, ideology, trust in technology, MTurk HIT record, attentiveness, self-reported suitability for the cataloging task, and indifference between tasks. Standard errors are in parentheses. See Table SI-8 for the full results †p>0.1 *p<0.05; **p<0.01, ***p<0.001

of the topic of AI, is affecting workers' attitudes. For instance, by making people feel more optimistic towards the future and thus more open to supporting innovations in public policy. Our experimental design allows us to test this possibility by dividing the placebo condition into positive and negative predictions about fashion trends. The results show that the coefficients for the tone of the information are not significant at any level. In contrast, the interaction term between positive information and AI is consistently positive and statistically significant at the 1% level across all models. This indicates that participants who were exposed to positive information, specifically about the consequences of AI, grew more supportive of using algorithms for decision-making in public policy.

Table SI-8 presents the results of the effect of exposure to information about AI on support for unrelated policy proposals, such as using background checks for gun purchases and deploying minimal quotas for women on company boards. As the table makes clear, no effect whatsoever was registered. Taken together, we can conclude that the shift to more supportive views is not the result of exposure to positive comments and frames in general but rather is directly tied to pertinent information about AI leading people to update their views on relevant policy questions.[16] This may partly be due to the limited role partisanship plays in the current debate over the regulation of AI. As our pretreatment survey shows, there is no significant difference in attitudes toward AI policy between Democrat and Republican workers (1.9, se= 0.92).

Next, we examine in which policy areas participants were more responsive to new information. The right panel of Figure 7 displays the marginal effects estimated separately for each policy decision. The results indicate that the pattern is not driven by one specific policy domain or context, tet attitudes did shift more in some areas than others. Specifically, individuals were more receptive to revising their policy views following exposure to new information related to decisions about the allocation of resources, such as foodstamps (p<0.01); homeless shelters (p<0.05), or increased personnel for enforcement of illegal construction (p<0.01). By dichotomizing the individual items, Table SI-9 shows that in these policy domains, the information led to a substantial change. For example, participants who encountered comments about the potential benefits of AI were 12 (p<0.01) percentage points more likely to support the use of AI for deciding where to increase enforcement, while those

[16] We also conducted a bounding exercise to address selective attrition. We assigned workers who did not complete the post-treatment survey either their pre-treatment outcome from wave 1 (lower bound) or their FA scores from responses given right after exposure to the treatment (upper bound). These measures thus assume either no change or complete change in attitudes, respectively. Table SI-10 shows that our main findings are robust to these different assumptions. This finding reinforces our conclusion that dropout between waves 2 and 3 does not pose a serious threat to estimating the treatment effects.

who learned about AI's potential risks were 8 percentage points less likely to support the use of the technology for making these decisions (p<0.05), compared to respondents who had received no information about AI.[17]

Conversely, this information had little effect on attitudes towards the use of AI in policies related to punishment, especially of individuals, such as granting parole or issuing restraining orders. Notably, in these domains, AI received the lowest levels of support in the pre-treatment survey. Indeed, previous research suggests that citizens are highly sensitive to human involvement in decisions that can have irreversible consequences on individuals' lives (Raviv, 2023). Our analysis indicates that in such domains, people are less likely to change their stance in response to new information.

5.4 Predispositions and Information Processing

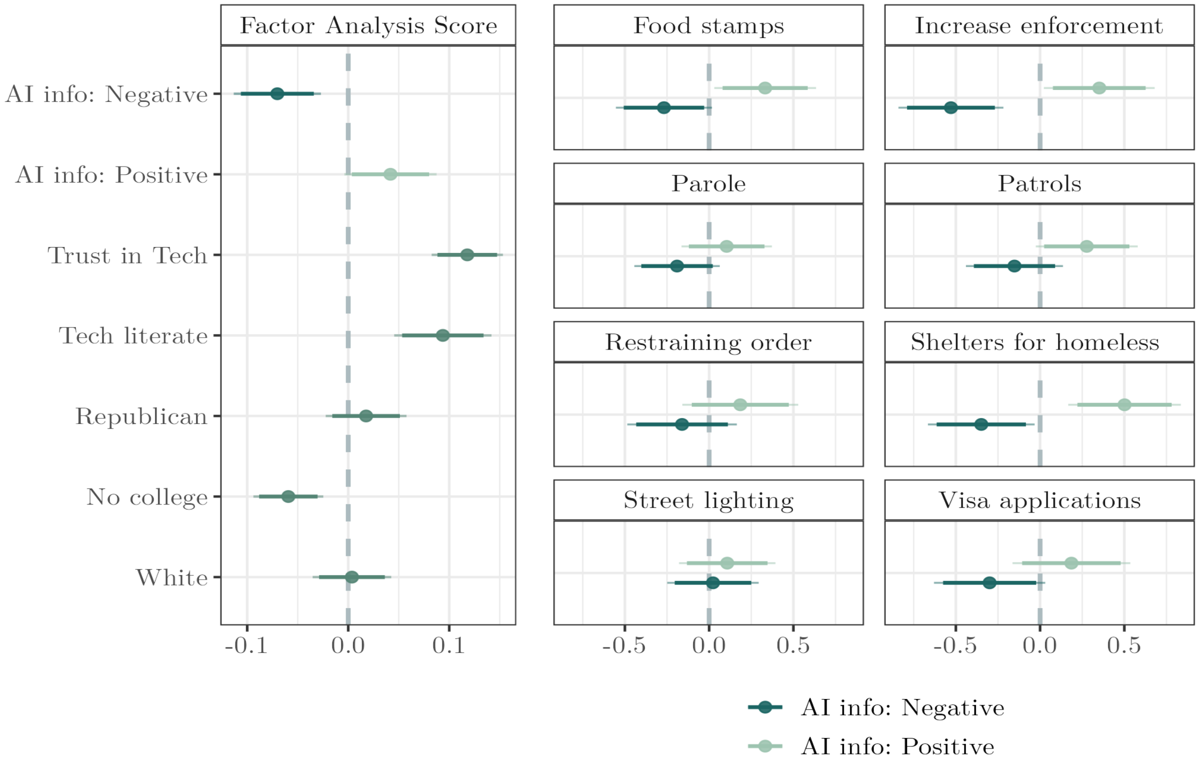
Our findings indicate that exposure to information about AI affects support for the use of the technology in policy areas where the public does not have a clear preference for a particular decision-maker. This raises the question of whether the information only reinforces the opinions of those who already agree with it or if it persuades those with opposing viewpoints.

To address this question, we divide our sample based on workers' predispositions, measured relative to the median score of the pre-treatment outcome. We then estimate the impact of information on AI for each group of workers separately, as well as the interaction between treatments and the predisposition. Table SI-11 reports the results. Recall that we included in the treatment a comment that was intentionally in the opposite direction of prevailing tone of the other comments, so that participants would have the option of "picking and choosing" evidence that is consistent with their predisposition.

We find that exposure to positive information about AI significantly increased support for its use in public policy implementation among both groups of workers, irrespective of

[17] This analysis was not pre-registered; we conducted it to help illustrate the substantive size of the effects.

Figure (7) Effects of Exposure to Information Treatments

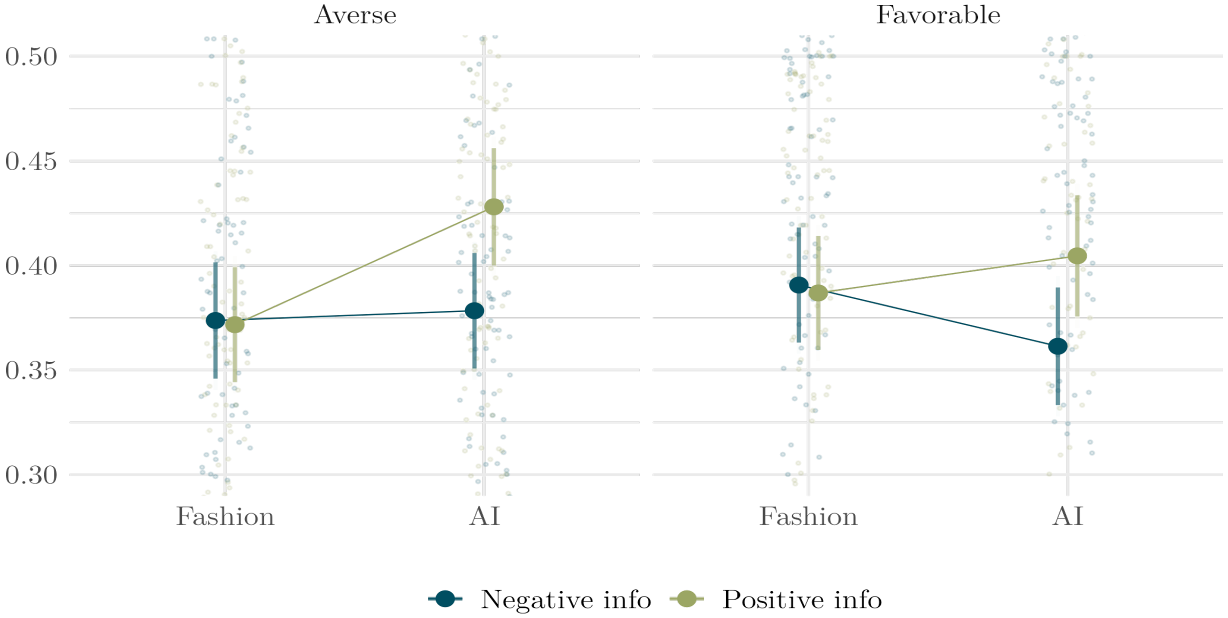


The figure shows the results of OLS regressions. The independent variables are indicators for positive information on AI, negative information on AI, or placebo information about fashion. The left panel shows the estimated treatment effects relative to the effects associated with key covariates, excluding the pre-treatment outcome. The right panel shows the treatment effects estimated separately for each policy domain. Models include controls for pre-treatment covariates, pre-treatment outcomes, as well as indicators for the experience and valence of information. We limited the sample to workers who were assigned to the human treatments to ensure a robust comparison. See Table SI-9 for the results. Thin bars represent 95% CI, and thick bars represent 90% CI.

their predispositions. The interaction between positive information on AI and negative predisposition is not statistically significant.

Figure 8 graphically illustrates the results, showing the predicted outcome by treatments and predispositions. Contrary to what motivated reasoning theory would suggest, the figure shows that workers who were initially skeptical of AI actually updated their views in response to reading positive information and grew more favorable of its use for policy implementation decisions. In contrast, exposure to negative information about AI appears to have had little

Figure (8) Treatment Effects by Predispositions



The figure shows the predicted FA Score of responses to the eight items in Wave 3, based on the interaction between the information treatment and predispositions. Error bars show 95% confidence interval. The model controls for decision-maker and experience treatments, the pre-treatment outcome (as a continuous measure) and demographic covariates. Column 5 in Table SI-11 shows the full results. Data points correspond to individual raw observations.

impact on participants' views.

We test for ceiling and floor effects by excluding the respondents identified in the baseline survey as either most opposed or most supportive of AI use in policy decisions (i.e., those in the upper and lowest deciles of the scale). We also re-ran the analysis while excluding respondents around the midpoint of the scale, as one might worry that they are simply indifferent. The findings remain consistent under these exclusions. Overall, then, our analysis indicates that rather than rejecting or ignoring information that challenges their prior views, participants updated their views in the direction of the information they received. This finding is consistent with research showing that people from different groups respond to persuasive information in the same direction (Coppock, 2022). One possible explanation for this finding is that attitudes toward AI in public policy are not deeply held, at least at this stage when the issue is not yet politicized, and thus are more likely to change when they encounter relevant information. We return to this issue in the concluding section below.

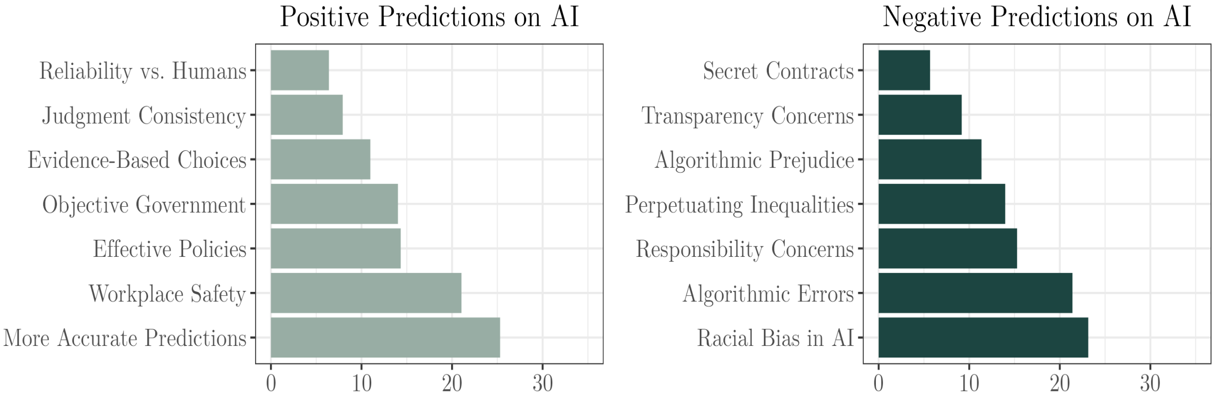
5.5 What Type of Information about AI Affects Attitudes?

To better understand the effect of information, we analyze which specific comments participants found most persuasive. This analysis is based on their responses, after they had completed the task, to a question asking what comment they found most convincing and why. We then constructed a dictionary for each comment, listing its key phrases and words and use the responses to the open-ended questions to identify the comments each treatment group found most persuasive.

Figure 9 presents the results. Among workers exposed to positive information, the most persuasive comments were those that emphasized AI's high degree of accuracy (26%) and its potential to enhance workplace safety (21%). Notably, these comments were considered far more persuasive than those highlighting the limitations of human decision-makers. For instance, only 7% of the workers cited the comment, "AI might lead to more consistent judgments than those made by humans, who may be influenced by emotional considerations or by fatigue,". Similarly, only 6% mentioned the comment emphasizing AI's reliability as compared to humans who can be influenced "by irrelevant factors, such as their mood." Among workers who received negative information, the figure shows that comments that addressed concerns about racial discrimination and unfairness (25.5%) or potential issues with utilizing aggregate data for individual decision-making (23%) were more frequently mentioned.

The findings suggest that workers' attitudes toward AI are responsive to specific pieces of information about its potential implications, but it is not the case that they changed their views in response to any positive assessment of the technology's merits.

Figure (9) Most Persuasive Comments among Workers Exposed to Information about AI



The figure shows the percentage of individuals in the positive and negative treatment groups (left and right panels, respectively) who cited each comment as the most persuasive. The comments were identified based on key phrases extracted from participants' open-ended responses.

6 Discussion

The growing use of AI-based algorithms in policy implementation is changing a fundamental component of democratic governance, namely the way important decisions affecting citizens' lives are made. It is therefore essential that the development and deployment of AI-based systems reflect the values and preferences of the public. Recognizing this important need, both governments and leading tech companies are advancing initiatives that foster public input in setting the norms and rules for the governance of AI. Yet such initiatives give rise to questions about what the public's views on this issue are and about how the views will evolve in response to personal experience with AI and exposure to information about the technology's potential impacts. This study provides the first systematic examination of these questions.

Our analysis indicates that people not only update their views on the use of AI in policy settings when presented with relevant new information, but do so even when the information does not conform with their prior views or inclinations. This type of openness to influence strikes us as far from obvious and may partly reflect the fact that the debate over AI regulation is not yet politicized. Indeed, as our baseline survey reveals, there is no significant difference in the attitudes of Republicans and Democrats on this issue, and another recent survey of policymakers also finds very little partisan differences on issues related to the regulation of AI (Schiff and O'Shaughnessy, 2023).These findings point to the potential—perhaps only a temporary one—for creating broad coalitions that span across the political spectrum and promote AI governance that is centered on safeguarding the public interest rather than those of partisan special interest groups.

Related to the point above, the findings also highlight the importance of informing the public early on about AI's potential benefits and risks, since the period of openness to information and to meaningful updating of views may be fleeting. Instead, people's attitudes might soon be shaped by partisanship, as happened with other policy issues that require expert knowledge but that underwent profound politicization (e.g., climate change, vaccinations).[18] The debate over AI regulation may soon undergo a similar dynamic.

Specifically, it is easy to imagine that business interests and large corporations are likely to have a strong interest in emphasizing AI's benefits and lobbying for weaker regulation, while civil society groups might put greater emphasis on the technology's potential harmful implications (e.g., on privacy, social justice) and push for deeper government involvement. The extent to which such messages will shape public opinion on the use of AI is an empirical question with potentially weighty implications, one that will surely require serious attention in the coming years.

Another key finding in our study is that participants did not seem to infer from their personal experience with AI as a decision-maker to the broader question of the appropriate

[18] Recall, for example, that in the 1960s there was a relatively broad consensus on environmental regulation, and a partisan divide began to emerge only in the 1990s (Hochschild, 2021). In fact, the Environmental Protection Agency was established in 1970 by a Republican president, Richard Nixon. But within three decades, the partisan gap had sharply increased: by 1990, 91% of Democrats but only 33 percent of Republicans expressed concern about climate change Brenan and Saad, 2018

use of AI in public policy decisions. One possibility is that people simply do not make the link between their personal experience and the broader policy question at hand. Yet another possibility is that people do make the connection but view the societal impact of AI in more normative lens, and hence go beyond their own interests or experiences when forming their attitudes. This possibility would be consistent with work that documented individuals' negative reactions toward algorithmic systems that risk public values such as fairness and transparency, even if they themselves are not likely to be directly affected by these decisions (Schiff, Schiff, and Pierson, 2022). One way to explore this question is to investigate how experience with AI-based decisions in more proximate public policy settings, such as being approved (or denied) a visa, a permit or a social benefit, influences preferences for replacing human decision-makers with AI in making policy implementation decisions. Such a study would speak to the generalizability of our findings regarding the disconnect between participants' personal experience and their preferences regarding the broader policy question.

Another promising direction for research would be to examine how exposure to AI algorithms in the labor market influences public opinion on other policy issues that are more directly connected to this experience. For example, toward policy interventions aimed at mitigating some of the negative effects of automation in the labor market, such as government-funded assistance and re-skilling programs. Specifically, experience with AI-as-boss could give workers a more concrete sense of what automation means for non-routine occupations. This in turn could affect their perceptions of the risks that automation poses, as well as shape their preferences regarding policy interventions designed to deal with these potential risks.

References

Anderson, Janna, Lee Rainie, and Alex Luchsinger (2018). “Artificial intelligence and the future of humans”. In: *Pew Research Center* 10.12.

Anelli, Massimo, Italo Colantone, and Piero Stanig (2019). “We were the robots: Automation and voting behavior in western europe”. In: *BAFFI CAREFIN Centre Research Paper* 2019-115.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner (2016). “Machine bias”. In: *Ethics of data and analytics*. Auerbach Publications, pp. 254–264.

Ansolabehere, Stephen, Jonathan Rodden, and James M Snyder (2008). “The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting”. In: *American Political Science Review* 102.2, pp. 215–232.

Anzia, Sarah F, Jake Alton Jares, and Neil Malhotra (2022). “Does Receiving Government Assistance Shape Political Attitudes? Evidence from Agricultural Producers”. In: *American Political Science Review* 116.4, pp. 1389–1406.

Araujo, Theo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese (2020). “In AI we trust? Perceptions about automated decision-making by artificial intelligence”. In: *AI & society* 35, pp. 611–623.

Austin, James E, Howard Stevenson, and Jane Wei-Skillern (2006). “Microfinance and Social Development: A Selective Literature Review”. In: *AI & SOCIETY* 21.4, pp. 355–364.

Bansak, Kirk, Jeremy Ferwerda, Jens Hainmueller, Andrea Dillon, Dominik Hangartner, Duncan Lawrence, and Jeremy Weinstein (2018). “Improving refugee integration through data-driven algorithmic assignment”. In: *Science* 359.6373, pp. 325–329.

Bansak, Kirk and Elisabeth Paulson (2023). “Public opinion on fairness and efficiency for algorithmic and human decision-makers”. In: *Working Paper*.

Bicchi, Nicolas, Aina Gallego, and Alexander Kuo (2023). *Workers support for policies to address digitalization-related risks*. Tech. rep. Joint Research Centre (Seville site).

Boudet, Hilary S (2019). “Public perceptions of and responses to new energy technologies”. In: *nature energy* 4.6, pp. 446–455.

Brenan, Megan and Lydia Saad (Mar. 28, 2018). *Global Warming Concern Steady Despite Some Partisan Shifts*.

Broockman, David E, Joshua L Kalla, and Jasjeet S Sekhon (2017). “The design of field experiments with survey outcomes: A framework for selecting more efficient, robust, and ethical designs”. In: *Political Analysis* 25.4, pp. 435–464.

Burbano, Vanessa C (2016). “Social responsibility messages and worker wage requirements: Field experimental evidence from online labor marketplaces”. In: *Organization Science* 27.4, pp. 1010–1028.

Christenson, Dino P and David M Glick (2013). “Crowdsourcing panel studies and real-time experiments in MTurk”. In: *The Political Methodologist* 20.2, pp. 27–32.

Cobb, Michael D and Jane Macoubrie (2004). “Public perceptions about nanotechnology: Risks, benefits and trust”. In: *Journal of Nanoparticle Research* 6, pp. 395–405.

Coppock, Alexander (2022). “Persuasion in parallel”. In: *Persuasion in Parallel*. University of Chicago Press.

Dietvorst, Berkeley J, Joseph P Simmons, and Cade Massey (2015). “Algorithm aversion: people erroneously avoid algorithms after seeing them err.” In: *Journal of Experimental Psychology: General* 144.1, p. 114.

Druckman, James N and Toby Bolsen (2011). “Framing, motivated reasoning, and opinions about emergent technologies”. In: *Journal of Communication* 61.4, pp. 659–688.

Egan, Patrick J and Megan Mullin (2012). “Turning personal experience into political attitudes: The effect of local weather on Americans’ perceptions about global warming”. In: *The Journal of Politics* 74.3, pp. 796–809.

Evgeniou, Theodoros, David R. Hardoon, and Anton Ovchinnikov (n.d.). *What Happens When AI is Used to Set Grades?*

Gallego, Aina, Alexander Kuo, Dulce Manzano, and José Fernández-Albertos (2022). “Technological risk and policy preferences”. In: *Comparative Political Studies* 55.1, pp. 60–92.

Guardian, The (2021). “Dutch Government Resigns Over Child Benefits Scandal”. In: Accessed: 2024-06-04. URL: https://www.theguardian.com/world/2021/jan/15/dutch-government-resigns-over-child-benefits-scandal.

Haring, Kerstin Sophie, David Silvera-Tawil, Katsumi Watanabe, and Mari Velonaki (2016). “The influence of robot appearance and interactive ability in HRI: a cross-cultural study”. In: *Social Robotics: 8th International Conference, ICSR 2016, Kansas City, MO, USA, November 1-3, 2016 Proceedings 8*. Springer, pp. 392–401.

Healy, Andrew and Neil Malhotra (2013). “Retrospective voting reconsidered”. In: *Annual Review of Political Science* 16, pp. 285–306.

Heilweil, Rebecca (2020). “Big tech companies back away from selling facial recognition to police. That’s progress”. In: *Vox, June* 11.

Hochschild, Jennifer (2021). *Genomic politics: how the revolution in genomic science is shaping American society*. Oxford University Press.

Horowitz, Michael C (2016). “Public opinion and the politics of the killer robots debate”. In: *Research & Politics* 3.1, p. 2053168015627183.

Horton, John J, David G Rand, and Richard J Zeckhauser (2011). “The online laboratory: Conducting experiments in a real labor market”. In: *Experimental economics* 14, pp. 399–425.

Kennedy, Ryan P, Philip D Waggoner, and Matthew M Ward (2022). “Trust in public policy algorithms”. In: *The Journal of Politics* 84.2, pp. 1132–1148.

Kurer, Thomas and Silja Hausermann (2022). “Automation Risk, Social Policy Preferences, and Political Participation”. In: *Digitalization and the welfare state*, p. 139.

Lapinsky, Stephen E, Randy S Wax, Randy Showalter, Manuel Martinez-Maldonado, David C Hallett, Peter D Austin, and Thomas E Stewart (2008). “Hepatocellular binding of drugs: correction for unbound fraction in hepatocyte incubations using microsomal binding or drug lipophilicity data”. In: *Drug metabolism and disposition* 36.7, pp. 1194–1197.

Lee, Chul-Joo, Dietram A Scheufele, and Bruce V Lewenstein (2005). “Public attitudes toward emerging technologies: Examining the interactive effects of cognitions and affect on public attitudes toward nanotechnology”. In: *Science communication* 27.2, pp. 240–267.

Lorinc, John (2022). *Dream States: Smart Cities, Technology, and the Pursuit of Urban Utopias*. Coach House Books.

Mahmud, Hasan, AKM Najmul Islam, Syed Ishtiaque Ahmed, and Kari Smolander (2022). “What influences algorithmic decision-making? A systematic literature review on algorithm aversion”. In: *Technological Forecasting and Social Change* 175, p. 121390.

Management, United States Office of and Budget (2020). *Memorandum on Guidance for Regulation of Artificial Intelligence Applications M-21-06*.

Margalit, Yotam and Moses Shayo (2021). “How markets shape values and political preferences: A field experiment”. In: *American Journal of Political Science* 65.2, pp. 473–492.

Mays, Kate K, Yiming Lei, Rebecca Giovanetti, and James E Katz (2021). “AI as a boss? A national US survey of predispositions governing comfort with expanded AI roles in society”. In: *AI & Society*, pp. 1–14.

McConnell, Christopher, Yotam Margalit, Neil Malhotra, and Matthew Levendusky (2018). "The economic consequences of partisanship in a polarized era". In: *American Journal of Political Science* 62.1, pp. 5–18.

Miller, Susan M and Lael R Keiser (2021). "Representative bureaucracy and attitudes toward automated decision making". In: *Journal of Public Administration Research and Theory* 31.1, pp. 150–165.

Mutz, Diana C (1994). "Contextualizing personal experience: The role of mass media". In: *The Journal of Politics* 56.3, pp. 689–714.

News, CBC (2023). *Hit pause on AI development, Elon Musk and others urge.*

O'Kane, Josh (2022). *Sideways: The City Google Couldn't Buy.*

Raviv, Shir (2023). "When Do Citizens Resist The Use of Algorithmic Decision-making in Public Policy? Theory and Evidence". In: *SSRN*.

Scheufele, Dietram A and Bruce V Lewenstein (2005). "The public and nanotechnology: How citizens make sense of emerging technologies". In: *Journal of nanoparticle research* 7, pp. 659–667.

Schiff, Daniel S, Kaylyn Jackson Schiff, and Patrick Pierson (2022). "Assessing public value failure in government adoption of artificial intelligence". In: *Public Administration* 100.3, pp. 653–673.

Schiff Daniel S, Schiff Jackson Kaylyn and Matthew O'Shaughnessy (2023). "Innovation, Ethics, and Public Participation: How Do US State Legislators View AI Policy?" In: *Working Paper*.

Schöll, Nikolas and Thomas Kurer (2023). "How technological change affects regional voting patterns". In: *Political Science Research and Methods*, pp. 1–19.

Stoutenborough, James W and Arnold Vedlitz (2016). "The role of scientific knowledge in the public's perceptions of energy technology risks". In: *Energy Policy* 96, pp. 206–216.

Taber, Charles S and Milton Lodge (2006). "Motivated skepticism in the evaluation of political beliefs". In: *American journal of political science* 50.3, pp. 755–769.

Toros, Halil and Daniel Flaming (2018). "Prioritizing homeless assistance using predictive algorithms: an evidence-based approach". In: *Cityscape* 20.1, pp. 117–146.

Ullman, Daniel and Bertram F. Malle (2017). "Human-Robot Trust: Just a Button Press Away". In: *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*, pp. 309–310. DOI: 10.1145/3029798.3038423.

Walsh, Bryan (2020). “How an AI grading system ignited a national controversy in the U.K.” In: *Axios*.

Wenzelburger, Georg and Anja Achtziger (2023). “Algorithms in the public sector. Why context matters”. In: *Public Administration* 101.1, 1–18.

White-House (2022). *Blueprint for an AI Bill of Rights–Making Automated Systems work for the American People*.

Winston, Ali (2018). “Palantir has secretly been using New Orleans to test its predictive policing technology”. In: *The Verge* 27.

Yeomans, Michael, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg (2019). “Making sense of recommendations”. In: *Journal of Behavioral Decision Making* 32.4, pp. 403–414.

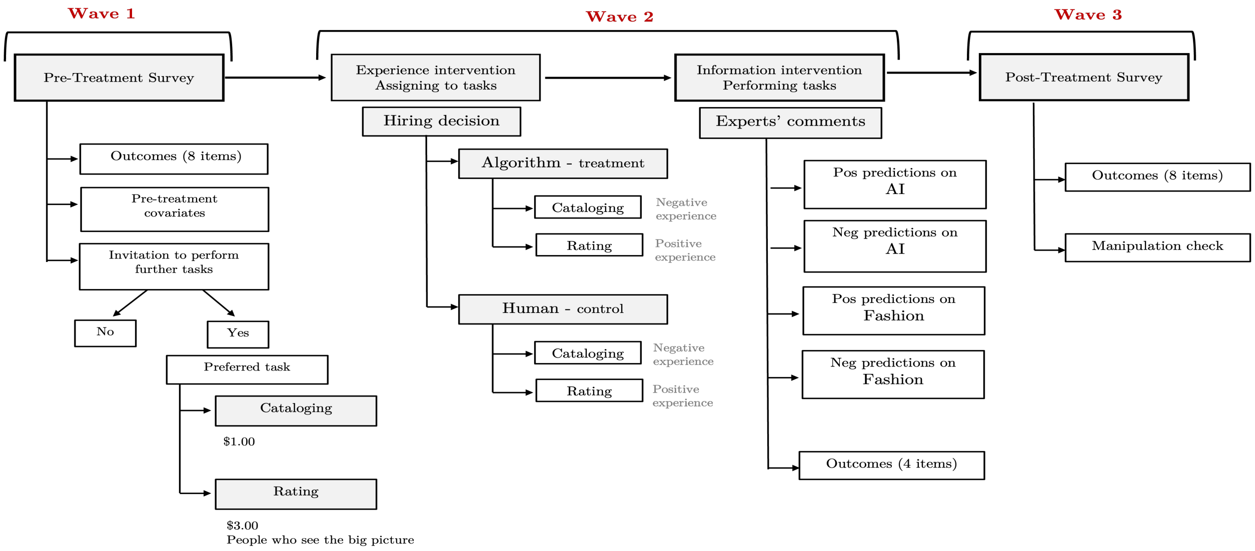
Yeung, Karen (2020). “Recommendation of the council on artificial intelligence (oecd)”. In: *International legal materials* 59.1, pp. 27–34.

Zhang, Baobao (2021). “Public Opinion Toward Artificial Intelligence”. In: *The Oxford Handbook of Artificial Intelligence Governance*. Ed. by TBA. Forthcoming. Oxford University Press.

Supplementary Materials
The Politics of Using AI in Policy: A Field Experiment

A Experimental Design

Figure (SI-1) Experimental Design



B Attrition

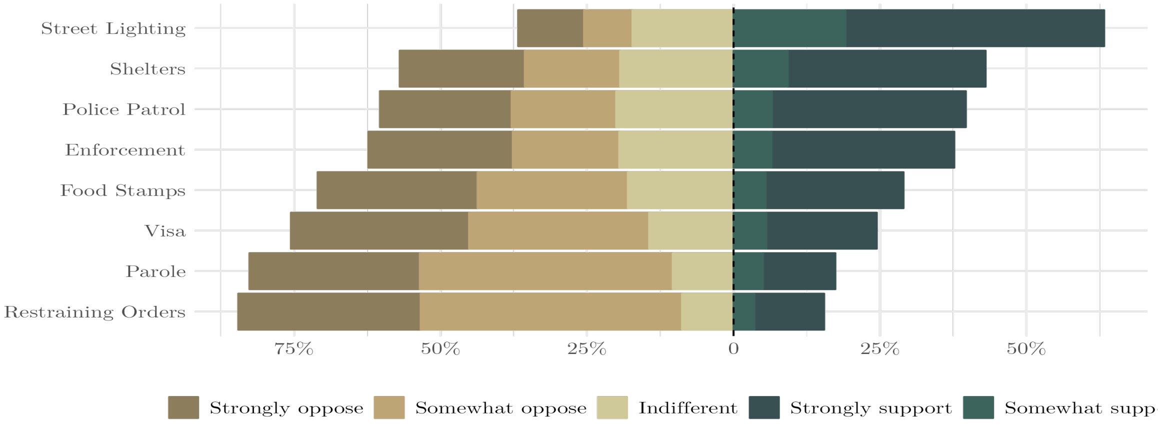
A total of 3,468 individuals participated in the baseline survey. We excluded 489 inattentive participants and 25 individuals who dropped out during the survey. Of the remaining participants, 222 (7%) declined to participate in future tasks and were removed from the panel study. To reduce potential noise in estimates from CACE and to adhere to budget constraints, we also excluded 324 participants who agreed to participate in future tasks but preferred the low-pay option (10% of the sample). In total, 2408 participants completed the pre-treatment survey, indicated a preference for the high-status and were randomly assigned to the experimental groups. Of them, 1875 entered the link (78%), and 1796 completed the task in wave 2 (74.5% of the invited participants).

Table SI-1 shows the number of participants who completed Wave 2 compared to dropouts as a function of treatment assignment. The results of the pairwise proportion test show significant differences in completion rates based on the type of experience (p=0.012), regardless of the identity of the decision maker. In contrast, there are no significant differences in completion rates in Wave 3 (all p-values > 0.05).

Table (SI-1) Attrition and Completion Rates by Treatment Assignment

	Algorithm X Neg	**Algorithm X Pos**	**Human X Neg**	**Human X Pos**
Completed wave 2	435	476	419	466
Dropped after clicked link	20 (4.3%)	12 (2.5%)	29 (6.5%)	18 (3.72%)
Completed wave 3	368	391	355	388
Dropped from wave 2 to 3	67 (15.5%)	85 (17.9%)	64 (15.3%)	78 (16.8%)

Figure (SI-2) Attitudes toward AI in public policy, pre-treatment



B.1 Descriptive Statistics

Table SI-2 presents descriptive statistics on pre-treatment values for a range of demographic and attitudinal variables, including all outcome variables used in our main analyses.

Table (SI-2) Descriptive Statistics

Characteristic	**N/Mean (%/SD)**
Overall	1513
Female	0.59 (0.49)
Age 18-34	581 (38.4%)
Age 35-45	459 (30.3%)
Age 45+	473 (31.3%)
White(%)	1160 (77%)
Low Education (Some College) (%)	430 (28%)
Conservative	488 (32.3%)
Moderate	338 (22.3%)
Liberal	687 (45.4%)
High Trust in Technology (%)	499 (33.0%)
Low Literacy	444 (29.3%)
Med Literacy	733 (48.4%)
High Literacy	336 (22.2%)
Chat GPT	0.16 (0.37)
Factor A (8 items)	0.39 (0.23)
Outcome 1: Parole	2.57 (1.85)
Outcome 1: Food Stamps	3.31 (1.89)
Outcome 1: Patrol	3.77 (1.84)
Outcome 1: Street Lighting	4.83 (1.78)
Outcome 1: Restraining Orders	2.41 (1.74)
Outcome 1: Visa	3.04 (1.90)
Outcome 1: Illegal Building	3.71 (1.85)
Outcome 1: Shelters	3.97 (1.89)

B.2 Attitudes toward AI in public policy, Pre-treatment

Figure SI-2 shows the preference distribution for the pre-treatment outcomes. We measured the responses on a seven-point scale and then classified them into five categories: Strongly Oppose (1), Oppose (2-3), Indifferent (4), Support (5-6), and Strongly Support (7).

B.3 Balance Tables

Table SI-3 presents descriptive statistics for participants in the baseline survey on a wide range of demographic and attitudinal variables, including all outcome variables used in subsequent analyses. The table includes only individuals who completed the post-treatment survey in wave 3, divided by their assignment to the information treatment. We present both descriptive results and statistical tests to examine the balance of covariates. The table shows that the sample is well balanced.

Table (SI-3) Balance table, Information (Content and Valence)

	Fashion Positive Info	Fashion Negative Info	AI Positive Info	AI Negative Info	p-value (Chi-S test)
n	384	376	366	374	
Female	0.62 (0.49)	0.61 (0.49)	0.55 (0.50)	0.56 (0.50)	0.129
Age Category (%)					0.904
18–34	154 (40.1)	143 (38.0)	137 (37.4)	140 (37.4)	
35–45	111 (28.9)	114 (30.3)	118 (32.2)	111 (29.7)	
45+	119 (31.0)	119 (31.6)	111 (30.3)	123 (32.9)	
White (%)	0.75 (0.44)	0.77 (0.42)	0.76 (0.43)	0.79 (0.41)	0.630
High Education (BA)	0.29 (0.46)	0.28 (0.45)	0.31 (0.46)	0.26 (0.44)	0.579
Political Orientation (%)					0.688
Conservative	121 (31.5)	130 (34.6)	113 (30.9)	119 (31.8)	
Moderate	91 (23.7)	79 (21.0)	90 (24.6)	76 (20.3)	
Liberal	172 (44.8)	167 (44.4)	163 (44.5)	179 (47.9)	
High Trust in Tech (%)	133 (34.6)	122 (32.4)	122 (33.3)	120 (32.1)	0.882
Tech Literacy (%)					0.958
Low Literacy	115 (29.9)	105 (27.9)	110 (30.1)	110 (29.4)	
Med Literacy	185 (48.2)	184 (48.9)	180 (49.2)	178 (47.6)	
High Literacy	84 (21.9)	87 (23.1)	76 (20.8)	86 (23.0)	
Factor A	0.40 (0.22)	0.41 (0.23)	0.39 (0.24)	0.38 (0.23)	0.191

Next, we examine whether the treatment groups differ in their background covariates based on their experience with the decision maker. Table SI-4 displays the pre-treatment demographic and attitudinal features of the participants, divided into four groups: Negative or positive experience with algorithmic DM, and negative or positive experience with human DM. The groups are well-balanced in their characteristics and prior opinions on AI in public policy, as most outcome variables show no significant differences across them.

Table (SI-4) Balance Table, by Type of Experience and Decision Maker

	Algorithm		Human		
	Negative experience	Positive experience	Negative experience	Positive experience	p-value
N	368	391	355	388	
Female	0.57 (0.50)	0.58 (0.49)	0.61 (0.49)	0.59 (0.49)	0.406
Age Category (%)					0.0.218
18–34	134 (36.4)	164 (41.9)	124 (34.9)	153 (39.4)	
35-45	122 (33.2)	118 (30.2)	109 (30.7)	106 (27.3)	
45+	112 (30.4)	109 (27.9)	122 (34.4)	129 (33.2)	
White = 1 (%)	0.74 (0.44)	0.77 (0.42)	0.81 (0.39)	0.75 (0.43)	0.262
High Education BA	0.27 (0.44)	0.28 (0.45)	0.29 (0.45)	0.30 (0.46)	0.842
Political Views (%)					0.479
Conservative	120 (32.6)	113 (28.9)	112 (31.5)	139 (35.8)	
Moderate	81 (22.0)	84 (21.5)	87 (24.5)	85 (21.9)	
Liberal	167 (45.4)	194 (49.6)	156 (43.9)	164 (42.3)	
Tech Literacy (%)					0.958
Low Literacy	105 (28.5)	118 (30.2)	111 (31.3)	106 (27.3)	
Med Literacy	178 (48.4)	184 (47.1)	168 (47.3)	198 (51.0)	
High Literacy	(18.2)	85 (23.1)	89 (22.8)	76 (21.4)	84 (21.6)
High Trust in Tech (%)	120 (32.6)	138 (35.3)	109 (30.7)	130 (33.5)	0.568
Factor A	0.40 (0.22)	0.39 (0.24)	0.40 (0.23)	0.39 (0.24)	0.858

C Experience with Algorithmic decision making

C.1 Alternative Measures of Attitudinal Outcomes

The outcome variable in the main analysis is a standardized measure based on a factor analysis of responses to the 8-item matrix asked in the post-treatment survey several days after receiving the treatments, where higher values indicate greater support for using AI. Below, we test the results using the following alternative measures of the outcome: (1) use of principal component analysis instead of factor analysis of the eight items asked in the third-wave survey and (2) factor analysis score of the 4 items collected in Wave 2, right after completing the task. Table SI-5 shows that the results remain very similar when using these alternative measures.

Table (SI-5) Effects of Experience on Alternative Outcomes

	Dependent variable:																	
	Factor Analysis Score - Wave 2									Principle Component Analysis - Wave 3								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
	(ITT)	(ITT)	(ITT)	(TOT)	(ITT)	(ITT)	(ITT)	(TOT)	(TOT)	(ITT)	(ITT)	(ITT)	(TOT)	(ITT)	(ITT)	(ITT)	(TOT)	(TOT)
Algorithmic DM	−0.00001 (0.010)	0.008 (0.008)	0.009 (0.008)	0.015 (0.012)	−0.002 (0.014)	0.008 (0.011)	0.010 (0.011)	0.018 (0.018)	0.023 (0.017)	0.002 (0.011)	0.005 (0.008)	0.006 (0.008)		−0.005 (0.015)	0.004 (0.011)	0.006 (0.011)	0.009 (0.019)	0.013 (0.019)
Algorithm X Negative exp					0.005 (0.021)	0.001 (0.015)	−0.001 (0.015)	−0.007 (0.026)	−0.012 (0.026)					0.016 (0.022)	0.001 (0.016)	−0.00002 (0.016)	−0.004 (0.028)	−0.007 (0.028)
Negative Experience	−0.003 (0.010)	0.006 (0.008)	0.007 (0.008)	0.006 (0.008)	−0.006 (0.015)	0.005 (0.011)	0.007 (0.011)	0.010 (0.015)	0.013 (0.015)	−0.021 (0.011)	−0.005 (0.008)	−0.003 (0.008)	−0.005 (0.008)	−0.028 (0.015)	−0.006 (0.012)	−0.003 (0.012)	−0.004 (0.017)	−0.002 (0.016)
AI Info	0.003 (0.010)	−0.020** (0.008)	−0.021** (0.008)	−0.018* (0.008)	0.003 (0.010)	−0.020** (0.008)	−0.021** (0.008)	−0.018* (0.008)	−0.018* (0.008)	−0.008 (0.011)	−0.031** (0.008)	−0.033** (0.008)	−0.033** (0.008)	−0.008 (0.011)	−0.031** (0.008)	−0.033** (0.008)	−0.033** (0.008)	−0.034** (0.008)
Negative Info		0.014 (0.008)	0.015 (0.008)	0.016* (0.008)		0.014 (0.008)	0.015 (0.008)	0.016* (0.008)	0.017* (0.008)		0.046** (0.008)	0.045** (0.008)	0.048** (0.008)		0.046** (0.008)	0.045** (0.008)	0.048** (0.008)	0.048** (0.008)
Pretreatment (PCA/FA)	0.786** (0.023)	0.739** (0.017)	0.699** (0.018)	0.752** (0.017)	0.786** (0.023)	0.739** (0.017)	0.699** (0.018)	0.752** (0.017)	0.712** (0.019)	0.795** (0.024)	0.731** (0.017)	0.688** (0.018)	0.744** (0.017)	0.795** (0.024)	0.731** (0.017)	0.688** (0.018)	0.743** (0.017)	0.702** (0.019)
Female			−0.002 (0.008)				−0.002 (0.008)		−0.004 (0.008)			−0.006 (0.009)				−0.006 (0.009)		−0.006 (0.009)
Age: 35-45			−0.004 (0.009)				−0.004 (0.009)		−0.001 (0.009)			−0.004 (0.010)				−0.004 (0.010)		0.001 (0.010)
Age: 45+			0.010 (0.010)				0.010 (0.010)		0.011 (0.010)			0.030** (0.010)				0.030** (0.010)		0.028** (0.010)
White			0.013 (0.009)				0.013 (0.009)		0.011 (0.009)			0.021* (0.010)				0.021* (0.010)		0.021* (0.010)
Low Education			−0.013 (0.009)				−0.013 (0.009)		−0.013 (0.009)			−0.021* (0.009)				−0.021* (0.009)		−0.023* (0.009)
Moderate			−0.010 (0.011)				−0.010 (0.011)		−0.015 (0.011)			−0.030** (0.011)				−0.030** (0.011)		−0.035** (0.012)
Liberal			−0.023* (0.009)				−0.023* (0.009)		−0.027** (0.009)			−0.034** (0.009)				−0.034** (0.009)		−0.037** (0.010)
Tech Literacy: Medium			0.017 (0.009)				0.017 (0.009)		0.014 (0.009)			0.001 (0.010)				0.001 (0.010)		0.001 (0.010)
Tech Literacy: High			0.018 (0.012)				0.018 (0.012)		0.013 (0.012)			0.013 (0.012)				0.013 (0.012)		0.013 (0.013)
Tech Trust			0.029** (0.009)				0.029** (0.009)		0.028** (0.009)			0.036** (0.009)				0.036** (0.009)		0.036** (0.010)
HIT percent			−0.007 (0.005)				−0.007 (0.005)		−0.011* (0.005)			−0.008 (0.006)				−0.008 (0.006)		−0.008 (0.006)
Inattentive			0.006 (0.008)				0.006 (0.008)		0.005 (0.008)			−0.007 (0.009)				−0.007 (0.009)		−0.007 (0.009)
Suit for Cataloging			−0.040* (0.017)				−0.040* (0.017)		−0.040* (0.018)			−0.035 (0.018)				−0.035 (0.018)		−0.035 (0.019)
Indifferent			−0.043 (0.023)				−0.043 (0.023)		−0.026 (0.025)			−0.046 (0.025)				−0.046 (0.025)		−0.025 (0.027)
Constant	0.102** (0.014)	0.124** (0.011)	0.141** (0.022)	0.115** (0.012)	0.103** (0.015)	0.125** (0.011)	0.140** (0.022)	0.113** (0.014)	0.147** (0.023)	0.091** (0.015)	0.122** (0.012)	0.154** (0.023)	0.115** (0.013)	0.095** (0.016)	0.122** (0.012)	0.154** (0.023)	0.115** (0.014)	0.147** (0.024)
Observations	760	1,500	1,497	1,433	760	1,500	1,497	1,433	1,430	760	1,498	1,495	1,432	760	1,498	1,495	1,432	1,429
R^2	0.598	0.569	0.580	0.580	0.598	0.569	0.580	0.579	0.590	0.603	0.557	0.576	0.568	0.603	0.557	0.576	0.568	0.587

Notes: LPM with standard errors in parentheses. DV are: PCA Score of 8 item (wave 3) in columns 1-8, and FA Score of 4 items (Wave 2) in columns 9-18. Columns 1, 5, 8, and 12 limit the sample to the placebo information, while the rest of columns control for information treatments and their valence. All models control for the pre-treatment outcome (Wave 1). *p<0.05; **p<0.01

C.2 Alternative Measures of Compliance

In the main analysis, we calculate treatment-on-the-treated estimates using IV regression. We define compliance as participants who indicated in the manipulation checks that: (1) they completed the rating task (high status-high pay task), and (2) the decision maker who assigned them to the task was the requester's algorithm.

In Table SI-6, we replicate the results using alternative measures of compliers: individuals who report that an algorithm, rather than a human, was responsible for assigning them to the task. This definition is broader and includes respondents who considered both the MTurk algorithm or the specific algorithm used by the requester.

Table (SI-6) Alternative Measures of Compliance

	Dependent variable:		
	Factor Analysis Score - Wave 3		
	(1)	(2)	(3)
TOT General Algorithm	0.013 (0.011)	0.013 (0.015)	0.016 (0.015)
ITT Negative Experience	0.005 (0.008)		
TOT Negative Experience		0.006 (0.016)	0.010 (0.016)
Negative Info	-0.019^{*} (0.008)	-0.019^{*} (0.008)	-0.020^{*} (0.008)
AI Info	0.014 (0.008)	0.014 (0.008)	0.014 (0.008)
Pretreatment (FA)	0.751^{**} (0.017)	0.752^{**} (0.017)	0.710^{**} (0.019)
Suit for Cataloging			-0.039^{*} (0.018)
Indifferent			−0.043 (0.024)
Constant	0.111^{**} (0.012)	0.110^{**} (0.014)	0.137^{**} (0.023)
Demographics	No	No	Yes
Observations	1,475	1,475	1,472
R^2	371.94	0.574	0.585
First stage F-statistic	471.27	355.29	110.55

C.3 Estimated Treatment Effects on Behavioral Outcomes, Full Results

In the main text, we assess whether the treatment was too weak to have any meaningful impact by examining its effect on the following behavioral outcomes:

- Accuracy in classifying comments with the opposite tone. We defined correct classification as classifying a comment with the opposite tone in the opposite direction compared to the dominant tone of the comments. For instance, for participants in the positive information treatment group, correct classification meant rating the comment with the opposite tone lower on the scale compared to the minimum rating among the positive comments. Conversely, for participants in the negative information treatment group, correct identification meant giving the comment with the opposite tone a higher rating than the maximum rating among the dominant tone comments.

- Time spent on the main classification tasks and the follow-up task. We used a binary variable that takes the value of '1' if the worker performed the task in a time longer than the median and 0 otherwise.

- Thoroughness in carrying out the task, measured by the number of clicks on the classification task. The task required at least eight clicks to complete, as there were eight comments to classify. We used a binary variable that takes the value of '1' if the worker made more than the median clicks, which is 13 clicks, meaning changing evaluation at least 5 times.

- Willingness to continue working with the same employer. We asked workers to suggest a wage for completing an additional task of similar scope and length. If a worker suggests a wage lower than the amount received for the current task, we use this as an indication of a high degree of willingness to continue working with the employer.

- Job satisfaction. We measure job satisfaction using an item that asks workers to rate their satisfaction with their task assignment. We used a binary variable that takes the value 1 if the worker was extremely satisfied with the task and 0 otherwise.

For ease of comparison across different outcomes, in Table SI-7 we report the full results of estimations using binary indicators and linear probability models. The table also shows the results for continuous measures.

D Exposure to new information about AI

D.1 Full results and Alternative Measures of Attitudinal Outcomes

Table SI-8 reports the full results of the main analysis reported in Table 2 including the controls. The outcome variable in the main analysis is a standardized measure based on a factor analysis of responses to the 8-item matrix asked in the post-treatment survey (Wave 3). Below, we test the results using the alternative measures of the outcome: 1) the first principal component of the eight items asked in the third-wave survey and 2) first factor of the 4 items collected in Wave 2, right after completing the task. As the table shows,

Table (SI-7) Effects of Experience on Behaviors and Alternative measures

	Dependent variable:																			
	Main Task Time				Follow-up Task Time				Clicks count				High Satisfaction				Accept Lower Pay		Correct Classification	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
Algorithmic DM	−0.113** (0.036)	−0.102** (0.033)	−24.158*** (5.534)	−23.256*** (5.205)	−0.098** (0.036)	−0.093** (0.035)	−33.515*** (7.846)	−32.368*** (7.576)	−0.081* (0.036)	−0.082* (0.036)	−1.764** (0.620)	−1.799** (0.615)	−0.149*** (0.028)	−0.146*** (0.028)	−0.218*** (0.063)	−0.206*** (0.062)	−0.064* (0.027)	−0.061* (0.027)	−0.068* (0.031)	−0.071* (0.030)
Algorithmic x Negative exp	0.044 (0.051)	0.050 (0.048)	14.041† (7.966)	14.257† (7.487)	0.070 (0.052)	0.070 (0.050)	28.964* (11.291)	28.441** (10.895)	0.010 (0.051)	0.016 (0.051)	0.387 (0.893)	0.489 (0.884)	0.142*** (0.040)	0.138*** (0.040)	−0.032 (0.091)	−0.050 (0.089)	0.057 (0.038)	0.056 (0.038)	0.103* (0.044)	0.117** (0.043)
Negative experience	−0.114** (0.037)	−0.109** (0.034)	−25.180*** (5.665)	−24.496*** (5.325)	−0.100** (0.037)	−0.093** (0.035)	−31.976*** (8.028)	−30.023*** (7.747)	−0.064† (0.037)	−0.063† (0.036)	−1.674** (0.635)	−1.672** (0.629)	−0.651*** (0.028)	−0.644*** (0.028)	−1.669*** (0.064)	−1.656*** (0.063)	−0.522*** (0.027)	−0.526*** (0.027)	−0.108*** (0.031)	−0.116*** (0.030)
AI info	0.029 (0.026)	0.030 (0.024)	2.664 (3.982)	2.817 (3.745)	−0.004 (0.026)	−0.005 (0.025)	5.492 (5.644)	6.014 (5.449)	−0.010 (0.026)	−0.005 (0.026)	−0.118 (0.446)	0.004 (0.442)	0.035† (0.020)	0.043* (0.020)	−0.018 (0.045)	0.005 (0.045)	−0.038* (0.019)	−0.039* (0.019)	−0.059** (0.022)	−0.068** (0.021)
Negative info	0.037 (0.026)	0.053* (0.024)	6.352 (3.983)	8.687* (3.745)	0.022 (0.026)	0.031 (0.025)	2.034 (5.645)	4.515 (5.449)	0.008 (0.026)	0.015 (0.026)	0.074 (0.447)	0.253 (0.442)	−0.012 (0.020)	−0.011 (0.020)	−0.036 (0.045)	−0.025 (0.045)	−0.047* (0.019)	−0.047* (0.019)	0.048* (0.022)	0.049* (0.021)
Female		0.003 (0.026)		−0.039 (4.052)		−0.038 (0.027)		2.659 (5.900)		0.083** (0.028)		1.174* (0.479)		0.053* (0.022)		0.143** (0.048)		−0.008 (0.021)		−0.042† (0.023)
Age 35-45		0.085** (0.029)		20.612*** (4.584)		0.053† (0.031)		12.390† (6.675)		−0.041 (0.031)		−0.136 (0.541)		0.035 (0.024)		0.087 (0.055)		−0.002 (0.024)		0.017 (0.026)
Age 45+		0.217*** (0.030)		28.367*** (4.703)		0.161*** (0.031)		28.160*** (6.841)		−0.073* (0.032)		−1.158* (0.556)		0.041 (0.025)		0.172** (0.056)		0.012 (0.024)		0.033 (0.027)
White		−0.052† (0.029)		−14.656** (4.510)		−0.087** (0.030)		−26.070*** (6.570)		−0.033 (0.031)		−0.404 (0.533)		−0.031 (0.024)		−0.076 (0.054)		0.004 (0.023)		0.050† (0.026)
Some College or less		0.091*** (0.027)		14.212*** (4.204)		0.010 (0.028)		12.155* (6.121)		0.050† (0.029)		1.046* (0.497)		0.022 (0.022)		0.034 (0.050)		0.007 (0.022)		0.017 (0.024)
Moderate		0.088** (0.033)		3.165 (5.210)		0.018 (0.035)		5.582 (7.581)		0.050 (0.036)		1.020† (0.615)		−0.051† (0.028)		−0.132* (0.062)		0.015 (0.027)		0.057† (0.030)
Liberal		0.039 (0.028)		5.929 (4.340)		0.018 (0.029)		4.383 (6.315)		0.001 (0.030)		−0.034 (0.513)		−0.042† (0.023)		−0.143** (0.052)		0.015 (0.022)		0.180*** (0.025)
Tech Literacy (Medium)		−0.033 (0.029)		−8.783* (4.461)		−0.008 (0.030)		3.275 (6.495)		0.001 (0.030)		−0.602 (0.527)		−0.003 (0.024)		−0.015 (0.053)		0.005 (0.023)		0.011 (0.025)
Tech Literacy (High)		−0.061† (0.036)		−13.790* (5.671)		−0.032 (0.038)		1.279 (8.254)		−0.060 (0.039)		−1.733** (0.670)		0.046 (0.030)		0.036 (0.067)		−0.006 (0.029)		−0.013 (0.032)
Trust in tech		−0.022 (0.027)		−6.509 (4.171)		−0.036 (0.028)		−3.151 (6.069)		−0.029 (0.028)		−0.605 (0.493)		0.023 (0.022)		0.135** (0.050)		−0.020 (0.021)		−0.048* (0.024)
MTurk activity		−0.022 (0.016)		−4.727† (2.549)		−0.028† (0.017)		−4.940 (3.709)		−0.020 (0.017)		−1.004*** (0.301)		−0.053*** (0.014)		−0.161*** (0.030)		0.014 (0.013)		0.061*** (0.015)
Inattentive		−0.278*** (0.026)		−41.790*** (4.106)		−0.230*** (0.027)		−52.575*** (5.975)		−0.034 (0.028)		−0.724 (0.485)		−0.035 (0.022)		−0.122* (0.049)		0.013 (0.021)		−0.078*** (0.023)
Suit Cataloging		0.106† (0.055)		20.395* (8.533)		0.044 (0.057)		6.316 (12.414)		−0.029 (0.058)		−0.433 (1.008)		−0.011 (0.045)		0.085 (0.102)		−0.104* (0.044)		0.046 (0.049)
Task Indifference		−0.070 (0.074)		5.807 (11.598)		0.024 (0.077)		20.910 (16.874)		0.089 (0.079)		1.454 (1.370)		−0.095 (0.062)		0.153 (0.138)		0.057 (0.060)		−0.039 (0.066)
Constant	0.569*** (0.031)	0.606*** (0.064)	144.744*** (4.796)	163.941*** (10.041)	0.574*** (0.031)	0.730*** (0.067)	181.683*** (6.795)	204.427*** (14.611)	0.522*** (0.031)	0.581*** (0.069)	17.466*** (0.538)	20.246*** (1.186)	0.725*** (0.024)	0.831*** (0.053)	4.694*** (0.055)	5.006*** (0.119)	0.585*** (0.023)	0.546*** (0.052)	0.825*** (0.026)	0.601*** (0.057)
Observations	1,499	1,496	1,499	1,496	1,498	1,495	1,498	1,495	1,499	1,496	1,499	1,496	1,500	1,497	1,500	1,497	1,498	1,495	1,499	1,496
R^2	0.020	0.159	0.030	0.159	0.010	0.092	0.019	0.102	0.010	0.036	0.016	0.053	0.368	0.385	0.487	0.513	0.315	0.320	0.017	0.086

Notes: LPM with standard errors in parentheses. Dependent variables in models 3-4, 7-8, 11-12, 15-16 are continuous measures of the behavioral outcomes. The independent variables are indicators for the treatments: algorithmic decision-maker, negative experience, and their interaction. All models control for information treatments. Even-numbered columns control for pre-treatment covariates. †p<0.1; *p<0.05; **p<0.01; ***p<0.001

the results remain consistent when estimating a linear probability model when using these alternative measures of the outcome.

Furthermore, Columns 17-19 presents the results of the effect of exposure to information about AI on support for unrelated policy proposals, such as using background checks for gun purchases and deploying minimal quotas for women on company boards. As the table makes clear, no effect was registered.

Table (SI-8) Effects of Information on Attitudes

	Dependent variable:																		
	FA Score - Wave 3						PCA Score - Wave 3				FA Score - Wave 2						Gun checks	Gender quotas	Affirmative quotas
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
AI X Positive info	0.079***	0.082***	0.082***	0.049**	0.050**	0.050**	0.075***	0.078***	0.048**	0.048**	0.074**	0.076**	0.077**	0.047**	0.049**	0.049**	0.103	−0.026	0.163
	(0.023)	(0.023)	(0.023)	(0.015)	(0.015)	(0.015)	(0.022)	(0.022)	(0.015)	(0.015)	(0.024)	(0.024)	(0.024)	(0.016)	(0.016)	(0.016)	(0.256)	(0.255)	(0.267)
AI Info	−0.036*	−0.039*	−0.037*	−0.012	−0.013	−0.012	−0.032*	−0.034*	−0.010	−0.010	0.004	−0.001	0.0001	0.021†	0.020†	0.020†	−0.263	0.222	0.010
	(0.016)	(0.016)	(0.016)	(0.011)	(0.011)	(0.011)	(0.015)	(0.015)	(0.011)	(0.011)	(0.017)	(0.016)	(0.016)	(0.012)	(0.011)	(0.011)	(0.176)	(0.173)	(0.182)
Positive Info	−0.019	−0.020	−0.020	−0.004	−0.003	−0.003	−0.018	−0.019	−0.003	−0.003	−0.001	−0.001	−0.001	0.008	0.009	0.008	0.011	0.213	0.348†
	(0.016)	(0.016)	(0.016)	(0.011)	(0.011)	(0.011)	(0.015)	(0.015)	(0.011)	(0.011)	(0.016)	(0.016)	(0.016)	(0.011)	(0.011)	(0.011)	(0.176)	(0.173)	(0.182)
Pretreatment FA (8 items)	0.736***	0.714***	0.711***	0.748***	0.714***	0.711***													
	(0.024)	(0.027)	(0.027)	(0.017)	(0.018)	(0.018)													
Pretreatment PCA (8 items)							0.724***	0.699***	0.738***	0.699***									
							(0.024)	(0.027)	(0.017)	(0.018)									
Pretreatment FA (4 items)											0.699***	0.669***	0.668***	0.729***	0.690***	0.689***			
											(0.024)	(0.026)	(0.026)	(0.017)	(0.018)	(0.018)			
Negative experience	0.006	0.008	0.008	0.006	0.007	0.007	0.005	0.007	0.005	0.007	−0.005	−0.004	−0.004	−0.006	−0.005	−0.004	−0.352**	−0.119	−0.095
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.012)	(0.012)	(0.012)	(0.012)	(0.011)	(0.011)	(0.128)	(0.126)	(0.132)
Algorithmic DM				0.009	0.011	0.011			0.008	0.010				0.004	0.006	0.006			
				(0.011)	(0.011)	(0.011)			(0.011)	(0.011)				(0.011)	(0.011)	(0.011)			
Algorithmic X Negative experience				0.0004	−0.001	−0.002			0.001	−0.001				0.0004	−0.001	−0.0005			
				(0.016)	(0.015)	(0.015)			(0.015)	(0.015)				(0.016)	(0.016)	(0.016)			
Female		−0.007	−0.007		−0.005	−0.005		−0.005		−0.003		−0.005	−0.005		−0.007	−0.007	0.266†	0.643***	0.654***
		(0.012)	(0.012)		(0.008)	(0.008)		(0.012)		(0.008)		(0.013)	(0.013)		(0.009)	(0.009)	(0.139)	(0.138)	(0.145)
Age: 35-45		−0.018	−0.017		−0.008	−0.007		−0.016		−0.007		−0.015	−0.015		−0.007	−0.006	−0.190	−0.132	−0.218
		(0.014)	(0.014)		(0.009)	(0.009)		(0.014)		(0.009)		(0.015)	(0.015)		(0.010)	(0.010)	(0.159)	(0.158)	(0.166)
Age: 45+		0.016	0.021		0.005	0.008		0.021		0.008		0.033*	0.033*		0.028**	0.028**	0.059	−0.326*	−0.175
		(0.014)	(0.014)		(0.009)	(0.010)		(0.014)		(0.009)		(0.014)	(0.015)		(0.010)	(0.010)	(0.158)	(0.157)	(0.165)
White		0.003	0.002		0.015	0.015		0.002		0.014		0.018	0.018		0.023*	0.022*	−0.041	−0.003	−0.006
		(0.014)	(0.014)		(0.009)	(0.009)		(0.014)		(0.009)		(0.015)	(0.015)		(0.010)	(0.010)	(0.159)	(0.158)	(0.166)
Low education		−0.027*	−0.027*		−0.016†	−0.015†		−0.026*		−0.014		−0.033*	−0.033*		−0.022*	−0.022*	−0.189	−0.074	−0.145
		(0.013)	(0.013)		(0.009)	(0.009)		(0.012)		(0.009)		(0.013)	(0.013)		(0.009)	(0.009)	(0.142)	(0.141)	(0.148)
Independent		−0.019	−0.017		−0.010	−0.010		−0.016		−0.009		−0.029†	−0.029†		−0.027*	−0.027*	0.583***	0.472**	0.443*
		(0.016)	(0.016)		(0.011)	(0.011)		(0.015)		(0.011)		(0.016)	(0.016)		(0.011)	(0.011)	(0.175)	(0.174)	(0.183)
Republican		−0.030*	−0.028*		−0.025**	−0.024**		−0.026*		−0.022*		−0.032*	−0.032*		−0.031***	−0.031**	1.291***	1.292***	1.400***
		(0.013)	(0.013)		(0.009)	(0.009)		(0.013)		(0.009)		(0.014)	(0.014)		(0.009)	(0.009)	(0.148)	(0.147)	(0.154)
Tech Literacy: medium		0.005	0.006		0.015	0.015†		0.006		0.016†		−0.002	−0.002		0.001	0.001	−0.079	−0.176	0.037
		(0.014)	(0.014)		(0.009)	(0.009)		(0.013)		(0.009)		(0.014)	(0.014)		(0.010)	(0.010)	(0.153)	(0.153)	(0.160)
Tech Literacy: high		0.014	0.012		0.017	0.016		0.012		0.018		0.012	0.011		0.014	0.014	0.290	0.510**	0.564**
		(0.018)	(0.018)		(0.012)	(0.012)		(0.017)		(0.012)		(0.018)	(0.019)		(0.012)	(0.012)	(0.197)	(0.196)	(0.205)
Trust in tech		0.014	0.013		0.028**	0.027**		0.015		0.028**		0.036**	0.036**		0.037***	0.036***	0.198	0.471**	0.378*
		(0.013)	(0.013)		(0.009)	(0.009)		(0.013)		(0.009)		(0.014)	(0.014)		(0.009)	(0.009)	(0.144)	(0.143)	(0.150)
MTurk intensity			−0.011			−0.007		−0.010		−0.007			−0.005			−0.008	−0.046	−0.030	−0.058
			(0.008)			(0.005)		(0.008)		(0.005)			(0.008)			(0.005)	(0.088)	(0.087)	(0.091)
Inattentive			0.014			0.006		0.015		0.006			−0.001			−0.007	−0.203	0.058	0.098
			(0.012)			(0.008)		(0.012)		(0.008)			(0.013)			(0.009)	(0.140)	(0.139)	(0.146)
Constant	0.126***	0.152***	0.171***	0.110***	0.112***	0.126***	0.134***	0.173***	0.117***	0.131***	0.128***	0.140***	0.151***	0.109***	0.114***	0.136***	5.252***	2.653***	2.775***
	(0.015)	(0.026)	(0.032)	(0.012)	(0.019)	(0.023)	(0.015)	(0.032)	(0.012)	(0.022)	(0.016)	(0.027)	(0.033)	(0.013)	(0.019)	(0.023)	(0.340)	(0.338)	(0.354)
Observations	741	741	741	1,500	1,500	1,497	741	741	1,500	1,497	741	741	741	1,498	1,498	1,495	741	741	741
R^2	0.561	0.573	0.574	0.576	0.585	0.584	0.557	0.569	0.571	0.579	0.535	0.557	0.557	0.561	0.578	0.577	0.125	0.160	0.158

Notes: LPM regressions with standard errors in parentheses. The dependent variable is the FA Score of 8 items in Wave 3 (columns 1-6), PCA score of items in Wave 3 (7-10), FA Score of 4 items in Wave 3 (columns 11-16), and Placebo outcomes: gun background checks, gender quotas for boards, and affirmative action for senior positions (17–19). The independent variables are indicators for the treatments: information on AI, Positive tone, as well as their interaction. The models are estimated for the human decision-maker condition (Columns 1-2, and 5-6, 11-13, 17-19), and the full sample (Columns 3-4 and 7-8). All models control for the experience with the DM. †p<0.1; *p<0.05; **p<0.01; ***p<0.001

Table SI-9 reports full results presented in Figure 7 in the main text. Model 1 shows the results presented in the right panel, while models 2-10 show the results by items presented in the left panel of the figure. Models 11-18 show the results when we dichotomized the items into an indicator of support.

Table (SI-9) Full results of Figure 5 and Alternative Measures by individual items

	Dependent variable:																	
	FA score (Fig 5)	Policy domains									Policy domains (binary)							
	(FS)	(FS)	(Patrols)	(Parole)	(Food-stamps)	(Streetlights)	(Enforcement)	(Restraining)	(Visa)	(Shelters)	(Patrols)	(Parole)	(Food-stamps)	(Streetlights)	(Enforcement)	(Restraining)	(Visa)	(Shelters)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
AI Info: positive	0.027† (0.016)	0.042† (0.023)	0.277† (0.154)	0.104 (0.137)	0.332* (0.154)	0.107 (0.146)	0.351* (0.168)	0.185 (0.175)	0.186 (0.178)	0.501** (0.170)	0.085* (0.038)	0.002 (0.026)	0.048 (0.035)	0.056 (0.046)	0.121** (0.041)	0.033 (0.032)	0.027 (0.037)	0.096* (0.043)
AI Info: negative	−0.033* (0.016)	−0.070** (0.022)	−0.152 (0.147)	−0.191 (0.129)	−0.268† (0.145)	0.023 (0.138)	−0.529*** (0.159)	−0.161 (0.166)	−0.299† (0.168)	−0.349* (0.161)	−0.039 (0.036)	−0.028 (0.024)	−0.019 (0.033)	0.036 (0.044)	−0.080* (0.039)	−0.026 (0.031)	−0.046 (0.035)	−0.079† (0.041)
Negative experience	0.017 (0.011)	0.023 (0.016)	0.072 (0.106)	0.100 (0.094)	0.007 (0.106)	0.083 (0.100)	0.081 (0.115)	0.151 (0.121)	0.115 (0.123)	−0.065 (0.117)	−0.012 (0.026)	0.022 (0.018)	0.006 (0.024)	0.011 (0.032)	0.026 (0.029)	0.015 (0.022)	−0.008 (0.025)	−0.033 (0.030)
Tone: Negative	0.007 (0.015)	0.029 (0.022)	0.022 (0.146)	−0.138 (0.129)	0.070 (0.146)	0.010 (0.138)	0.314* (0.159)	0.145 (0.166)	0.201 (0.169)	0.315† (0.161)	0.012 (0.036)	−0.008 (0.024)	0.024 (0.033)	−0.024 (0.044)	0.051 (0.039)	0.036 (0.031)	0.041 (0.035)	0.072† (0.041)
Algorithmic DM	0.006 (0.011)																	
Pretreatment: Patrol			0.635*** (0.030)								0.088*** (0.007)							
Pretreatment: Parole				0.572*** (0.028)								0.058*** (0.005)						
Pretreatment: Foodstamps					0.568*** (0.029)								0.069*** (0.007)					
Pretreatment: Streetlights						0.496*** (0.029)								0.131*** (0.009)				
Pretreatment: Enforcement							0.568*** (0.032)								0.070*** (0.008)			
Pretreatment: Restraining Order								0.461*** (0.036)								0.048*** (0.007)		
Pretreatment: Visa									0.449*** (0.034)								0.044*** (0.007)	
Pretreatment: Shelters										0.416*** (0.031)								0.076*** (0.008)
Feamle	−0.006 (0.012)	−0.005 (0.017)	−0.270* (0.115)	−0.021 (0.102)	−0.019 (0.115)	0.013 (0.109)	−0.075 (0.125)	0.038 (0.131)	−0.010 (0.133)	0.126 (0.127)	−0.034 (0.028)	0.042* (0.019)	0.009 (0.026)	0.064† (0.034)	0.013 (0.031)	0.060* (0.024)	0.032 (0.028)	0.050 (0.032)
Age: 35-45	−0.001 (0.013)	−0.015 (0.020)	−0.067 (0.132)	−0.073 (0.117)	−0.081 (0.132)	−0.027 (0.125)	−0.261† (0.143)	−0.045 (0.150)	−0.134 (0.152)	−0.136 (0.146)	0.006 (0.032)	−0.023 (0.022)	−0.024 (0.030)	−0.029 (0.039)	−0.051 (0.035)	−0.012 (0.028)	−0.032 (0.031)	0.011 (0.037)
Age: 45+	0.003 (0.014)	−0.001 (0.020)	0.245† (0.131)	0.059 (0.117)	0.095 (0.131)	0.304* (0.124)	−0.099 (0.142)	0.038 (0.149)	0.125 (0.152)	−0.017 (0.144)	0.038 (0.032)	0.004 (0.022)	−0.011 (0.030)	0.069† (0.039)	0.019 (0.035)	−0.002 (0.028)	0.015 (0.031)	0.057 (0.037)
White	0.008 (0.013)	0.004 (0.020)	0.235† (0.131)	0.002 (0.117)	0.077 (0.131)	0.164 (0.124)	−0.034 (0.143)	−0.061 (0.150)	−0.153 (0.152)	−0.179 (0.145)	0.028 (0.032)	−0.031 (0.022)	−0.034 (0.030)	0.041 (0.039)	−0.033 (0.035)	0.010 (0.028)	−0.052† (0.031)	−0.070† (0.037)
Low education	−0.048*** (0.012)	−0.059*** (0.018)	−0.275* (0.117)	−0.274** (0.104)	−0.132 (0.117)	−0.094 (0.111)	−0.302* (0.127)	−0.215 (0.134)	−0.236† (0.135)	−0.113 (0.129)	−0.058* (0.029)	−0.033† (0.019)	−0.012 (0.027)	−0.083* (0.035)	−0.067* (0.031)	−0.015 (0.025)	−0.019 (0.028)	−0.033 (0.033)
Independent	−0.024† (0.013)	−0.031† (0.019)	0.125 (0.125)	−0.088 (0.110)	−0.211† (0.124)	0.065 (0.118)	0.003 (0.136)	−0.250† (0.142)	−0.293* (0.144)	−0.057 (0.137)	0.029 (0.030)	−0.008 (0.021)	−0.067* (0.028)	0.039 (0.037)	0.018 (0.034)	−0.050† (0.026)	−0.021 (0.030)	−0.007 (0.035)
Republican	0.039** (0.014)	0.018 (0.020)	0.075 (0.135)	−0.030 (0.120)	0.179 (0.135)	0.020 (0.128)	0.350* (0.147)	0.066 (0.154)	0.020 (0.156)	0.245 (0.149)	0.023 (0.033)	0.052* (0.022)	0.062* (0.031)	0.037 (0.040)	0.079* (0.036)	0.020 (0.029)	0.032 (0.032)	0.066† (0.038)
Tech Literacy: medium	0.044*** (0.013)	0.027 (0.019)	0.059 (0.127)	0.160 (0.112)	0.086 (0.126)	0.008 (0.120)	0.057 (0.137)	0.037 (0.144)	−0.116 (0.146)	0.190 (0.140)	0.074* (0.031)	0.039† (0.021)	0.021 (0.029)	0.050 (0.038)	0.051 (0.034)	0.013 (0.027)	0.007 (0.030)	0.020 (0.035)
Tech Literacy: high	0.084*** (0.017)	0.094*** (0.025)	0.053 (0.164)	0.403** (0.147)	0.419* (0.163)	−0.067 (0.155)	0.129 (0.177)	0.420* (0.187)	0.151 (0.190)	−0.024 (0.180)	0.096* (0.040)	0.088** (0.027)	0.107** (0.037)	0.082† (0.049)	0.112* (0.044)	0.100** (0.035)	0.035 (0.039)	0.009 (0.046)
Trust in tech	0.135*** (0.012)	0.118*** (0.018)	0.149 (0.122)	0.192† (0.108)	0.310* (0.121)	0.295* (0.116)	0.184 (0.132)	0.224 (0.139)	0.327* (0.141)	0.346* (0.135)	0.081** (0.030)	0.031 (0.020)	0.044 (0.028)	0.099** (0.036)	0.085** (0.033)	0.059* (0.026)	0.083** (0.029)	0.116*** (0.034)
MTurk intensity	−0.025*** (0.007)	−0.029** (0.011)	−0.131† (0.072)	−0.107† (0.064)	−0.202** (0.072)	0.029 (0.068)	−0.086 (0.079)	−0.009 (0.082)	−0.065 (0.084)	−0.110 (0.080)	−0.038* (0.018)	−0.040*** (0.012)	−0.012 (0.016)	−0.006 (0.022)	−0.015 (0.019)	0.008 (0.015)	−0.025 (0.017)	−0.012 (0.020)
Inattentive	0.015 (0.012)	0.012 (0.017)	0.033 (0.116)	0.110 (0.103)	0.156 (0.116)	−0.023 (0.109)	−0.007 (0.126)	0.103 (0.132)	0.090 (0.134)	0.182 (0.128)	−0.030 (0.028)	0.029 (0.019)	−0.013 (0.026)	−0.049 (0.035)	−0.034 (0.031)	−0.003 (0.024)	0.012 (0.028)	0.007 (0.032)
Constant	0.379*** (0.029)	0.411*** (0.042)	1.430*** (0.294)	1.277*** (0.255)	1.471*** (0.291)	2.352*** (0.291)	1.846*** (0.324)	1.343*** (0.330)	2.007*** (0.333)	2.362*** (0.329)	−0.140† (0.072)	−0.024 (0.048)	−0.074 (0.066)	−0.264** (0.092)	−0.095 (0.080)	−0.115† (0.061)	0.052 (0.069)	−0.106 (0.083)
Observations	1,497	741	741	741	741	741	741	741	741	741	741	741	741	741	741	741	741	741
R^2	0.159	0.156	0.448	0.466	0.433	0.345	0.370	0.263	0.269	0.259	0.265	0.255	0.203	0.294	0.181	0.136	0.106	0.179

Notes: IV are indicators for positive information on AI, negative information on AI, or placebo information about fashion. †p<0.1; *p<0.05; **p<0.01; ***p<0.001

D.2 A Bounding Exercise

To address selective attrition in our study, we conducted a bounding exercise to assign values to workers who did not complete the post-treatment survey. We employed two approaches for these individuals: Lower Bound: For the conservative measure, we assigned all attrited individuals their pre-treatment outcome, which was collected at wave 1. This approach assumes no change in attitudes for those who dropped out. Upper Bound: For the more permissive measure, we assigned all attrited workers their FA scores measured immediately after receiving the treatment in wave 2.

Table (SI-10) Estimating bounds of the treatment effect

	Dependent variable:							
	Lower bound FA score pre-treatment outcome (wave 1)				Upper bound FA score posttreatment outcome (wave 2)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI X Positive info	0.041**	0.043**	0.069**	0.072**	0.053**	0.054**	0.067**	0.070**
	(0.013)	(0.013)	(0.019)	(0.019)	(0.014)	(0.014)	(0.020)	(0.020)
Info about AI (ref: Fashion)	−0.010	−0.011	−0.033*	−0.035*	−0.011	−0.011	−0.027	−0.028*
	(0.009)	(0.009)	(0.013)	(0.013)	(0.010)	(0.010)	(0.014)	(0.014)
Positive info (ref: Negative)	−0.004	−0.005	−0.018	−0.019	−0.001	−0.002	−0.011	−0.012
	(0.009)	(0.009)	(0.013)	(0.013)	(0.010)	(0.010)	(0.014)	(0.014)
Pre-dispositions (PCA wave 1)	0.787**	0.760**	0.770**	0.754**	0.750**	0.721**	0.743**	0.726**
	(0.014)	(0.016)	(0.021)	(0.023)	(0.015)	(0.017)	(0.022)	(0.024)
Algorithm DM	0.008	0.009			0.007	0.009		
	(0.009)	(0.009)			(0.010)	(0.010)		
Experience Positive	−0.001	−0.002			0.004	0.003		
	(0.013)	(0.013)			(0.014)	(0.014)		
Algorithm X Positive Exp	0.005	0.006	0.006	0.007	0.004	0.005	0.005	0.006
	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
Female		−0.006		−0.009		−0.002		0.001
		(0.007)		(0.011)		(0.008)		(0.011)
35-45		−0.005		−0.012		−0.005		−0.014
		(0.008)		(0.012)		(0.009)		(0.013)
45+		0.011		0.020		0.011		0.017
		(0.008)		(0.012)		(0.009)		(0.013)
Race/Ethnicity (White)		0.012		0.002		0.013		0.006
		(0.008)		(0.012)		(0.008)		(0.012)
Education Level (Low)		−0.011		−0.020		−0.009		−0.021
		(0.007)		(0.011)		(0.008)		(0.011)
Political Orientation: Moderate		−0.007		−0.015		−0.012		−0.008
		(0.009)		(0.013)		(0.010)		(0.014)
Political Orientation: Liberal		−0.022**		−0.025*		−0.029**		−0.026*
		(0.008)		(0.011)		(0.008)		(0.012)
Technology Literacy (Medium)		0.008		0.004		0.014		0.009
		(0.008)		(0.012)		(0.008)		(0.012)
Technology Literacy (High)		0.009		0.006		0.014		0.016
		(0.010)		(0.015)		(0.011)		(0.016)
Trust in Technology		0.020**		0.010		0.021*		0.010
		(0.008)		(0.011)		(0.008)		(0.012)
MTurk Activity		−0.006		−0.008		−0.007		−0.009
		(0.004)		(0.007)		(0.005)		(0.007)
Inattentive		0.008		0.014		0.004		0.015
		(0.007)		(0.010)		(0.008)		(0.011)
Constant	0.092**	0.107**	0.109**	0.145**	0.107**	0.124**	0.119**	0.143**
	(0.010)	(0.019)	(0.013)	(0.027)	(0.011)	(0.021)	(0.014)	(0.029)
Model	Minimal	Socio-demog	Mturk HITs	Minimal	Socio-demog	Mturk HITs		
Observations	1,796	1,793	885	885	1,796	1,793	885	885
R^2	0.628	0.633	0.608	0.617	0.572	0.579	0.565	0.574
Model	Minimal	Socio-demog	Minimal	Socio-demog	Minimal	Socio-demog	Minimal	Socio-demog

D.3 Effects of Information Treatments by Predispositions

Table (SI-11) Effects of Information Treatments by Predispositions

	Dependent variable:			
	(1)	(2)	(3)	(4)
AI x Positive Info	0.066^{*}	0.055^{*}	0.067^{**}	0.071^{*}
	(0.027)	(0.028)	(0.026)	(0.028)
AI x Positive Info x Averse	−0.014	−0.003	0.002	−0.003
	(0.037)	(0.037)	(0.038)	(0.040)
AI x Averse	0.024	0.013	0.022	0.025
	(0.026)	(0.026)	(0.027)	(0.028)
Positive Info x Averse	0.008	−0.004	−0.006	0.008
	(0.026)	(0.026)	(0.027)	(0.028)
Positive info	−0.014	−0.002	−0.014	−0.019
	(0.018)	(0.019)	(0.018)	(0.019)
AI Info	$-0.035^{\text{Ď}}$	−0.025	$-0.035^{\text{Ď}}$	$-0.039^{\text{Ď}}$
	(0.019)	(0.020)	(0.018)	(0.020)
Pretreatment outcome	-0.293^{***}	-0.239^{***}	-0.258^{***}	-0.325^{***}
	(0.019)	(0.019)	(0.019)	(0.020)
Algorithmic DM	0.010	0.010	0.010	0.014
	(0.009)	(0.009)	(0.010)	(0.010)
Negative Experience	0.009	$0.015^{\text{Ď}}$	0.007	0.010
	(0.009)	(0.009)	(0.010)	(0.010)
Constant	0.554^{***}	0.497^{***}	0.555^{***}	0.569^{***}
	(0.014)	(0.015)	(0.014)	(0.015)
Observations	1,500	1,351	1,351	1,350
R^2	0.383	0.334	0.344	0.427
Sample	Fill	Excluding top 10%	Excluding bottom 10%	Excluding midpoint (45-55)

Notes: OLS regressions with se in parentheses. The DV is the FA score of 8 items in Wave 3. The IV variables are indicators for the treatments: info about AI (Fashion as ref), positive tone (negative tone as ref), an indicator for predisposition (relative to the median measures) and their interaction.

†p<0.1; *p<0.05; **p<0.01; ***p<0.001

E Research Ethics

This study is a field experiment conducted on the online labor marketplace, MTurk. The experiment design, the treatments and the survey instruments were all reviewed and approved by the Institutional Review Board (IRB) before the study was initiated.

The study was conducted with adherence to the current standards for research transparency and ethics, including the American Political Science Association's "Principles and Guidance for Human Subjects Research," which were approved by the APSA Council in April 2020.

At the beginning of the pre-treatment and the post-treatment surveys (waves 1 and 2), each participant was provided a consent form which informed them that participation in the study was voluntary and that they could withdraw at any time without penalty. No identifying data, such as names or email addresses, was collected, ensuring the anonymity of the data used for analysis and replication.

After the study was complete, all participants were sent a debriefing letter to their MTurk interface. The letter explains the study and the treatment they received. The wording of the letter is provided below.

Debriefing Letter

We are getting back to you following your participation in our research, in which you performed the following HITs: - A short survey about views on social issues - A short task of placing predictions on a scale ranging from very negative to very positive - A short survey about views on social issues.

Thank you for your participation in our study! It is greatly appreciated.

In addition to the task, you may recall that we also asked you quite a few questions about your views on the use of algorithmic decision-making in different policy contexts. As the questions in the surveys you completed made clear, public agencies are increasingly relying on algorithmic systems to make important decisions. These systems seem to be expanding into more and more areas of our lives. The actual purpose of this academic study was to learn how experience with these algorithmic systems and exposure to new information about this technology and its potential implications shape views on AI usage in public policy.

As part of this study, you were informed that you were assigned to the task by an algorithm/our team, which found you suitable to perform this specific task. We would like to let you know that the task assignment was randomly assigned, as the academic focus was to learn about the way the identity of the decision-maker - algorithm vs. a human being - affects participants' views on these technologies and their incorporation in a range of public policies. Since this was a random assignment, no significance should be attached to the degree of suitability found for you for the specific task you were assigned.

If you have any questions regarding this study, its purpose, or procedures, or if you would like to receive a copy of the final report of this study when it is completed, please feel free to contact us at academic.research.tasks@gmail.com.

F Survey Instruments

Wording of Pre-treatment Survey Items

Definitions: Before beginning, please read the following definitions that are relevant to the survey:

- The pretrial stage in the criminal justice system is the period between arrest and trial.
- A predictive algorithm is computer software that makes decisions without human instruction, relying on massive amounts of data.
- Homelessness is defined as living somewhere that is below a minimum quality standard or that you can be evicted from with little or no warning.

Screener: Please select the definition that did not appear among the previous three definitions.

- The pretrial stage in the criminal justice system is the period between arrest and trial. (1)
- A predictive algorithm is computer software that makes decisions without human instruction, relying on massive amounts of data. (2)
- Screeners are workers in child welfare who respond to the hotline calls reporting child abuse allegations. (3)
- Homelessness is defined as living somewhere that is below a minimum quality standard or that you can be evicted from with little or no warning. (4)

Policy Decisions: Please indicate the extent to which you support or oppose having each policy decision made by a predictive algorithm rather than by a human being.

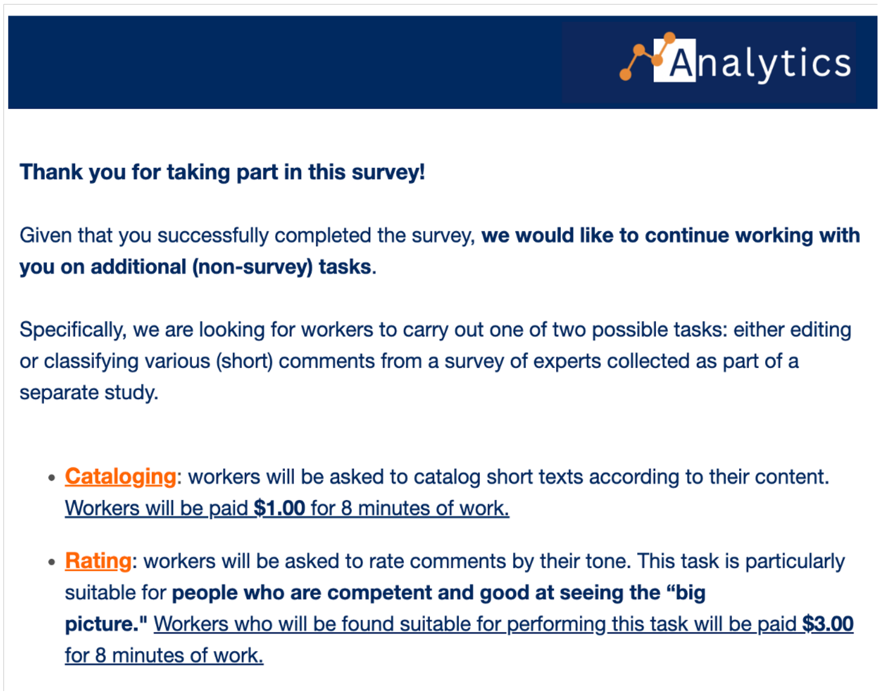
- (`Parole`) Deciding whether to grant parole.
- (`Food stamps`) Deciding which individuals should receive food stamps.
- (`Patrols`) Deciding where police forces should patrol.
- (`street lighting`) Deciding where to place street lighting.

Please indicate the extent to which you support or oppose having each policy decision made by an algorithm rather than by a human being.

- (`Restraining order`) Determining whether a restraining order should be issued
- (`Visa applications`) Deciding whether to approve immigrant visa applications

- (`Enforcement`) Deciding where to increase police enforcement of illegal construction.
- (`Homeless shelters`) Deciding where to build shelters for homeless people
- (`Screenert`) Deciding between options. Please tick the answer '5'.
- (`Invitation 1`: If you had to choose between these two tasks, which one would you prefer to perform? Cataloging - $1.00 for 8 minutes of work. (1) Rating - $3.00 for 8 minutes of work. (2) If
- (`Invitation 2`: In addition, please indicate which task you think suits you better. Cataloging - $1.00 for 8 minutes of work. (1) Rating - $3.00 for 8 minutes of work. (2)
- (`Invitation 3`: If you had to choose between these two tasks, which one would you prefer to perform? Cataloging - $1.00 for 8 minutes of work. (1) Rating - $3.00 for 8 minutes of work. (2)

Figure (SI-3) Screen capture of the invitation to perform additional tasks



Thank you for taking part in this survey!

Given that you successfully completed the survey, **we would like to continue working with you on additional (non-survey) tasks**.

Specifically, we are looking for workers to carry out one of two possible tasks: either editing or classifying various (short) comments from a survey of experts collected as part of a separate study.

- **Cataloging**: workers will be asked to catalog short texts according to their content. Workers will be paid **$1.00** for 8 minutes of work.
- **Rating**: workers will be asked to rate comments by their tone. This task is particularly suitable for **people who are competent and good at seeing the "big picture."** Workers who will be found suitable for performing this task will be paid **$3.00** for 8 minutes of work.

F.1 Wave 2

Table (SI-12) Negative and Positive Predictions on AI

	AI Negative	**AI Positive**
1	I think that using algorithms that rely on aggregate data to make decisions about individuals may lead to many errors, especially in cases where someone doesn't fit the typical profile.	I believe artificial intelligence will make more reliable decisions than humans, whose decisions are often influenced by irrelevant factors, such as their mood.
2	The lack of transparency in the way algorithmic decision-making systems are being used may generate frustration among those affected by the decisions.	Artificial intelligence relies on massive amounts of data to make predictions. This can lead to a high degree of accuracy.
3	Because algorithms learn by analyzing historical data, sources of inequality from the past will also be programmed into future outcomes.	Artificial intelligence (AI) can improve workplace safety. AI doesn't get stressed, tired, or sick—three major causes of human accidents in the workplace.
4	Many of these AI systems are secret. In Wisconsin, for example, the algorithm was developed by a private company and has never been publicly disclosed because it is considered proprietary.	Artificial intelligence might lead to more consistent judgments than those made by humans, who may be influenced by emotional considerations or by fatigue.
5	AI may purposely exclude all references to race and ethnicity, but these systems still consider factors that correlate with race, such as low-income neighborhoods or employment history. As a result, the algorithm's outputs can be racially discriminatory.	Artificial intelligence helps humans make more rational choices based on evidence and accumulated knowledge.
6	Despite perceptions that algorithms are somehow neutral and uniquely objective, they can often reproduce and amplify existing prejudices.	By reducing the need for human discretion, algorithms may help deploy government resources in a more objective manner.
7	Using AI to make important decisions without human oversight can make it difficult to determine who is responsible for the outcomes.	AI can help policymakers identify patterns and trends that may not be immediately obvious, leading to more effective policies.
8 (opposite)	Artificial intelligence relies on massive amounts of data to make predictions. This can lead to a high degree of accuracy.	The lack of transparency in the way algorithmic decision-making systems are being used may generate frustration among those affected by the decisions.

Table (SI-13) Negative and Positive Predictions on Fashion

	Fashion Negative	**Fashion Positive**
1	Similar to the trend in fast food, I expect people to continue consuming lower-quality, inexpensive clothing rather than a few higher-quality and costlier pieces. These consumption habits will have very adverse effects on the environment.	Fashion brands are expected to move to more sustainable fabrics and manufacturing methods, which means garments will be more eco-friendly and longer-lasting.
2	Most fashion brands are not expected to move to more sustainable fabrics and manufacturing methods, which means garments will be less eco-friendly than they should be.	Similar to the trend in organic food, I expect people to start consuming a few higher-quality yet expensive pieces rather than many lower-quality pieces. These consumption habits will have very beneficial effects on the environment.
3	In the next decade; we will see less of an emphasis on comfort in clothing design, which will likely cause individuals, especially women, to feel less comfortable with their appearance.	In the next decade, we will see a greater emphasis on comfort in the design of clothing, which will cause individuals, especially women, to feel more comfortable with their appearance.
4	Instead of manufacturing new products, more and more brands are investing in new resale business models and offering second-hand goods. This may lead to a drop in prices for the consumer.	Fashion will become more inclusive and diverse, with more options for different body types and sizes.
5	Fashion will become more inclusive and diverse, with more options for different body types and sizes.	As more people work from home, fashion brands are adapting their styles to better suit the home environment.
6	As more people work from home, fashion brands are adapting their styles to better suit the home environment.	I expect to see more emphasis on earth colors, especially in men's suits, but also in womenswear.
7	I expect to see more emphasis on earth colors, especially in men's suits, but also in womenswear.	Instead of manufacturing new products, more and more brands are investing in new resale business models and offering second-hand goods. This may lead to a drop in prices for the consumer.
8 (opposite)	Instead of manufacturing new products, more and more brands are investing in new resale business models and offering second-hand goods. This may lead to a drop in prices for the consumer.	Fashion consumerism, especially among teenagers, will continue to grow in the coming decade and will contribute significantly to environmental degradation.

F.2 Wave 3

- (Post treatment outcomes With respect to each policy decision in the list below, please indicate the extent to which you support or oppose having that decision made by a predictive algorithm rather than by a human being.
- (Parole) Deciding whether to grant parole.
- (Food stamps) Deciding which individuals should receive food stamps.
- (Patrols) Deciding where police forces should patrol.
- (street lighting) Deciding where to place street lighting.

Please indicate the extent to which you support or oppose having each policy decision made by an algorithm rather than by a human being.

- (Restraining order) Determining whether a restraining order should be issued
- (Visa applications) Deciding whether to approve immigrant visa applications
- (Enforcement) Deciding whether to increase police enforcement
- (Homeless shelters) Deciding where to build shelters for homeless people
- (Manipulation check 1 (asked at the very end of the wave 3 survey: Thank you for your participation in our study! It is very much appreciated. A few days ago, we invited you to carry out one of two possible tasks: cataloging task for $1.00 or rating task for $3.00 ($1 + $2 bonus). Please indicate which task you were assigned to perform: (1) The rating task for $3.00 (2) The cataloging task for $1.00.
- (Manipulation check 2 By whom were you assigned to this task: (1) A predictive algorithm (2) Daniel, a member of the HR team (3) Shir, a member of the HR team (4) Danielle, a member of the HR team
- (Manipulation check 2 (follow up, if "a predicted algorithm" was selected) Which algorithm was used to assign you to the task: (1) The regular M-Turk algorithm (2) A specific algorithm used by the HIT requester.