The Politics of Using AI in Public Policy: Experimental Evidence

Yotam Margalit*and Shir Raviv†

September 19, 2023

Abstract

The use by government agencies of AI in guiding important policy decisions (e.g., on policing, welfare, education) has generated backlash and led to calls for greater public input in AI regulation. But what does the public's input on this topic entail? Does personal experience with the technology or learning about its implications change people's views on using AI for guiding policy implementation? We study these questions experimentally in an online labor marketplace. We track the attitudes of over 1,500 workers, where the boss who allocates them to tasks, 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' behavior on the job, but not their policy attitudes. In contrast, exposure to information about the technology generated significant attitudinal change. The findings provide insight into the formation of attitudes on AI regulation and the potential politicization of this issue.

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

†Data Science Institute, Columbia University.

The pre-analysis plan is available at: https://aspredicted.org/blind.php?x=XNX$_K$$CB$.

The online appendix is available at: https://www.dropbox.com/scl/fi/s7ktu1loo1a5k4bl2vior/appendix.pdf

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; O'Kane, 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; Sheehan and Pittman, 2016). Notably, in some of these cases the AI-based technology was deemed to offer a significant improvement over prior methods,

[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 (Lorinc, 2022; O'Kane, 2022).

but it was nonetheless withdrawn in response to public opposition.

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 implementation decisions across an array of public 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 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

[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).

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 in response to relevant information on AI and its societal implications.

For example, those who learned about the positive aspects of AI grew more favorable towards the use of the technology in policy implementation decisions, particularly in policy decisions about allocation of public resources, such as food stamps for the poor, shelters for the homeless, or police patrols. In contrast, learning about AI's potential drawbacks increased opposition to the use of AI in specific domains, especially in criminal justice. Interestingly, workers who were initially skeptical of AI updated their views more strongly in response to positive information than those who initially supported AI. 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 from our study 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 decisions. The reasons for this 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 research on AI ethics that examines the trade-offs involved in using the technology in the public sector. On the one hand, AI systems may improve the performance and fairness of public services by reducing human errors and biases or making it easier to expose them systematically (Kahneman, Sibony, and Sunstein, 2021; Sunstein, 2022; Kleinberg et al., 2018). On the other hand, such systems also raise ethical concerns about their opacity, accountability, and the risk of perpetuating inequalities and biases inherent in historical data (Burrell, 2016; Lepri et al., 2018). Notably, much of this research implicitly assumes that, without knowledge about these trade-offs, people view algorithms as an attractive solution, reflecting a broader tendency to identify mathematics

as objective and accurate (e.g., O'neil, 2016; Pasquale, 2020).[3] Yet this assumption remains inadequately tested, with some scholars questioning its veracity (e.g., Starke et al., 2022). We provide novel evidence on this question, showing that people are not fixed in their predispositions and do update their views when encountering information about these trade-offs.

Our study also contributes to the growing literature on the determinants of public opinion regarding the use of algorithmic decision systems 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

[3]For example, as Eubanks (2018) stated, "one of the great benefits of these tools for governments is it allows them to portray the decisions they are making as neutral and objective..." Noble (2018) suggested that "The reason why thinking that predicting technology, risk assessment score is more fair, is that people believe that algorithms and math are unbiased and objective and fair. So there's a very easy logic to understand why the public would get behind this, right?" Similarly, O'neil (2016) noted that: "algorithms are opinions embedded in code. It's really different from what most people think of algorithms. They think algorithms are objective and true and scientific."

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

To understand the public's views on the use of AI in implementing important public policy decisions, one might look for insights from earlier research on attitude formation and the public adoption of other emerging technologies. But as we describe below the insights one could draw from the literature are less than conclusive.

One strand of research emphasizes a cognitive process of learning and holds that people's attitudes often shift as they acquire more knowledge about new technology (e.g., Yeomans et al., 2019). Examining public acceptance of energy technologies or biotechnology, studies suggest that information and technological literacy are key to the way people weigh the costs, risks, and benefits of new technologies, and thus greater knowledge can lead to a change in attitudes (Stoutenborough and Vedlitz, 2016).

This conjecture seems particularly relevant at this early stage of the public debate over AI regulation, i.e., when most people still know little about AI and there are no widely accepted elite positions that can cue public opinion on the matter (Stamm, Clark, and Eblacas, 2000; Cobb and Macoubrie, 2004). 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.[4]

Another strand of research underscores the affective dimension, and contends that information alone about a new technology is rarely sufficient to lead to attitudes change. Instead, people need to have also a motivation to process the information (Scheufele and Lewenstein,

[4]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).

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 nor to reason about its various uses. Moreover, the complexity and novelty of the technology may render it difficult to comprehend. If that is the case, information about AI may have little bearing on how people view the technology and its implications.

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). In fact, 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.[5] 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.

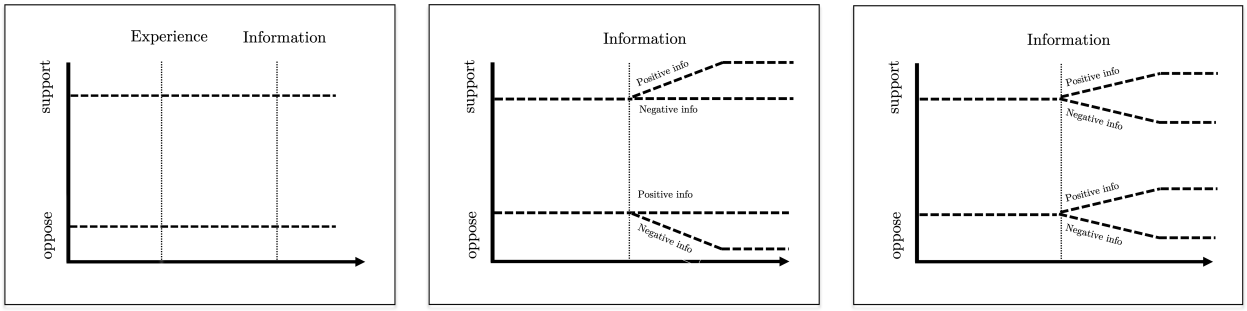
To illustrate the differences between these theoretical approaches, suppose two people watch a news segment of experts discussing the ProPublica report that revealed racial bias in COMPAS, the risk assessment algorithm discussed earlier. Imagine one person is initially more favorable towards AI, while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

[5]For example, Druckman and Bolsen (2011) examine how framing and information shape public views on carbon-nanotubes and genetically modified foods. They find that providing factual information does little to enhance public knowledge or support for these technologies. Rather, people tend to process new information in a biased manner that confirms their prior beliefs, especially once they have formed clear opinions on the issue.

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 abou AI and its societal implications. The vertical axis indicates the probability of favoring AI-based algorithms in the implementation of policy decisions.

The vertical axis of each graph represents the probability of supporting the use of algorithms rather than humans in making high-stakes decisions in the public sector. The left-most panel (Figure 1(a)) shows the attitude of both individuals remaining stable, irrespective of the information they encounter. In contrast, the middle panel 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 pior opinions while ignoring the rest. Finally, the right panel shows support for AI-use change in the direction implied by the new information, irrespective of people's initial stance. For instance, learning about the biased outcomes of the COMPAS algorithm would make them both more skeptical about using AI technology 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. Indeed, in recent years, people are increasingly exposed to algorithmic systems in their daily lives, whether it is a chatbot that answers questions, a recommendation system that suggests what to watch or buy, or a bank's credit score system that determines eligibility for a loan. These interactions may affect how people think about AI.

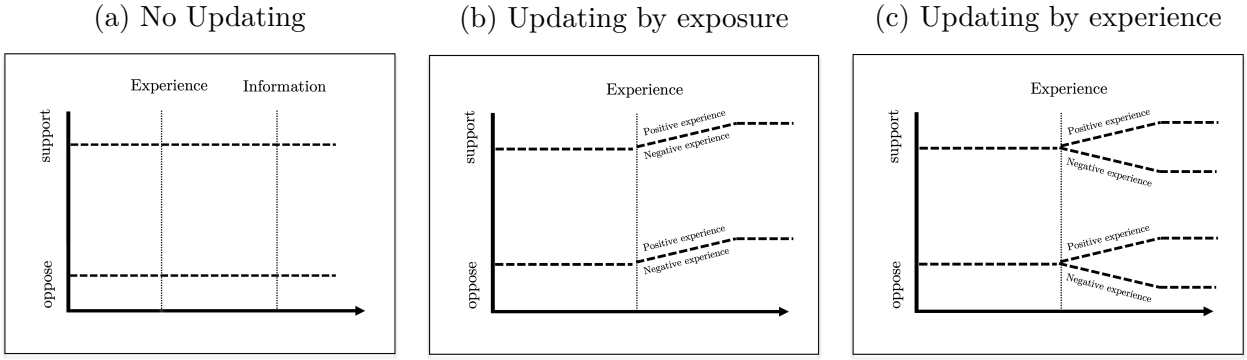
Specifically, direct experience with algorithmic decision-making may foster a sense of fa-

miliarity 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 and its capabilities, subsequently increasing their support for the use of AI applications in other domains. Indeed, earlier experimental studies have shown a similar response to interaction with other advanced technologies (e.g., semi-autonomous cars) (Austin, Stevenson, and Wei-Skillern, 2006; Lapinsky et al., 2008).

Alternatively, the nature of the experience, i.e., whether positive or negative, plays a more prominent role in shaping opinions. A negative experience could be, for example, receiving consistently faulty information from a ChatGPT-like system or being denied a loan request by a bank's AI-based decision-making system. By this view, personal experience provides a vivid and salient heuristic, one that is more accessible than other sources of information. If this is the case, people's attitudes toward the incorporation of AI-based tools in public policy could very much be a function of their satisfaction with the algorithm-based decisions 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' political attitudes and 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).

Figure (2) Experiencing algorithmic decision making - Trajectories of Preferences



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

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 algorithms 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).

While theoretically intuitive, we know little about the way personal experience with AI influences preferences toward the broader question of the desirability of employing the technology in public policy decisions. Specifically, we are interested in whether people generalize from their own experiences to the policy realm or whether they treat these experiences as less relevant for the policy domain. While plausible, 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.

Returning to the example of the two individuals who have different initial opinions on the use of AI-based algorithms in public policy decisions. How would a personal experience with algorithmic decision-making affect their views? Figure 2 illustrates possible paths of attitude change that may follow as a result of such an experience. 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 review of existing research reveals that 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

We begin this section with a high-level overview of the experimental design. We then delve into a detailed description of each of the experiment's components.

Given that people increasingly encounter AI-based applications in carrying out various labor market functions (e.g., employee recruitment, task allocation, or quality assessment), we chose as our experimental setting Amazon's Mechanical Turk (MTurk) platform. MTurk is the world's largest online labor market, providing employers ("requesters") with access to a large base of potential employees that are hired to perform a range of discrete on-demand tasks. This setting is similar to that offered in other general-purpose labor markets online

platforms such as Upwork, Fiverr, and Guru. 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). In fact, evidence suggests that findings from MTurk experiments are comparable to those from more traditional platforms (Horton, Rand, and Zeckhauser, 2011).

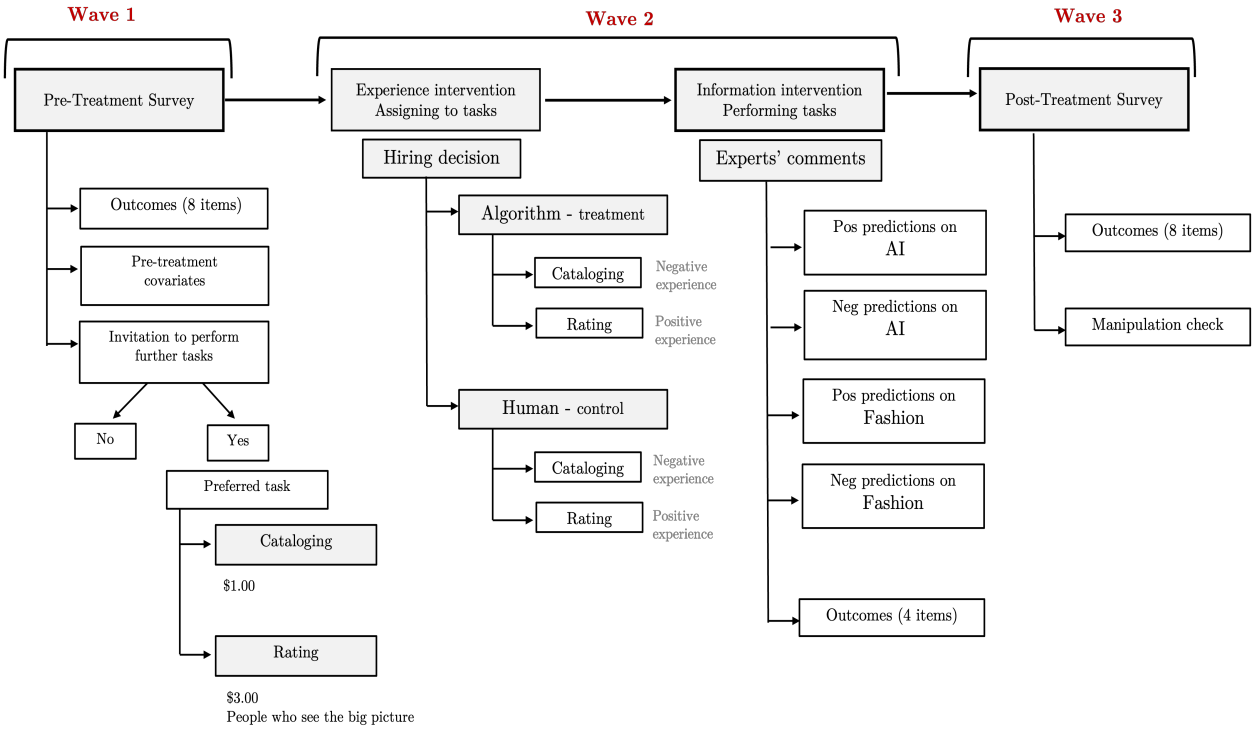
To assess the impact of exposure to AI-guided decision-making, we invited 2375 American workers to perform paid tasks that we initiated as employers on the platform. The experiment consisted of a factorial design of three treatments: the identity of the boss assigning the workers to the tasks (human or an AI algorithm), the nature of the experience (good or bad), and the informational content of the task. All three treatments, which we describe in detail below, were randomly assigned.

To document participants' views on relevant issues and track attitudinal changes over time, the experiment included a three-wave panel survey, one of which was administered before the participants completed the task, and the other two were fielded after the task's completion.

To evaluate the impact of personal experience with algorithmic decision-making on 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 algorithmic decision-making, we provided participants with the option of signing up to perform either a high-paying, higher-status task or to a low-paying, lower-status task. As expected, they almost uniformly requested to perform the first, higher-paying task. The second intervention then consisted of whether they were selected (by the algorithm or the HR person) to their preferred task or not.[6] Finally, to assess the impact of exposure to new information on the updating of

[6] As we pre-registered, we excluded the small number of participants who requested the lower-status task, resulting in two treatment groups: those with a positive experience assigned to the high-paying-status desired task and those with a negative experience assigned to the low-paying-status undesired task.

Figure (3) Experimental Design



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 sequence of the experiment is captured in Figure 3. In this section, we discuss the rationale for and procedure of each step in more detail.

3.1.1 Pre-Treatment Survey

In February 2023 we recruited workers for our study via Amazon's Mechanical Turk platform. Participants were invited to participate in a "short survey on social issues" and offered a

payment of $0.8. The baseline survey included several pre-treatment outcomes. Specifically, we asked respondents to indicate the extent to which they support or oppose the use of a predictive algorithm instead of a human to make determinations in various policy contexts. To minimize the possibility of demand effects influencing how participants answered the questions, we added to the survey a host of unrelated items with the aim of blurring the focus of the study. We also collected information relevant as pre-treatment covariates, such as age, race, ideology, education, technological literacy, and trust in institutions. For the exact wording of the survey items, see Appendix A.

Invitation to Perform Additional Tasks

At the end of the survey participants received an invitation to continue work with the same employer ("requester" in M-Turk parlance) 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 between the two tasks.[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.[8] This allows us later to compare the work performance of participants in the different treatment groups. 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 more positive or more negative experience with the employer's decision regarding the allocation of work.

[7] To account for cases where participants had no meaningful preference between the two tasks, we offered a third option, "don't care." Indifferent participants who chose this option were then forced to choose between the two tasks. We used this indication of indifference as a control in our analysis (see Appendix C).

[8] We wrote the description of the two tasks in a way that ultimately described well the actual work the participants were asked to perform.

Figure SI-1 in the appendix shows a screen capture of the invitation, with the exact wording of the questions. As the Figure shows, we designed a distinct interface with the Analytics logo that was used for both Waves 1 and 2. In doing so, our intention was twofold: (1) to enhance the sense of an employer-worker setting; (2) to help differentiate the first two waves from the third wave, to which the participants were invited by a different employer and which used a different user interface. This was done to minimize the chance of participants connecting between the surveys and potentially being strategic in the way they answer the post-treatment survey.[9]

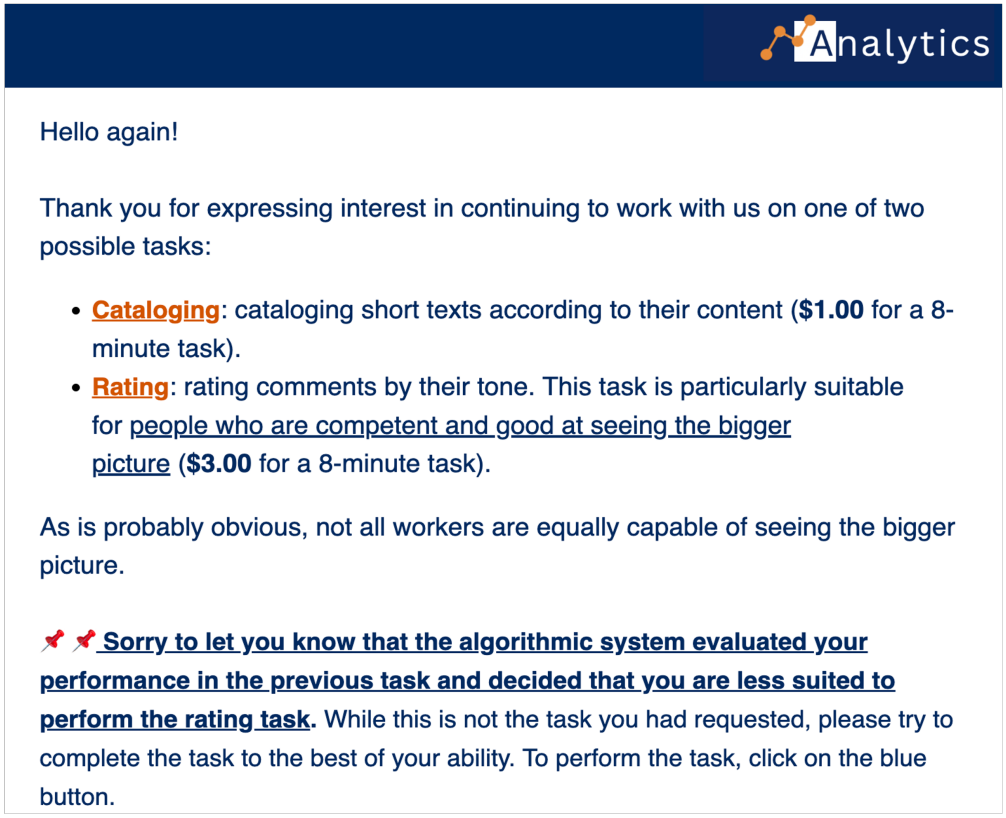
3.1.2 Experience with Algorithmic Decision-making

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 in this initial message, informing them that they would 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

[9] 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.

Figure (4) Screen capture of the reply message: a negative experience with algorithmic decision maker



negative experience treatment were told that the decision-maker evaluated their performance in the previous task and deemed them unsuited for the rating task. 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.[10]

To assess the effects of working under human or algorithmic decision-making, 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.[11] Additionally, we included in the pre-treatment

[10] 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.

[11] Specifically, to avoid gender-based bias we alternated across respondents the member's name between "Danielle" and "Daniel".

survey questions 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.

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 form of 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 algorithmic decision-maker, allowing the participants to express dissatisfaction with the decision. Of the total sample of 1864 participants, 36% opted to share their feelings. Perhaps unsurprisingly, a large majority of those (76%) were respondents who had a negative experience and expressed disappointment or frustration at being denied the higher-paying option.[12]

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. Specifically, workers criticized the algorithm's decision, saying, 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 decision-maker 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

[12]In the open-ended texts, one participant wrote, "I feel that I put 100% effort into all these HITS; therefore, I feel 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."

is clueless about me. I know myself better than anyone."

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 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 four-to-seven 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, 1864 individuals completed the wave 2 survey, and 1541 completed both post-treatment surveys (i.e. waves 2 and 3). As noted earlier, 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 in the Appendix presents descriptive statistics on pretreatment values for a range of 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.

4.2 Attrition

Table SI-1 in the Appendix reports attrition and completion rates for waves 2 and 3 by treatment assignment. The data covers all participants who received an invitation for wave 2, had opened the link, and were subsequently exposed to the treatments. As expected, the table shows differences in the completion rates of wave 2 based on the type of experience (positive vs. negative) but not by the identity of the decision maker (human vs. algorithm). Participants assigned to the less attractive task had a higher dropout rate of 7-8%, compared to the 3-4% attrition rate for participants with a positive experience ($p = 0.012$). This finding remains consistent across both human and algorithmic decision-makers.

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 (depending on the group), a high rate compared to previous research that utilized in MTurk panels (e.g., Christenson and Glick, 2013).

4.3 Outcome Variables

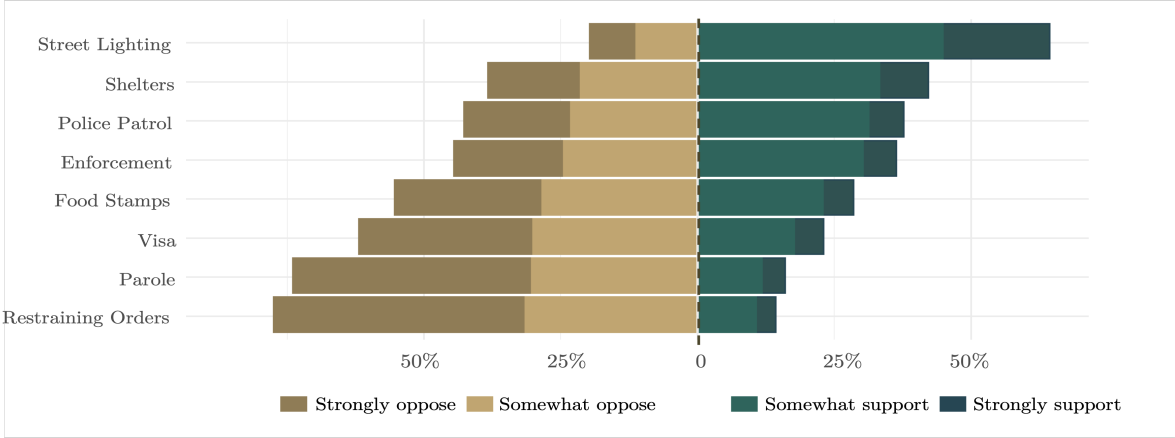
Our primary dependent variable examined individuals' attitudes toward reliance on algorithmic decision-making systems in the implementation of public policy. Specifically, we asked respondents to indicate their support or opposition to using predictive algorithms instead of human decision-makers in a 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 of police enforcement, 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.[13] 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).[14]

[13] Our results remain similar when using instead PCA (principal component analysis) and when removing each time one of the eight items from the calculation and computing the first principal component for the remaining items. See Appendix C

[14] 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. By organizing the questions in this manner, we strive to eliminate potential bias in participant evaluations and ensure consistency in the outcomes measured across all 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



The figure presents 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). The distribution calculation takes into account the indifferent category. Table SI-3 reports the full distribution.

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 Americans 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 stil with a small proportion expressing strong support.

5.2 Effects of Experience on Attitudes

Our experiment is designed to isolate the causal effect of personal experience with AI algorithms and of information about the technology on people's attitudes regarding the use of AI for making policy implementation decisions.

With regard to the first causal effect, we estimate the average treatment effect of exposure to algorithmic decision-making on AI-related attitudes, as measured by the post-treatment survey conducted several days after the assignment. Our outcome variable is a standardized measure based on a factor analysis of responses to questions about reliance on AI for implementation decisions in various policy settings, where higher values indicate greater support for using AI. 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 taking account of 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 algorithmic decision-making did not affect workers' attitudes toward the use of AI in public policy. Across all specifications, the coefficient of algorithmic decision-maker is consistently very small and below statistical significance, ranging from 0.002 (t=0.385) to 0.011 (t=1.46).

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

Table (1) Effects of Experience on Attitudes

	Dependent variable:								
	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	0.002	0.010	0.011	0.018	−0.002	0.009	0.011	0.022	0.025
	(0.011)	(0.008)	(0.008)	(0.013)	(0.015)	(0.011)	(0.011)	(0.018)	(0.018)
Negative experience	−0.001	0.007	0.008	0.007	−0.005	0.006	0.008	0.011	0.014
	(0.011)	(0.008)	(0.008)	(0.008)	(0.015)	(0.015)	(0.011)	(0.016)	(0.016)
Algorithm X Negative experience					0.008	0.001	0.000	−0.006	−0.010
					(0.021)	(0.015)	(0.015)	(0.015)	(0.024)
Pretreatment outcome	0.778**	0.746**	0.703**	0.758**	0.788**	0.746**	0.703**	0.759**	0.715**
	(0.024)	(0.017)	(0.018)	(0.017)	(0.024)	(0.017)	(0.018)	(0.017)	(0.019)
Constant	0.098**	0.119**	0.130**	0.109**	0.100**	0.119**	0.130**	0.107**	0.136**
	(0.014)	(0.011)	(0.020)	(0.012)	(0.015)	(0.011)	(0.020)	(0.014)	(0.021)
Observations	773	1,528	1,524	1,460	773	1,528	1,524	1,460	1,456
R^2	0.593	0.570	0.582	0.580	0.593	0.570	0.582	0.580	0.592
Demographic Controls	No	No	Yes	No	No	No	Yes	No	Yes
Sample	Fashion	Full	Full	Full	Fashion	Full	Full	Full	Full
F-test (TOT DM)	–	–	–	1029 **	–	–	–	344 **	355**
F-test (TOT Experience)	–	–	–	–	–	–	–	956**	965**
F-test (TOT DM x Experience)	–	–	–	–	–	–	–	502**	518**

Notes: Linear regression models with standard errors in parentheses. The dependent variable is the factor analysis score of 8 items in Wave 3. The independent variables are indicators for the treatments: algorithmic decision-maker, 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. 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. Table SI-10 reports the full results. *p<0.05; **p<0.01

a function of exposure to algorithms per se, but rather of the nature of exposure (i.e., positive or negative). Indeed, previous studies on human-computer interactions have shown that people become less willing to rely on algorithmic recommendations after a negative experience, such as finding an error in the algorithm's output (e.g., Dietvorst, Simmons, and Massey, 2015). 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 algorithmic decision-maker 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. Appendix A.1 shows that the manipulation checks successfully distinguish between workers by their assigned decision maker (DM), as over 74% 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.

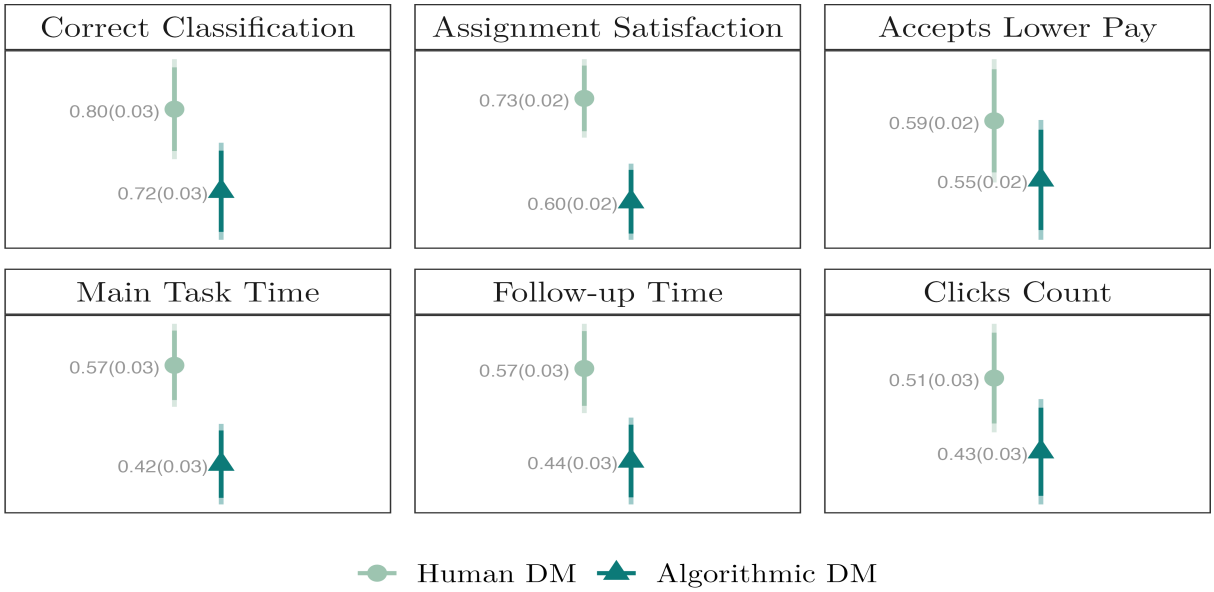
Again, the results indicate that workers who interacted with an AI algorithm as their employer did not significantly differ from other workers in their support for employing AI in public policy decisions. The estimated effect on the treated is, as expected, larger compared to the effect on all participants assigned to the treatment, but it is well below statistical significance. These results further suggest that personal interactions with AI do not affect attitudes toward the broader question of using such algorithms in public policy implementation.

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 therefore explore the treatment effects 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.4 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-9 in the Appendix. Figure 6 shows the predicted

Figure (6) Effects of Experience on Behaviors



The figure shows the predicted score of each behavioral outcome based on intent-to-treat analyses that regress them on a binary indicator for experience with algorithmic decision-making, 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-9.

values using this regression model.[15]

Our analysis reveals that workers' personal experiences with algorithmic decision-making 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), and put less effort in performing the task, and were significantly less likely to correctly classify the comments (p<0.01).

Taken together, the results suggest that the null effects of personal experience with algorithmic decision-making on AI policy-related 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.

[15]To make the interpretation easier, we converted all outcomes to indicator variables. As Table SI-10 shows, the results hold when using different measures, such as continuous measures of the number of clicks or the time spent on the task.

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 technology.

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 explore this question by randomly exposing workers to different types of relevant information. Recall that the task that all workers were assigned to perform required them to read a set of comments and then rate how positive or negative the implied assessments in those short texts were. But while some participants received eight short texts that predominantly focused on positive aspects of AI technology, others received texts of similar length that focused on potentially negative aspects of the technology, while still others were exposed to placebo-like comments about the future of fashion.

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[16] Columns 4-6 include the full sample, controlling for the decision-maker treatments.

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

Table (2) Effects of Information on Attitudes

	Dependent variable:					
	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^{***}	0.047^{**}	0.049^{**}	0.048^{**}
	(0.023)	(0.023)	(0.023)	(0.015)	(0.015)	(0.015)
Info about AI (ref: Fashion)	-0.036^{*}	-0.039^{*}	-0.038^{*}	-0.012	-0.013	-0.013
	(0.016)	(0.016)	(0.016)	(0.011)	(0.011)	(0.011)
Positive info (ref: Negative)	-0.019	-0.020	-0.018	-0.004	-0.004	-0.003
	(0.016)	(0.016)	(0.016)	(0.011)	(0.011)	(0.011)
Pre-dispositions (wave 1)	0.736^{***}	0.714^{***}	0.707^{***}	0.745^{***}	0.708^{***}	0.704^{***}
	(0.024)	(0.027)	(0.027)	(0.017)	(0.018)	(0.018)
Constant	0.127^{***}	0.152^{***}	0.177^{***}	0.112^{***}	0.117^{***}	0.131^{***}
	(0.015)	(0.026)	(0.029)	(0.012)	(0.018)	(0.021)
Model	Minimal	Demographics	Mturk HITs	Minimal	Demographics	Mturk HITs
Observations	741	741	741	1,528	1,528	1,525
R^2	0.561	0.572	0.577	0.573	0.582	0.582

Notes: The dependent variable 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-11 for the full results †p>0.1 *p<0.05; **p<0.01, ***p<0.001

The results show that exposure to positive information about the potential 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 0.035 to 0.044 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.[17] Notably, the figure shows that the treatment effect of positive 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-13 in the Appendix.

One concern might be that the valence of the informational content itself, regardless 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-12 in the appendix 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.[18]

[17] 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.

[18] We also conducted a bounding exercise to address selective attrition. To this end, we assigned workers who did not complete the post-treatment survey either their pre-treatment outcome from wave 1 (lower bound) or their factor analysis 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-15 shows that our main findings are robust to these different assumptions about the impact of attrition. This finding reinforces our conclusion that dropout between waves 2 and 3 does not pose a serious threat to

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).

When looking at which policy areas were particularly sensitive to the 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. However, attitudes did shift more in some areas than others. Specifically, exposure to positive information significantly increased willingness to support the use of predictive algorithms (rather than human decision makers) in deciding how to allocate food stamps (p<0.01), where to assign police officers to patrols (p<0.05), as well as where to locate shelters for the homeless (p<0.05). In contrast, learning about the negative implications of AI significantly decreased support for using algorithms to decide where to beef up police enforcement and, to a lesser extent, to decide on parole for defendants.

By dichotomizing the individual items, we found that in these policy domains, the positive information treatment led to a considerable increase in the share of support for the policy, not merely to a hardening of views (see Table SI-14. For instance, participants who encountered comments about the potential benefits of AI were seven percentage points more likely to support the use of AI for deciding where to locate police patrols compared to those who had received no information about AI (p<0.01).[19]

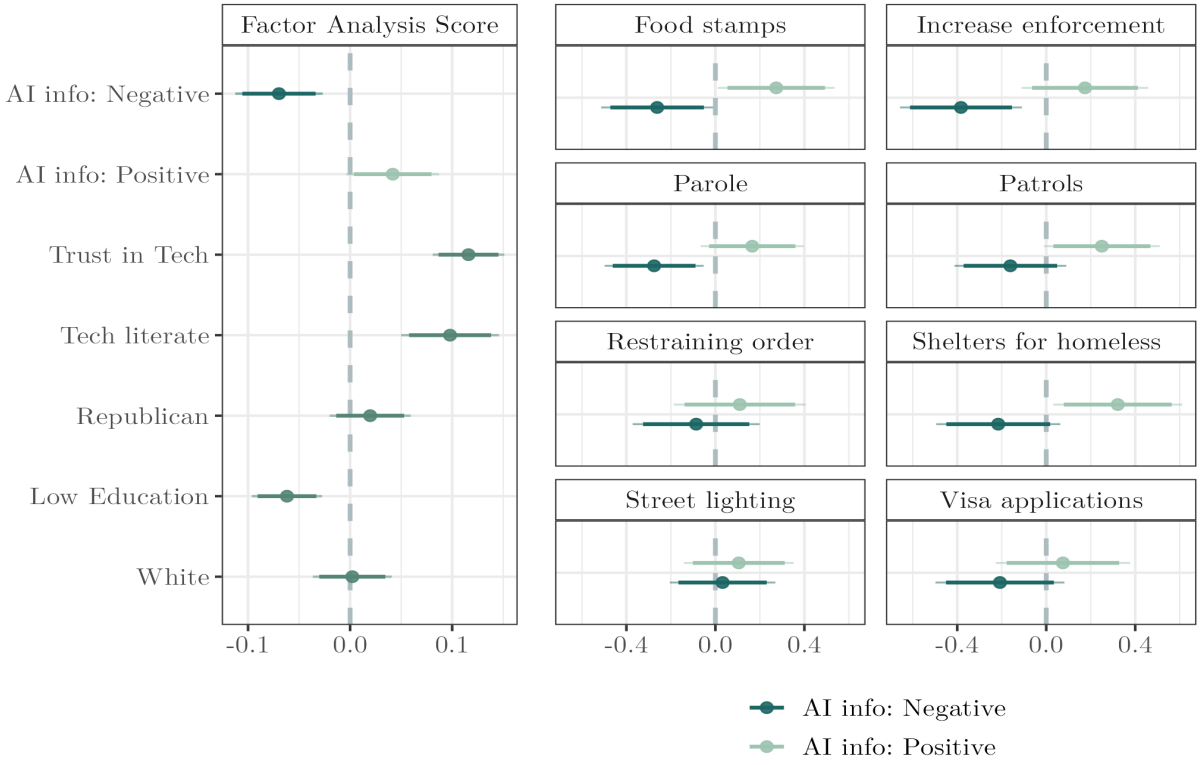
5.4 What Type of Information about AI Affects Attitudes?

To better understand the effect of information on workers' attitudes towards the use of AI, in this section we analyze which specific comments the participants found most persuasive. After they had completed the task, we asked them to indicate which comment they found

estimating the treatment effects of information.

[19]These analyses of dichotomized versions of the individual items were not pre-registered; we conducted them to help illustrate the substantive size of the effects.

Figure (7) Effects of Exposure to Information Treatments

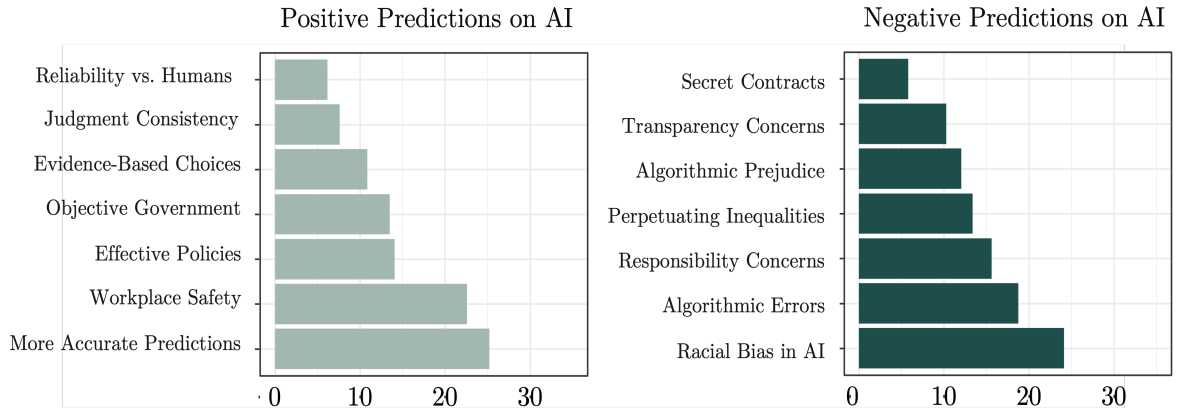


The figure shows the results of OLS regressions. The independent variables are indicators for positive information on AI, negative information on AI, or placebo information about fashion. The left panel shows the estimated treatment effects relative to the effects associated with key covariates, excluding 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 outcome, as well as indicators for the type of experience treatment and valence of information. We limited the sample to workers who were assigned to the human decision maker treatments to ensure a more robust comparison. See Table SI-13 for the full results. Thin bars represent 95% confidence intervals, and thick bars represent 90% CI.

most convincing and to explain why. We then constructed a dictionary for each comment, listing its key phrases and words. We use the responses to the open-ended questions to identify the comments each treatment group found the most persuasive.

Figure 8 presents the results. Among workers exposed to positive information about AI, the most persuasive comments were those that emphasized AI's high degree of accuracy (25%) and its potential to enhance workplace safety (21%). Notably, these comments were considered far more persuasive than those highlighting the limitations of human decision-

Figure (8) 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.

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 human decision-making that can be influenced "by irrelevant factors, such as their mood."

Among workers who received predominantly negative information about AI, the figure shows that comments that addressed concerns about racial discrimination and unfairness (23%) or potential issues with utilizing aggregate data for individual decision-making (19%) were more frequently mentioned. Interestingly, the most cited comment that workers in the negative information treatment group described as the most persuasive was one with a contrary message. In fact, 30% cited a positive comment that stated, "Artificial intelligence relies on massive amounts of data to make predictions. This can lead to a high degree of accuracy," as most convincing.[20]

[20]One participant, for example, explained, "I find this most convincing simply because most predictive applications currently in use (outside of AI) do not have access to the massive amounts of data that AI systems do. I believe that this is an area where AI will really shine."

These findings suggest that workers' attitudes towards AI are responsive to specific information about its potential implications but are not changing their views in response to any positive assessments of the technology's merits.

5.5 Predispositions and Information Processing

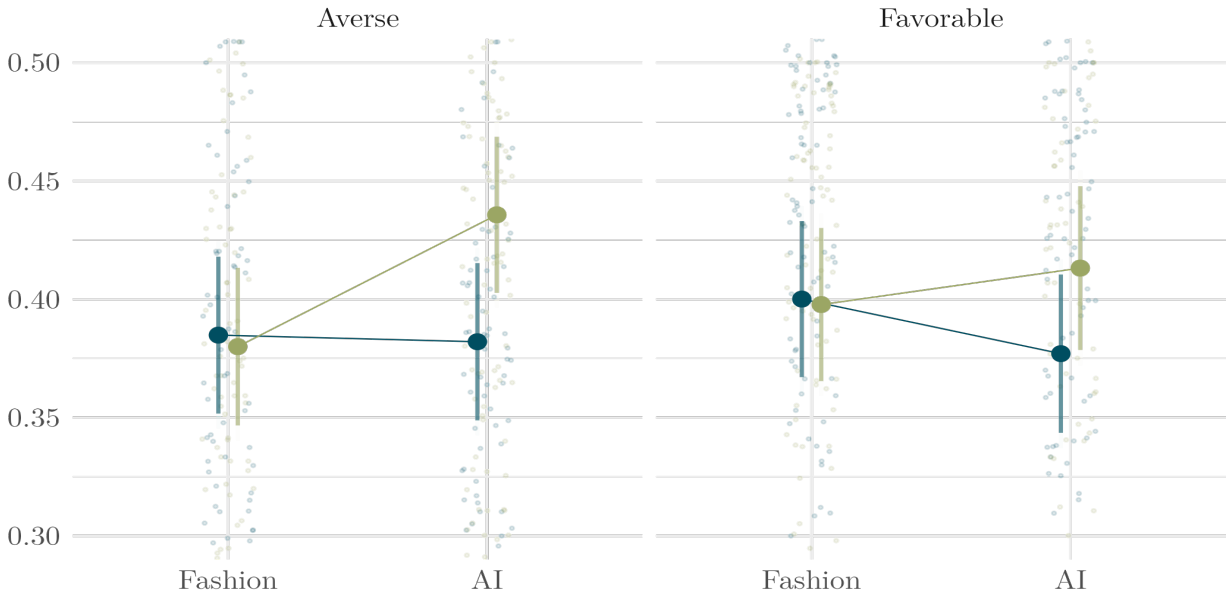
Our results demonstrate that exposure to information about the positive aspects of AI made workers, on average, more amenable toward the use of the technology in public policy decisions. Yet it remains unclear whether the positive information persuaded those who were previously apprehensive about the use of AI or whether, instead it simply reinforced the opinions of those already predisposed to agree with the information. 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-16 reports the results.

We find that positive information about AI significantly increases support for its use in public policy implementation for both groups of workers, irrespective of their predisposition. The interaction between positive information on AI and negative predisposition is not statistically significant.

Figure 9 graphically illustrates the tabular results, showing the predicted outcome by treatment group and predisposition. Contrary to what motivated reasoning theory would suggest, the figure shows that workers who were skeptical of AI actually updated their views in response to positive information about AI and grew more favorable of its use for policy implementation decisions (0.62, $p<0.01$). In contrast, negative information about AI's potential implications appears to have had little impact on participants' views.

We test for ceiling and floor effects by excluding the respondents identified in the baseline survey as either most opposed and most supportive of AI use in policy decisions (i.e., those in

Figure (9) 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% and 90% confidence intervals. The model controls for decision-maker and experience treatments, the pre-treatment outcome (as a continuous measure) as well as key demographic covariates. Column 5 in Table SI-16 shows the full results. Data points correspond to individual raw observations.

the upper and lowest deciles of the scale). The findings remain consistent under this exclusion as well. We also re-ran the analysis while excluding respondents around the midpoint of the scale, as they may be suspected of being indifferent. Even so, we find that the effect does not change substantially. Overall, then, our analysis indicates that rather than rejecting or ignoring information that challenges their prior views, participants updated their preferences 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.

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.[21] 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).[22] These findings point

[21] For example, OpenAI, the organization behind ChatGPT, recently launched a grant program to support ten projects that foster public deliberation on "how to establish a democratic process for determining the rules that AI systems should follow, within legal boundaries."

[22] In fact, the U.S. Congress recently demonstrated (rare) bipartisan cooperation on several bills related to AI regulation, such as the National AI Commission Act. This Act which would establish a commission of 20 members from both parties to review the current regulatory approach to AI development and implementation (Ghose, 2023).

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 special interest partisan 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). [23] The debate over AI regulation may undergo similar dynamics.

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 on issues of social justice and will therefore 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 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 issue at hand. Yet another possibility is that people do make the connection but view the societal impact of AI in more normative terms and hence go beyond their own interests or experiences when forming their attitudes. This possibility would be consistent with work that documented individuals'

[23] 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

negative reactions toward algorithmic systems that risk public values such as fairness and transparency, even if they themselves are not 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, 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 particpants' 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 AI-related 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 automation poses as well as shape their preferences for policy interventions 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*.

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.

Burrell, Jenna (2016). "How the machine 'thinks': Understanding opacity in machine learning algorithms". In: *Big data & society* 3.1, p. 2053951715622512.

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.

Eubanks, Virginia (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

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.

Ghose, Anindya (2023). *AI Regulation Is Coming To The U.S., Albeit Slowly*.

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.

Kahneman, Daniel, Olivier Sibony, and Cass R Sunstein (2021). *Noise: a flaw in human judgment*. Hachette UK.

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.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein (2018). "Discrimination in the Age of Algorithms". In: *Journal of Legal Analysis* 10, pp. 113–174.

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.

Lepri, Bruno, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck (2018). "Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges". In: *Philosophy & Technology* 31, pp. 611–627.

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.

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

Noble, SU (2018). *Algorithms of oppression: how search engines reinforce racism: nyu press*.

O'neil, Cathy (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

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

Pasquale, Frank (2020). *New laws of robotics*. Harvard University Press.

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.

Sheehan, Kim Bartel and Matthew Pittman (2016). *Amazon's Mechanical Turk for academics: The HIT handbook for social science research.* Melvin & Leigh, Publishers.

Stamm, Keith R, Fiona Clark, and Paula Reynolds Eblacas (2000). "Mass communication and public understanding of environmental problems: the case of global warming". In: *Public understanding of science* 9.3, p. 219.

Starke, Christopher, Janine Baleis, Birte Keller, and Frank Marcinkowski (2022). "Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature". In: *Big Data & Society* 9.2, p. 20539517221115189.

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.

Sunstein, Cass R (2022). "The Use of Algorithms in Society". In: *Available at SSRN 4310137*.

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.