When Do Citizens Resist The Use of AI Algorithms in Public Policy? Theory and Evidence

Shir Raviv*

November 26, 2024

Abstract

In recent years, there has been a significant rise in the use of algorithmic decision-making systems (ADS) to assist or replace human decision-making in a wide range of policy areas as policing, criminal sentencing, and social welfare assistance. How do citizens view the incorporation of this technology in guiding high-stakes decisions? I introduce a new theory to explain the conditions under which citizens view ADS as legitimate, fair, and accurate, and test it using a series of original experiments embedded in a national U.S. survey. Using evidence on a wide range of decisions and policy domains, I show that citizens exhibit aversion to the use of ADS in decisions that are seen as designed to sanction rather than to assist, and when they are required to make inferences about individuals rather than collectives. Evidence from a second experiment suggests that the employment of ADS in such contexts can significantly undermine the legitimacy of policy decisions they inform. Overall, the theory and evidence I present provide novel insights into the way AI-based tools can be used in public policy and the political implications of this growing phenomenon.

*Data Science Institute, Columbia University. sr4093@columbia.edu
The pre-analysis plan is available at https://osf.io/u7qs6/?show=view.

Introduction

In November 2020, Californians voted on a referendum to replace the state's policy of cash bail for pretrial release with an algorithmic system. Under the new law, local courts would decide whether those arrested and charged with a crime should be kept in custody or released while awaiting trial, based on an algorithmic assessment of the defendant's likelihood of showing up for the court date, the seriousness of their crime, and their likelihood of recidivism (Pislar and Puleo, 2020). Despite evidence regarding the potential effectiveness of the algorithmic system in reducing crime among released defendants without imprisoning additional people (Kleinberg, Mullainathan, and Raghavan, 2016), voters decisively rejected this proposition by a wide margin (56% versus 44%). What explains this opposition? Does the public response reflect a general resistance to algorithms in the public domain or is it contingent on the context, reflecting opposition to the specific use of these tools in the criminal justice system?

These questions are particularly pertinent, given the growing use of algorithmic decision-making systems (ADS) in a wide array of policy contexts. In the last few years, public authorities are increasingly relying on such AI-based algorithms—software that autonomously makes decisions without explicit human instruction, relying on data-driven inferences instead—to determine questions such as where to focus policing efforts, which child abuse allegations to investigate, who qualifies for public housing or how to allocate welfare benefits (e.g., Eubanks, 2018; Meijer, Lorenz, and Wessels, 2021; Robertson, Nguyen, and Salehi, 2021).

In this paper, I introduce a novel theory to explain public attitudes towards the use of ADS in policy implementation. I argue that people do not think uniformly about this technology. Rather, their views on the accuracy and fairness of these systems vary as a function of (1) the objective of the decision at stake, specifically, whether seen as assisting or sanctioning; and (2) the population directly affected by the decision: individuals versus collectives. I put this framework and its implications to empirical tests using data from two original, pre-registered experiments embedded in a national U.S. survey.

The first experiment systematically examines when people accept the use of ADS in government as appropriate and how they balance considerations of accuracy and fairness by randomizing the decision context in which the algorithm is used. Results provide strong support for the theory: people exhibit aversion to ADS, especially in decisions that are designed to sanction rather than assist, as well as when they are required to make inferences regarding individuals rather than collectives. These findings are generalizable to a wide range of decisions in various domains, including policing, public education, immigration, social welfare, and criminal justice. The analysis also highlights the tradeoff people face when considering the accuracy and fairness of ADS in decisions that assist individuals and those that sanction collectives. In these contexts, the weight given to each consideration appears to follow the pattern predicted by the theory: respondents were significantly less tolerant of ADS when used to inform sanctioning decisions with less reversible outcomes, even though they are considered to improve the accuracy of decision-making.

While public opinion does not always drive the adoption of policies, public knowledge and approval of using ADS in public policy implementation are crucial for establishing their legitimacy. This is evident in recent high-profile cases where governments and municipalities have reversed or abandoned initiatives that used ADS due to public backlash. For example, both New Orleans and Los Angeles have discontinued predictive policing algorithms following public outcry over issues such as transparency or bias (Winston, 2018; Sainato and Chiu, 2021). In the Netherlands, public outcry against the SyRI algorithm, which was found to disproportionately target welfare fraud among ethnic minority families, ultimately contributed to the government's resignation (International, 2021).

To empirically assess the political implications of these findings, I present results from a second experiment that examined whether the use of ADS affects public support for policy actions. By asking respondents to evaluate the same policies that (randomly) involved either an algorithmic or a human decision-maker, the experiment provides a useful way to evaluate public views in comparison to the status quo, assessing whether citizens care about the use of ADS and consider it when evaluating policy issues.[1] This question is particularly important as in most cases, unlike the California referendum, government agencies deploy ADS without informing or consulting with the public.

The results suggest that the use of algorithms in contexts where citizens view them as inappropriate can undermine the legitimacy of policy interventions. ADS significantly reduce support for policies that involve decisions that sanction individuals, such as deciding which child abuse allegations to investigate. In contrast, we see the opposite trend for policies that involve decisions about assistance, especially when applied to the collective, such as providing additional funding to certain schools for educational programs. Interestingly, the results also suggest that in cases where there is a tradeoff between considerations of fairness and accuracy, the hybrid use of algorithmic evaluation and human decision-making appears to be an attractive solution. While relying solely on the algorithm decreases support for the policy of selectively allocating patrols, using ADS as a support tool for policy officers actually increases support for the policy. These findings have important practical implications for the current discussion over the regulation of AI.

Beyond its practical implications, the study contributes to a growing literature on the political ramifications of the recent advancements in AI and digitization, which has primarily focused on labor market disruptions (e.g., Gallego and Kurer, 2022). This study provides insights into an important yet under-explored domain where AI-based technology increasingly influences citizens' lives, highlighting its implications for democratic governance, therefore

[1] The two experiments were embedded within the same survey. I discuss the sequence of the survey experiments in detail in the subsequent sections

underscoring the need for a more comprehensive research agenda in political science.

The study also speaks to the literature on the determinants of individuals' attitudes toward AI. Most of the experimental work on this matter focuses on the views and reactions of the users or the operators who interact directly with the algorithms (Lee, 2018; Waggoner and Kennedy, 2022). More recently, studies shifted focus to the general public who are subjected to algorithmic decisions without an option of opting out (Zhang and Dafoe, 2019; O'Shaughnessy et al., 2023). The findings presented in this paper add to the scant but rapidly growing research that underscores the contingent nature of mass attitudes (Araujo et al., 2020; Miller and Keiser, 2021; Schiff, Schiff, and Pierson, 2021; Schiff et al., 2023; Wenzelburger and Achtziger, 2023). By showing how the perceived fairness and accuracy of the same algorithmic systems can differ depending on the particular type of decision they are informing, this study adds more nuanced and systematic insights that transcend various policy areas.

Contextual Attitudes Toward Using AI Algorithms in Governance

The integration of ADS in high-stake policy domains has triggered a debate about the potential benefits and risks of these systems (Schiff et al., 2020). Proponents contend that as algorithms provide data-driven analysis on a scale, scope, and time frame that humans cannot offer, they can help deploy government resources and public services more efficiently, objectively, and accurately (Lepri et al., 2018). However, recent research has cast doubt on this idea, highlighting a range of ethical concerns, including racial bias, discrimination against marginalized groups, the perpetuation of societal inequities, a lack of transparency and accountability; and privacy violations (e.g., Barocas, Hardt, and Narayanan, 2017).

Much of this debate has focused on whether ADS can improve the accuracy and fairness of policy decision-making. Accuracy, in this context, refers to the degree to which ADS achieves the intended outcome, such as correctly identifying low-risk defendants or students with learning difficulties. Fairness, on the other hand, is more elusive. It includes procedural aspects, such as neutrality, consistency, and transparency (Tyler, 2006), which may align with accuracy, when less biased decisions are both more accurate and fairer. However, it also involves more substantive aspects that go beyond accuracy, such as promoting equal opportunities and accountability (Reich, Sahami, and Weinstein, 2020). The latter relates to the consequences of the decisions, specifically, the extent to which they affect or constrain citizens' lives.

How do citizens evaluate the fairness and accuracy of ADS? Most of the empirical work assumes that people's views of algorithms are quite fixed, either as a function of their predispositions toward the technology (Dietvorst, Simmons, and Massey, 2018; Zhang and Dafoe, 2019), or their prior knowledge about AI (Horowitz and Kahn, 2024). Other research highlights the design features of the technology, such as the quality or amount of data the algorithm is trained on or its degree of transparency (Waggoner et al., 2019; Kennedy, Waggoner, and Ward, 2022). Recent studies have shown that individuals' evaluations of ADS vary depending on the context it is used (Horowitz, 2016; Lee, 2018; Logg, Minson, and Moore, 2019; Araujo et al., 2020). Building on this contextual evidence, I argue that individuals' expectations and assumptions about the accuracy and fairness of using ADS in government largely depend on two key features of the decision. The first dimension relates to the target of the decision, namely, the population that the decision directly affects. In particular, I distinguish between decisions that target *individuals*—such as whom to stop for speeding or whom to provide with social benefits—and decisions imposed on *collectives* (i.e., groups or areas), such as which neighborhoods to patrol or which schools should receive further funding assistance.

The second dimension relates to the decision's objective, particularly whether it seems designed to sanction or benefit. *Assisting* decisions involve providing social services or public goods, such as determining where to build a new public park or who is eligible for public housing. Conversely, *sanctioning* decisions involve imposing penalties or restrictions on targeted groups or individuals, such as increasing law enforcement against illegal immigration or removing a child from their parent's care. Drawing on the distinction between negative and positive liberty proposed by Berlin (1969), sanctioning decisions directly restrict one's choices and behaviors, impacting their negative liberty. On the other hand, assisting decisions influence the conditions and resources that empower individuals or groups to pursue their goals or interests, thereby relating to their positive liberty.

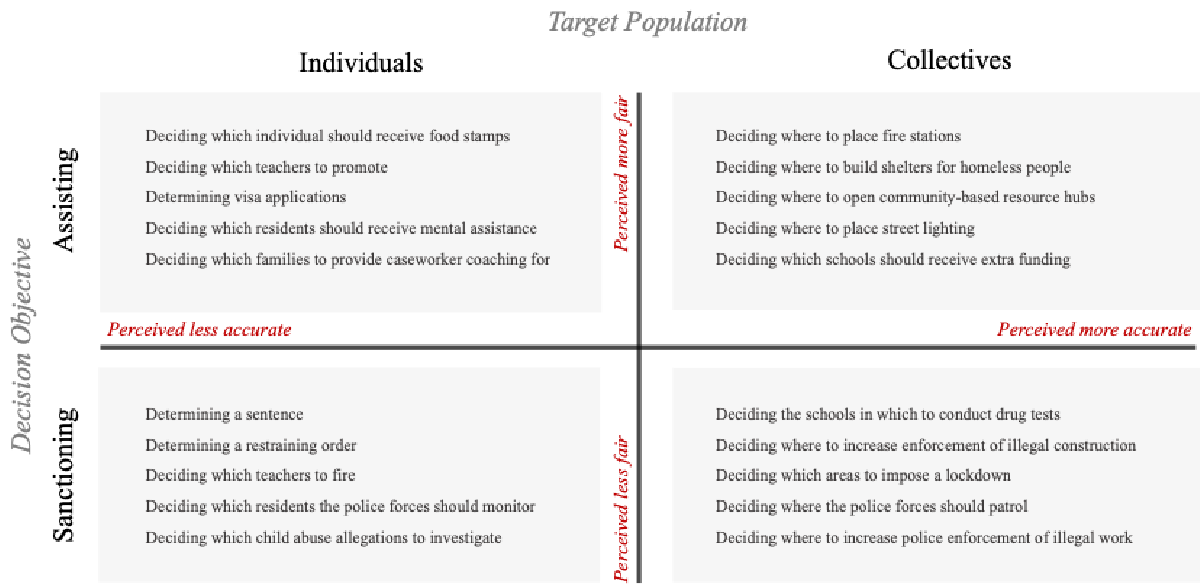
The distinction between assisting and sanctioning decisions is not always clear-cut. One could argue that determining eligibility for a social benefit or resource can be viewed as sanctioning rather than assisting. However, the theory assumes that there is a fundamental difference between decisions that "do not give" (assisting) and those that "take away" (sanctioning). The difference is derived from the potential change in the status quo, which has implications for the decision's consequences, particularly the extent to which the decision outcome is reversible. To validate this theoretical framework, I conducted a survey on MTurk, asking 150 respondents to categorize six randomly selected decisions into the four types derived from the theory, saying nothing about the identity of the decision maker. The results, reported in Figure A-4, show that respondents' answers are significantly consistent with this 2x2 classification.

Although the two dimensions are not all-encompassing, they provide a useful starting point for understanding contextual variation in preferences.[2] As Figure 1 shows, there are many examples of real-world decisions in the public sector that can be classified into this two-by-two framework.

I contend that ADS are more likely to be seen as improving accuracy when applied to

[2]Building on this framework, further study should examine these distinctions as a spectrum, where, for example, some decisions may be perceived as more assisting than sanctioning. Another useful direction is to study heterogeneity across individuals in classifying policy decisions, which could affect their views on using ADS in these contexts.

Figure (1) Four Types of Decisions in Public Policy



Notes: This figure applies the theoretical framework to real-world examples. 1

collectives rather than individuals, as they excel at processing large amounts of data but may overlook individual nuances and exceptional circumstances that human judgment and discretion can better address. In terms of fairness, the impersonal nature of ADS may be perceived as an advantage in assisting decisions that distribute benefits, as it reduces the risk of favoritism and corruption associated with human decision-makers. However, in sanctioning decisions, the same impersonality of ADS may be seen as human accountability and reasoning are deemed more important for ensuring fairness. In what follows, I characterize each of the four decision types in terms of accuracy and fairness and derive observable implications for the perceived legitimacy of using ADS in each of the four decision types.

1. Assisting collectives

In terms of accuracy, the fact that algorithmic systems rely on big data to make predictions about aggregate cases can be perceived as highly accurate, especially compared to the limited ability of humans to capture, aggregate, and process such a massive amount of information

(Green and Chen, 2019). Indeed, research suggests that people view algorithms relying on big data as inherently trustworthy, using this 'big data effect' as a heuristic to gauge the algorithm's quality (Waggoner et al., 2019).

Assuming that algorithms make decisions based on rules applied consistently over time and across different parties and situations, several studies suggest that such technology has the potential to increase not only accuracy but also fairness in human decision-making (e.g., Sunstein, 2019; Helberger, Araujo, and Vreese, 2020).

By consistently adhering to predefined impersonal procedures, algorithms can reduce the potential corruption and favoritism in decision-making, which is particularly relevant to decisions designed to assist collectives, given their distributive nature– specifically that they are often determined in isolation from each other or from a general policy rule, and have short-term gains for specific groups while less immediate and less visible costs to the whole society (Lowi, 1964). Furthermore, by analyzing extensive historical data, algorithms can identify disadvantages and circumstances beyond individual control. Using this data-driven assessment to inform decisions about resource distribution can promote equality of outcome, a substantive aspect of fairness.

Taken together, when people form judgments about the use of ADS in decisions of this kind, they are not expected to perceive meaningful tradeoffs between accuracy and fairness considerations. The upper right panel of Table 1 indicates that the use of algorithms is expected to improve both accuracy and fairness in decision-making designed to assist collectives.

2. Sanctioning collectives

For the same reasons noted in the context of assisting, data-driven algorithms appear to be highly accurate in locating areas or communities likely to face major challenges. Nonetheless, using these assessments and predictions—as accurate as they may be—to sanction and punish targeted communities rather than assist them with the resources they need can be perceived as unfair in some substantive respects.[3]

The key concern is that using ADS for sanctioning purposes can have a long-lasting impact and may adversely affect historically disadvantaged groups, thereby undermining equality of opportunity. Unlike decisions that assist collectives, where ADS can potentially promote equality of outcomes by redressing or compensating communities or areas suffering from past injustices, using these data-driven assessments to sanction groups reflects and therefore perpetuates such injustices (Barocas, Hardt, and Narayanan, 2017). A growing concern in this context is that ADS could lead to feedback effects in the sense that they not only predict events but also contribute to their future occurrence (Brayne and Christin, 2021). Consider, for example, the predictive policing algorithm widely used by U.S. police departments to assign patrols. This algorithmic system relies on linkages between locations, events, and historical crime rates to predict the areas where crimes are most likely to occur in the future. This can lead to a negative feedback loop in which police disproportionately patrol areas with historically high crime rates, resulting in more arrests in those locations, which then become the algorithm's new training data, confirming and reinforcing its earlier predictions (Ferguson, 2017).[4]

The key point here is that the same algorithmic system, which assesses the risk of crime in a particular area, may be perceived as fair in decisions that assist collectives (e.g., deciding where to put more streetlights or where to build a community-based resource center) but significantly unfair in decisions that sanction collectives (e.g., deciding the schools in which to conduct more drug and alcohol testing).

[3]Indeed, recent studies in international relations document cases of public support for using ADS in national security decisions targeting collectives (Horowitz and Kahn, 2024), particularly for defense, but less so for offensive purposes (e.g., autonomous weapons systems) (Horowitz et al., 2023)

[4]The concern that algorithmic systems not only predict future events but also shape the conditions they are designed to predict aligns with policy feedback theory, which posits that by distributing resources, policies can shape political behavior over time (Pierson, 1993).

The observable implication is that using ADS for decisions that sanction groups involves a potential tradeoff: it may be seen as more accurate but also as unfair. Since these are highly consequential decisions, I expect that fairness considerations will outweigh accuracy considerations and thus trigger greater opposition to ADS in this context.

3. Assisting individuals

The main characteristic of decisions that assist individuals is that they are usually made at the "street-level bureaucracy," a term that refers to the layer of bureaucracy, including judges, teachers, social workers, and police officers, that directly interacts with citizens and makes everyday decisions (Lipsky, 2010). These decisions often involve nuances or extenuating circumstances, making it impossible to prescribe (and thus code) a correct response ahead of time for all cases and situations.

Human bureaucrats can flexibly refine the contours of their decision boundaries before deciding on a novel or marginal case. Yet for algorithms that aggregate data, such reflexivity can only occur *after* the system has received feedback or additional training data, and more importantly, after an incorrect decision has occurred (Binns, 2019). Data-driven algorithms are, by their nature, simplifications that cannot account for all possible relevant facts about subjects and thus necessarily treat people as members of groups rather than as individuals (Brauneis and Goodman, 2018). Consequently, due to their difficulty in identifying borderline and exceptional cases, algorithms may be perceived as less accurate than humans in making decisions about individuals.[5]

The very discretion that allows humans to tailor decisions to unique situations can also lead to potential misuse—whether intentional or not—based on personal biases, favoritism, or

[5]Note that this heightened concern about individual-level accuracy reflects laypeople's intuitions and may not always align with the logic of statistics or expert opinions. The idea is that people are more aware of, and therefore worry more about, idiosyncratic elements when decisions are granular. Such individual variances are perceived to be averaged out at the aggregate level, thereby raising fewer concerns in the context of collective decisions.

other irrelevant factors (Danziger, Levav, and Avnaim-Pesso, 2011; Alkhatib and Bernstein, 2019). The rule-based, data-driven approach of ADS ensures that all individuals are treated equally under the same criteria and, therefore, can be perceived as fair from a procedural standpoint.

Overall, people are expected to weigh a trade-off between accuracy and fairness when evaluating the use of ADS in assisting individuals. As I will show, since the repercussions of these decisions on individuals' lives and opportunities are more reversible than those in sanctioning decisions, people might be more willing to accept the use of ADS, balancing the potential loss in accuracy with gains in procedural fairness.

4. Sanctioning individuals

As with decisions that assist individuals, the inability of algorithms to adapt to novel or marginal circumstances is expected to lead people to perceive them as less accurate when sanctioning individuals (Young, Bullock, and Lecy, 2019).

In terms of fairness, the black box nature and inherent opacity of ADS, which makes it difficult to explain their output, even for programmers, also makes it difficult for ordinary citizens to access and challenge their decisions (Pasquale, 2015). Such access, though, is necessary to ensure accountability in decision-making, namely, the notion that the decision maker is obligated to explain and justify a decision to the subjects to whom the decision relates. A lack of accountability is expected to produce a strong sense of unfairness, especially in decisions of this type, as any potential error would be highly significant both for an individual's life (e.g., a false positive that wrongfully convicts someone innocent) and for society's safety (e.g., a false negative that finds a guilty individual innocent).[6]

[6] An example of an algorithmic decision system is a risk assessment algorithm, which can be used to determine whether a child should be removed from the care of their parents based on the risk of future maltreatment (Cuccaro-Alamin et al., 2017). In this case, the lack of a human being with a conscience making such a fateful decision about someone's life can be perceived as unfair.

Table (1) Classifying attitudes toward ADS in the public sector

		Target Population: Individuals	Target Population: Collectives
Objective	Assisting	*(1)* *Trade-off: AI less accurate but fairer than humans*	*(2)* *No trade-off: AI more accurate and fairer than humans*
		Reversible outcomes	
	Sanctioning	*(3)* *No trade-off: AI less accurate and less fair than humans*	*(4)* *Trade-off: AI more accurate but less fair than humans*
		Less reversible outcomes	

Let us return to the example with which this paper begins: the proposition of replacing California's bail system with an algorithmic system that predicts the risk of defendants committing future crimes. As in Kafka's novel *The Trial*, in which the protagonist Josef K. is arrested, charged, sentenced, and ultimately punished without knowing the charges or meeting the prosecutor, ADS could place individuals in a similarly Kafkaesque position in which they feel they are at the mercy of an entity they do not understand, and whose decisions are not transparent or explained. Accordingly, as shown in the lower right panel of Figure 1, for sanctioning decisions that have less-reversible repercussions for the lives and liberties of individuals, I expect that people on average view ADS as both less fair and less accurate compared to other contexts.

To conclude, Table 1 lists the characteristics of each type of decision in terms of accuracy and fairness and the potential tradeoff between them. As shown in the table, I expect that citizens, on average, will consider ADS to be fair and accurate for decisions designed to assist collectives, but reject these systems as both less accurate and less fair, when employed to sanction individuals. Finally, in the case of tradeoffs between accuracy and fairness, I expect citizens to be highly sensitive to the presence of a human decision-maker in decisions with less-reversible consequences, and thus they will be less tolerant of the use of ADS. In the next sections, I empirically assess these theoretical predictions.

Research Design

To evaluate the theory and its observable implications, I designed two original experiments embedded in a national U.S. survey. The sample consisted of 1,590 adults, recruited in March–April 2022 by the survey company Dynata (formerly Survey Sampling International - SSI), which is commonly used in social science research (Malhotra, Monin, and Tomz, 2019; Read, Wolters, and Berinsky, 2021). SSI used quota sampling to approximate the US adult population with respect to gender, age, education, and race/ethnicity. Table A-1 in the Appendix shows the characteristics of the sample compared with those of the general US population. The table indicates that the sample is representative along the quota dimensions. For more details about the sample see Appendix A.

The survey includes two experimental studies. The *Decision-Context Experiment*, which directly tests the theory by assessing how the perceived appropriateness, accuracy, and fairness of ADS vary across contexts by randomizing the decision context in which the algorithmic system is implemented, and the *Decision-Maker Experiment*, which evaluates the implications of these views, by assessing how the use of an ADS affects public support for policy actions it informed, by randomizing the identity of the decision maker. Respondents participated in all experiments, but treatment assignment in each experiment was independent.

Figure 2 illustrates the flow of the survey design.[7] To minimize priming effects, I presented the *Decision-Context Experiment*, which explicitly compares ADS to human decision-making, at the end of the survey. The key concern was that such direct questions would have artificially diverted respondents' attention to the identity of the decision-maker when evaluating the policy proposals asked in the *Decision-Maker Experiment*.[8]

[7] A pre-analysis plan detailing the design and hypotheses was preregistered prior to fielding the survey, and can be found in Appendix G.

[8] Table A-18 in the Appendix shows that the results from the Decision-Context experiment remain the same when controlling for the treatments received in the Decision-Maker Experiment.

To ensure respondents have a similar definition of a predictive algorithm in mind, the following description was provided at the beginning of the survey: "A predictive algorithm is a computer software that makes decisions without human instruction, relying on a massive amount of data." To reduce concerns about experimenter demand effects, I communicated this definition indirectly, along with two other definitions relevant to the survey.[9]

Decision Context Experiment

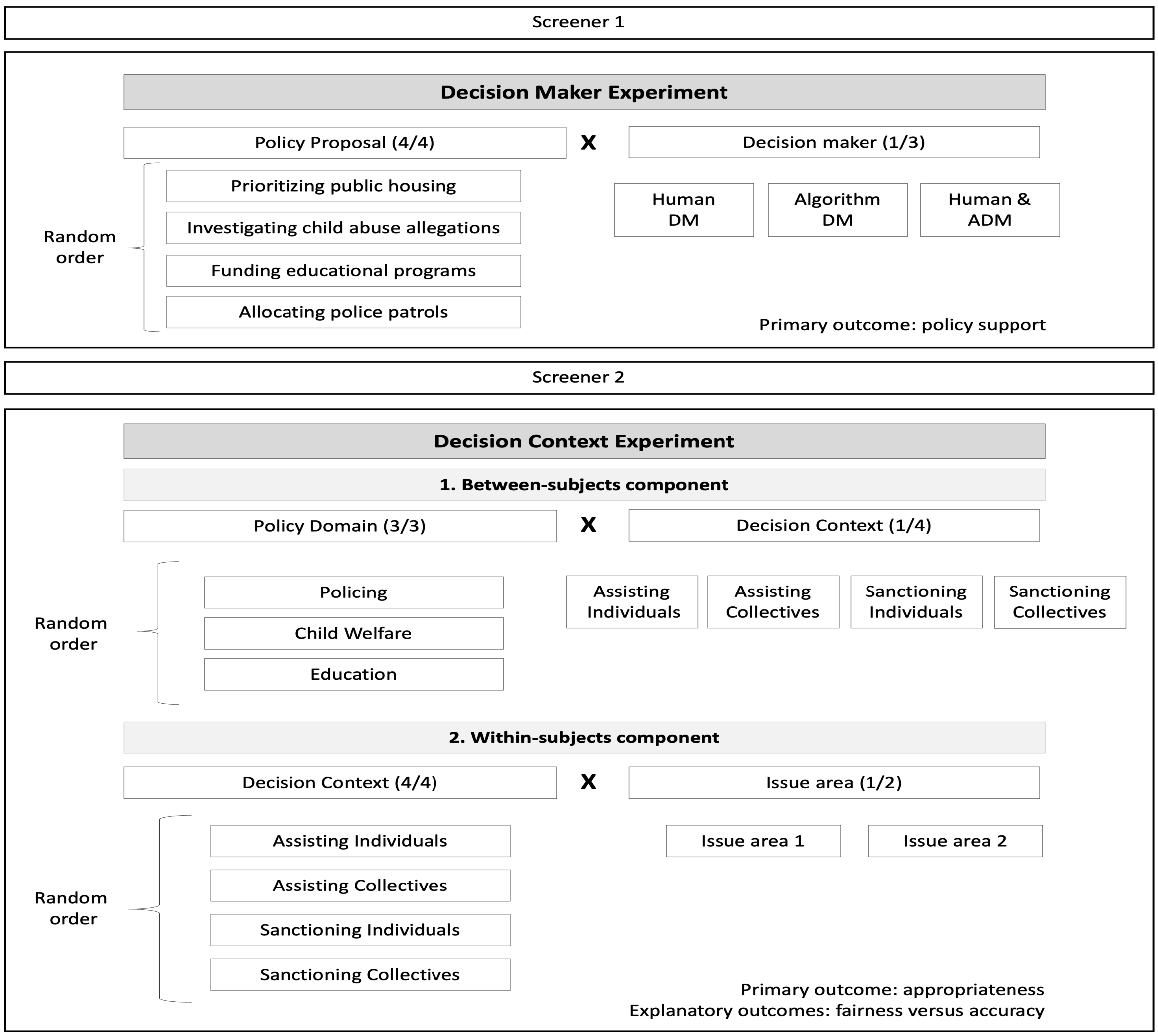
The Decision Context Experiment directly tests the theory by examining contextual variation in people's views on the use of ADS in public policy implementation across various policy domains and issue areas. Respondents were presented with a matrix of several randomly selected policy decisions and were asked to evaluate the appropriateness and, in a follow-up question, to assess the perceived accuracy and fairness of using ADS in each decision. The matrix includes two experimental components.

Between-Subject Component. Respondents first evaluate decisions from three high-stakes policy domains: policing, education, and child welfare, presented in a random order on the same matrix. Each policy domain was independently randomized along two theoretical dimensions: (1) whether the decision assists or sanctions and (2) whether the decision targets individuals or collectives.[10] Table 2 provides the wording of the questions by policy domains.

[9] This item was also used as a screener. On the next page, respondents were presented with four definitions and asked to indicate which one did not appear on the previous page. Those who failed to answer correctly were immediately removed from the study before the randomization to the "decision-maker" experiment. The survey included an additional screener question before the "decision-context experiment." The question asked: "People are very busy these days and many do not have time to follow what goes on in the government. We are testing whether people read questions. To show that you've read this much, answer both 'extremely interested' and 'very interested."' Only respondents who passed both pre-treatment screeners completed the full survey and were included in the analysis. Furthermore,I incorporated an additional non-screening attention check within the matrix of the decision-context experiment, which asked respondents the following: "It's important that you pay attention to this study, please tick 5."

[10] Notably, all respondents were asked about all three policy domains. By starting with the three policy domains, the experiment incentivizes respondents to compare ADS across policy domains rather than to focus on differences in the type of decision within domains as the theory predicts. This approach thus provides a hard test for the theory

Figure (2) Survey Design



Notes: Figure 2 shows the sequence of the experiments embedded in the survey, the randomization procedures taken within each experiment, and the outcomes included in each experiment.

Within-subject component. Respondents were presented respondents with four additional items on the same matrix, each corresponding to a different decision type: assisting individuals, assisting collectives, sanctioning individuals, and sanctioning collectives. For each decision type, the issue area was randomly assigned to one of two issue areas to ensure the broad applicability of the theory across various relevant policy decisions. For example, all respondents were asked to evaluate the use of ADS in one out of two decisions that sanction individuals: either deciding sentencing based on a prediction of the individual's risk of committing a future crime, or deciding to issue a restraining order based on a prediction of the individual's risk of assaulting their partner.[11] Rather than isolating the effect of a specific policy issue, this component aimed to assess a systematic variation within individuals across the four decision types while covering a wider range of policy interventions and issue areas beyond those used in the between-subject component. In that sense, this component provides additional correlational evidence that complements the primary between-subject component.[12]

Balance tests, as shown in Table A-5, confirm that all conditions are balanced across key demographic covariates, including gender, race, age, educational attainment, and technological literacy (see Appendix A for measurement details). To account for potential spillover effects, I randomized the order of the items in which the items were presented to respondents within each matrix component. Table A-7 shows the results remain robust when controlling for order effects.

In this experiment, the main dependent variable is the perceived appropriateness of using ADS in public policy implementation. The wording for the question reads as follows:

[11] Please see Table A-2 for question and treatment wordings.

[12] Ideally, both the decision type of decision and the policy domain would have been fully randomized across all decisions within and between subjects. However, finding comparable real-world examples for each of the four decision types within the same policy domains. Still, there is a wide range of other relevant real-world examples in which ADS are being used and discussed that are worth examination but are not fully comparable across domains or types of decisions. The within-subject component addressed this tradeoff between internal validity and the desire for a broader policy scope.

"We ask that you read the description of several policy decisions. For each please indicate how appropriate it is to have that decision made by an algorithm rather than by a human being," with answers ranging on a seven-point scale from 1 "extremely appropriate" to 7 "extremely inappropriate."[13] As preregistered, I dichotomize this variable to facilitate the interpretation of the results in a clear and politically substantive way. The variable is coded as 1 for respondents who found the use of ADS appropriate (above the middle "indifferent" category) and 0 otherwise. This approach allows me to estimate the proportion of the population open to ADS use in a clear and politically meaningful manner.[14]

To disentangle the key considerations when evaluating these algorithmic systems, I use two additional outcomes. Respondents rated the *fairness* and *accuracy* of ADS in each of the previous decisions on a seven-point scale ranging from "extremely inaccurate/unfair" (1) to "extremely accurate/fair" (7). These two questions were presented side-by-side in the same matrix and in randomized order to minimize potential order effects. To estimate the proportion of the population that perceived ADS as accurate or fair, I dichotomized these variables, assigning a value of '1' if the respondent chose any of the three categories above the midpoint on the scale and '0' otherwise.[15]

Results: Effect of decision context on perceived appropriateness

My main interest is in evaluating how the perceived appropriateness of using ADS in public policy changes based on the decision type along two dimensions: (1) the subject of the decision and (2) the objective of the decision. I begin with analyzing data from the between-subjects component, which independently manipulates the decision type within three policy domains. For each domain, I calculate the average treatment effects (ATEs) of the two

[13]I intentionally avoided using the term "legitimate" in the question due to its strong legal connotations, which could influence respondents to consider the legality of ADS uses rather than their personal judgment and sense of right and wrong.

[14]Table A-7 reports results using alternative cut-off points or the full seven-point scale.

[15]See Tables A-3 and A-11 for summary statistics of the three outcomes.

Table (2) Decision Wordings Randomized in the Between-Subjects Component

	Assisting	Sanctioning
Public Education		
	Assisting	**Sanctioning**
Individuals	Deciding which teachers to promote based on an assessment of their effectiveness in improving students' grades.	Deciding which teachers to fire based on an assessment of their effectiveness in improving students' grades.
Collectives	Deciding which schools should receive extra funding for alcohol and drug education programs, based on the risk of juvenile crime in that area.	Deciding at which schools to conduct drug and alcohol tests, based on an assessment of the risk of juvenile crime in that area.
Policing		
	Assisting	**Sanctioning**
Individuals	Deciding which residents should receive certain social services and mental health assistance, based on an assessment of their likelihood of shooting someone with a gun.	Deciding which residents the police forces should monitor, based on an assessment of their likelihood of shooting someone with a gun.
Collectives	Deciding where to place street lighting, based on an assessment of the risk of crime in the area.	Deciding where the police forces should patrol, based on an assessment the risk of crime in the area.
Child Welfare		
	Assisting	**Sanctioning**
Individuals	Deciding where to open community resource centers, based on an assessment of the risk of child abuse and neglect in neighborhoods.	Deciding where police forces should increase enforcement, based on an assessment of the risk of child abuse in neighborhoods.
Collectives	Deciding which families to provide caseworker coaching and mental health services, based on an assessment of the risk of child abuse.	Deciding which child abuse allegations to investigate, based on an assessment of the risk of child abuse.

Notes: This table details the treatment conditions included in the between-subject experiment. Respondents received decisions from three policy domains, each independently randomized into one of the four types of decisions.

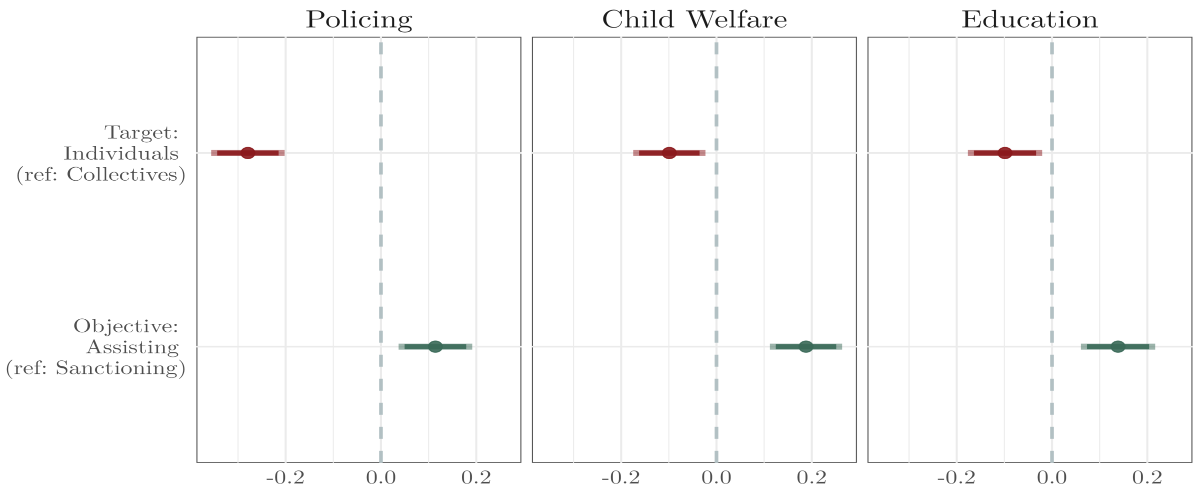
dimensions.[16]

Figure 3 presents estimates from linear three probability models (LPM) studying the effect of the two theoretical dimensions on the probability of finding the use of ADS appropriate in each policy area: education, policing, and child welfare. The analysis uses data from the first item randomly presented to respondents. Results are reported in columns 1, 3, and 5 of Tables A-6. To enhance statistical power, Table A-7 replicates the results using pulled data from all policy domains and controlling for the presentation order of the items. Table A-9 confirms that all results remain substantively similar when using a multilevel analysis that accounts for both between and within-subject variation.

Consistent with the theory, the results show that people are distinctly less tolerant of ADS when they target individuals rather than collectives. This negative effect is statistically

[16]The mean value of the three dependent variables and associated confidence interval by the four types of decisions are reported in Table A-4.

Figure (3) Effects of decision contexts on perceived appropriateness of ADS, across domains



Notes: The figure shows marginal effects estimated separately for each policy area: education, child welfare, and policing, using data collected from the first item randomly presented to respondents. The dependent variable takes the value of '1' if the respondent indicates that it is appropriate to use ADS in this area and '0' otherwise. The independent DV are indicators for the context of the decision: the subject on which the decision is made and the objective of the decision. Base categories are decisions on collectives and decisions that sanction. The full analysis can be found in Table A-6, specifically in columns 1, 4 and 7.

significant and substantively meaningful across all three policy areas ($p< 0.05$). For example, in child welfare, using an algorithmic system to assess the risk of child abuse in a specific family instead of a neighborhood significantly decreases the probability of viewing it as appropriate by 10 percentage points.

When looking at the objective of the decision, Figure 3 shows that ADS face significantly less resistance when used for assistance rather than sanctioning. Across all three policy domains, respondents were significantly more likely to view ADS as appropriate when informing assisting rather than sanctioning decisions ($p< 0.001$). As Table A-6 shows, the estimates are statistically significant across policy domains, ranging from 11 percentage points in policing to 19 percentage points in child welfare. The results are also substantively large. For instance, in public education, an algorithmic system assessing teachers' effectiveness in improving students' grades was accepted by only 15 percent of respondents when used to decide which teachers to *fire*, compared to 34 percent when used to decide which teachers to *promote*.

I conducted a set of tests to confirm the robustness of the findings. As Table A-6 shows, controlling for demographic characteristics, such as age, gender, education, and race, and other attitudinal covariates, such as technological literacy or prior knowledge of AI, does not alter these results. Tables A-7 and A-8 confirm that the results remain consistent when using logistic regression or alternative measures of the outcome. Moreover, to ensure that respondents were attentive to the treatments, I measured the response time for each question (Read, Wolters, and Berinsky, 2021). As table A-8 shows, the findings are robust when controlling for both fast, likely inattentive respondents who rush through surveys and very slow respondents who may be distracted and exhibit longer response times. [17] The results also hold when controlling for inattentive respondents using the non-screening attention check embedded within the same matrix of the experiment.

To confirm the generalizability of these findings beyond the specific items used in the between-subjects design, I analyze data from the within-subject component, which covers a wider range of issue areas, including decisions about restraining orders, criminal sentences, providing food stamps, study assistance, allocating shelters for the homeless, fire stations, enforcing illegal instructions, and illegal work. I employed an LPM regressing a binary outcome for the perceived appropriateness of using ADS on indicator variables for the two theoretical dimensions—the subject and the objective of the decision—and their interaction while controlling for the issue area randomized for each decision and using fixed effects for respondent. The results, reported in Table A-13, are highly consistent with the main findings and further support for the theory, showing a robust association between the type of decision and the perceived appropriateness of using ADS in this range of other policy issues. Once again, ADS is significantly less likely to be deemed appropriate in decisions

[17] This analysis was not pre-registered.

involving sanctions rather than assistance ($p < 0.01$) and in decisions applying to individuals rather than collectives ($p < 0.01$). Again, results are of a similar magnitude when using the alternative outcome measure (columns 2-4), and when using a linear mixed model with random intercepts for different policy issues and for each respondent (columns 5-6).

Additional Results: Fairness-Accuracy Trade-offs

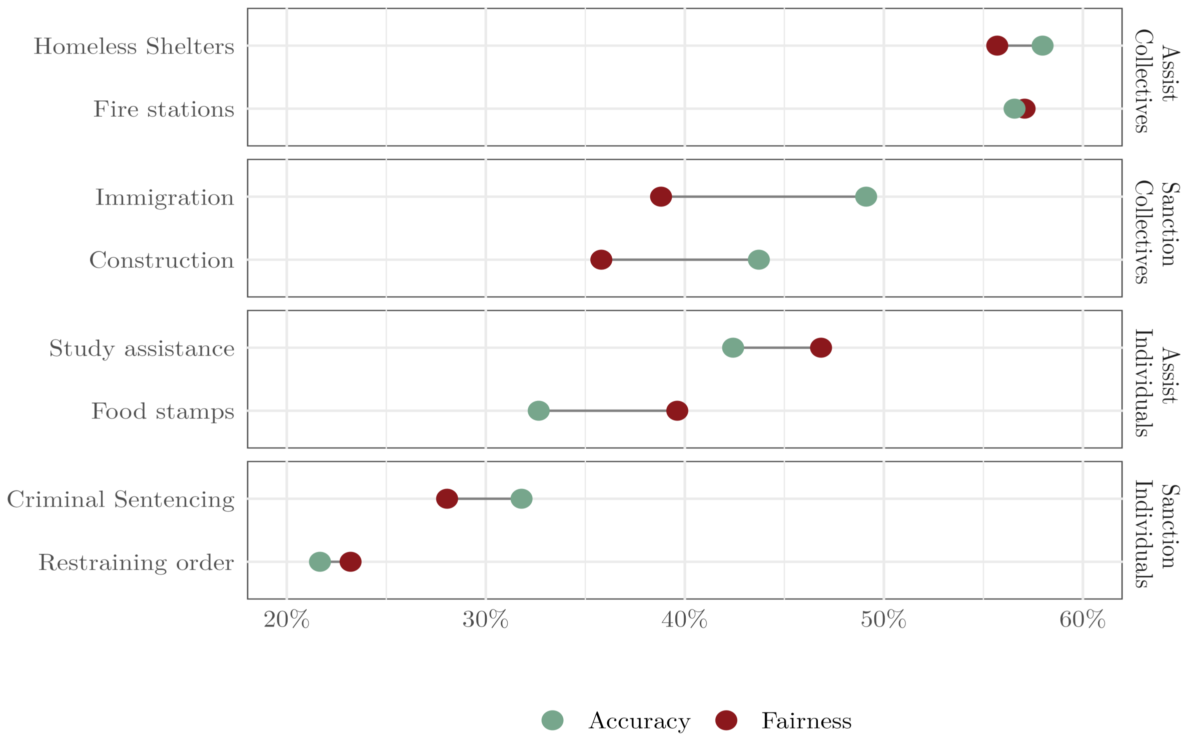
The results indicate that public opinion on ADS in government varies across and within policy domains, depending on (1) the decision's target and (2) its objective. Respondents are more likely to find ADS appropriate for assisting rather than sanctioning decisions and when targeting collectives rather than individuals. Indeed, using ADS to inform decisions assisting collectives received the highest acceptance rate, while using ADS in decisions sanctioning individuals received the lowest share.[18] The theory explains this variation by considering individuals' expectations about the perceived accuracy and fairness of ADS and the potential trade-offs between these considerations in two other decision types: assisting individuals and sanctioning collectives. To explore these mechanisms, I examine the perceived fairness and accuracy of using ADS in each of the four decision types using data from the within-subject component. Figure 4 compares the proportion of respondents who deem ADS use fair with those who regard it as accurate, for each decision type and issue area. Table A-14 formally tests these differences using paired t-tests.

Figure 4 gives rise to several notable findings. First, consistent with the main results, ADS are seen most favorably when used to assist collectives such as determining the location of a new fire station or homeless shelter. In this context, ADS received the highest ratings on both fairness and accuracy, with no significant trade-off between the two.[19] Similarly, there is minimal trade-off between fairness and accuracy when ADS are used to sanction individuals

[18] See Tables A-4 and A-12 for the full descriptive results.

[19] This finding is notable given that the question format explicitly asked respondents to compare these dimensions, potentially incentivizing them to identify differences.

Figure (4) Perceived Fairness versus Accuracy, by Decision Context and Issue Area



Notes: This figure shows the share of respondents who evaluate ADS as accurate (green dots) compared to the share of respondents who evaluate it as fair (red dots), across the four decision types and issue areas included in the within-subject component.

(p=0.264). The bottom panel shows that in both issue areas–criminal sentencing and restraining order, the shares of respondents who perceived ADS to be fair and accurate do not exceed 28%-31% respectively – about 26-27% lower than when ADS used in decisions assist collectives ($p < 0.01$).[20] One potential concern is that the results may reflect people's general aversion towards these decisions, regardless of the decision-maker. The "Decision-Maker" experiment addresses this concern by isolating the effect of the decision-maker (human vs. algorithm) on policy support.

The remaining two decision types—assisting individuals and sanctioning collectives—elicit

[20]The strong disapproval is also evident in the between-subject experiment. As tables A-4 shows, the percentage of respondents who found the use of ADS in this context to be appropriate is significantly low, ranging from 15 (child welfare and education) to no more than 20 (in policing).

more ambivalent opinions, with respondents grappling with trade-offs between fairness and accuracy, consistent with theoretical predictions. For decisions assisting individuals (e.g., determining eligibility for food stamps or study assistance), the perceived fairness of using ADS is significantly higher than its perceived accuracy (p<0.05 and p<0.001, respectively). Conversely, for decisions sanctioning collectives (e.g., increasing enforcement for illegal construction or work), perceived accuracy significantly outweighs perceived fairness (p<0.001). This aligns with the theory, which suggests that while ADS may improve accuracy in such contexts due to their ability to process vast amounts of data, using them to sanction rather than assist targeted communities can be seen as unfair.

The between-subjects analysis further supports these findings. Figure A-3 illustrates the predicted appropriateness, fairness, and accuracy of ADS for each decision type, based on a mixed-effects model that regress these outcomes on indicators for the decision-type treatments using random intercepts for the policy domain and the respondent. Notably, when ADS are used to sanction collectives, a significant difference emerges: while perceived accuracy is higher, respondents find this use significantly less appropriate. The confidence intervals reveal a significant gap between perceived accuracy and appropriateness, whereas perceptions of appropriateness and fairness do not significantly differ. This pattern aligns with the theoretical expectation that even if perceived to improve accuracy, ADS are less tolerated for decisions involving sanctions given their high-stake consequences which are less reversible compared to other types of policy decisions. The result is also consistent with previous work showing that the public prior values of fairness when contemplating the use of ADS in government (Schiff, Schiff, and Pierson, 2021).[21]

What does this finding tell us about potential reactions and backlash from citizens when

[21] It is important to note that while this analysis provides suggestive evidence in line with the patterns predicted by the theory, it does not estimate the relative effects of accuracy and fairness on the assessment of ADS appropriateness. This is because the design treats them as three dependent variable outcomes. Experimentally isolating these two considerations will be an important task for future research.

they become more aware of algorithms' role in informing public policies? The fact that people, in principle, oppose using ADS does not necessarily indicate that they would, in practice, react negatively to policy decisions involving ADS. I now turn to assess the political implications of these findings.

Decision Maker Experiment

To learn about the political implications of these findings, the decision-maker experiment independently manipulates the identity of the decision-maker and the policy contexts. Unlike direct survey questions that capture explicit preferences, this experiment aims to indirectly capture citizens' attitudes. By asking different respondents to evaluate the same policies, which (randomly) involve either an algorithmic or a human decision-maker, this experiment allows me to assess whether in fact, citizens care about the use of ADS, and how they consider these systems when thinking about concrete policy issues.

Specifically, respondents were asked to evaluate four policy proposals presented in random order: (1) prioritizing housing based on disability rather than waiting period; (2) investigating allegations based on the risk of child abuse instead of investigating all allegations; (3) allocating police patrols based on the risk of crime rather than allocating patrols at random; and (4) providing extra funding for alcohol and drug education programs for selected schools identified as problematic. The wording of each policy proposal and the treatment conditions are presented in Table 3. As the table shows, these four scenarios were chosen according to two distinguishing features derived from the theory, and were based on real-world initiatives of using ADS that are currently being promoted or implemented in the public sector.

The key aim of this experiment is to evaluate the reaction of citizens to the use of ADS in public policy, assessing whether they actually care about the use of algorithmic systems when thinking about policy issues. Therefore, the key dependent variable measures support for

Table (3) Policy scenarios and experimental treatments

	Individuals	Collectives
Assisting	***Public Housing*** **The issue:** Homelessness has increased over the past decade. The number of people currently homeless exceeds the number of affordable housing units available to them. **Policy solution:** To manage this shortage, some propose that [treatment condition] should decide which individuals receive housing first, prioritizing those with the most severe disabilities for assistance, regardless of the time they have been waiting on the list.	***Public Education*** **The issue:** In recent years, violent crime among juveniles has increased nationwide. Many of these crimes have been committed under the influence of drugs and alcohol. **Policy solution:** To address this problem, some propose that [treatment condition] should decide which schools receive additional funding for alcohol and drug education programs based on an assessment of the risk of juvenile crime in the area.
Sanctioning	***Child Welfare*** **The issue:** The number of calls reporting suspected child abuse or neglect is very high. Yet, some of them turn out to be false. Policy solution: To manage the high number of reports, some propose that instead of investigating every allegation, [treatment condition] should decide which allegation to investigate based on a preliminary assessment of the family's risk of child abuse or neglect.	***Policing*** The issue: As part of the fight against rising crime in the U.S., many police departments are concentrating their efforts on preventing incidents from occurring by increasing deterrence, instead of reacting to incidents after they occur. Policy solution: As part of this approach, some propose that instead of random patrols, [treatment condition] should decide where police officers patrol based on a prediction of where crimes are most likely to occur.

Notes: This table provides the wording of the policy scenarios and the experimental conditions. The full text of all questions in the survey is available in the Appendix.

the proposed policy. Respondents were asked to indicate the degree to which they support or oppose a policy proposal, with answers on a five-point scale ranging from "strongly oppose" to "strongly support." As preregistered, to facilitate the interpretation of the results, I recoded the scale to a binary measure with a value of 1 for positive answers ("strongly support" or "somewhat support") and 0 otherwise.[22] Importantly, the experiment was designed such that respondents would view the policy itself as the center of the question, not the identity of the decision maker. For example, in the proposal to prioritize housing, respondents were asked to indicate whether they supported or opposed the prioritization of public housing based on an individual's disabilities rather than time spent on the waiting list. In other words, the decision maker who prioritized housing was not the object of interest.

All respondents evaluated the same four policy proposals. For each case, I independently randomized the identity of the decision-maker implementing the policy decisions: a human officer in the control group and a predictive algorithm in the treatment group.[23] The primary

[22]Using this binary outcome allows me to capture the potential shifts in respondents who were initially indifferent about the policy—this is a key segment that could determine political outcomes.

[23]Summary statistics, and balance tests across experimental conditions are reported in Table A-15.

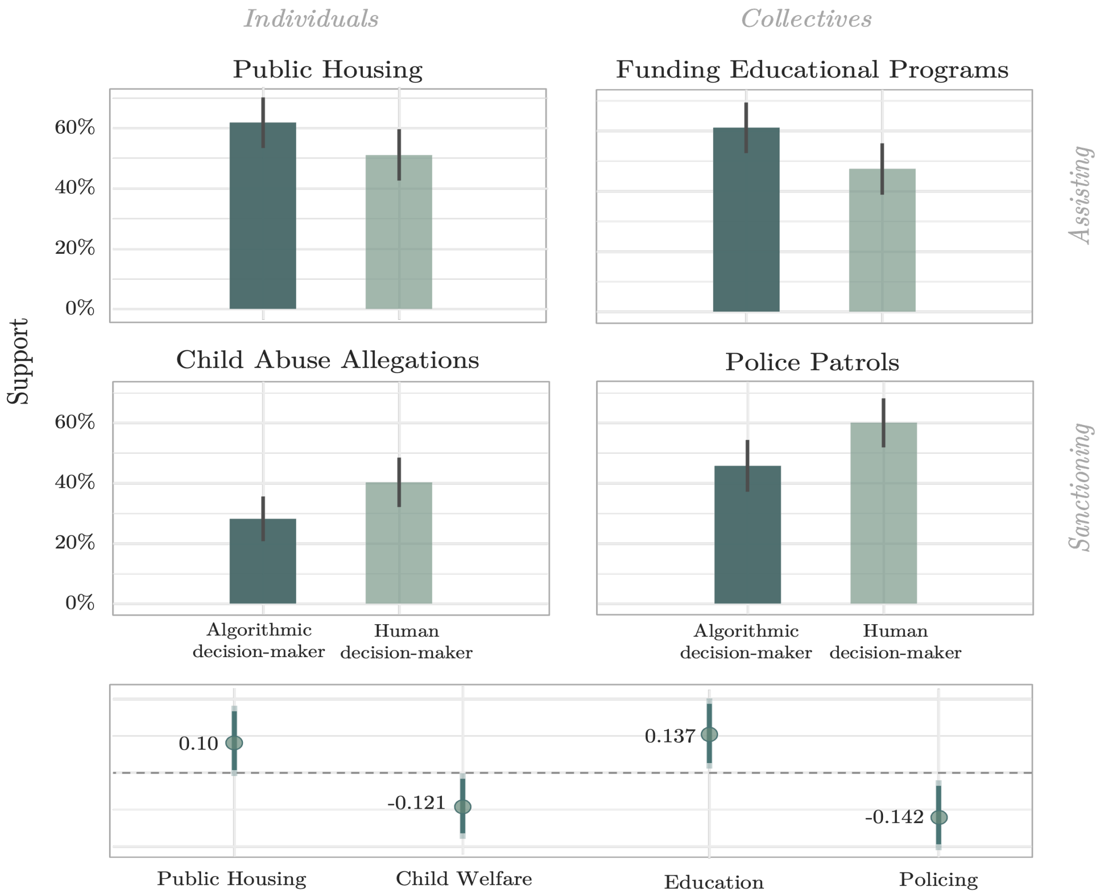
goal of the experiment is to test the theory, which contrasts algorithmic decision-makers with human decision-makers. Yet a prominent form in which is currently used is as tools to assist human decision-makers in "hybrid" decision-making process. To reflect this common practice, I also included in the experiment a second hybrid treatment condition in which a human is assisted by an ADS. The full results are reported in the Appendix C.3.1 and will be discussed later in more detail.

Results: Effect of decision-maker on policy support

Do ADS affect public support for policy decisions? I estimate the average treatment effects of the decision-maker treatment on support for the four policy proposals. To identify the initial reactions and avoid spillover effects, in which respondents evaluate the subsequent policy proposals compared to the decision makers presented in the previous scenario, the primary analysis is based only on data collected from the first scenario that was presented to participants. This means that the analysis is based on a between-subjects design in which both the identity of the decision maker and the policy area were varied in a random fashion. Table A-18 replicates the analysis with data from all scenarios, controlling for the order of the scenarios. The results remain in the same direction, but the magnitude of effects is somewhat weaker.

Figure 5 shows the percentage of respondents who support each policy proposal as a function of the decision-maker treatment: human versus algorithmic. Consistent with the theoretical expectations, the results show that people do not respond uniformly to the use of ADS in public policy. Respondents who were presented with an algorithmic decision-maker were, on average, 14 percentage points less likely to support the proposal to allocate police forces to patrol than those who were presented with a human decision-maker ($p<0.05$). This effect is both statistically and substantively significant, decreasing support from 60% to less than 46%.

Figure (5) Average policy support, by decision-maker and context



Notes: The figure shows average support for each proposal as a function of the decision-maker condition. The sample includes responses to the first scenario. Error bars indicate 95% CI. The bottom panel shows the results of LPMs, without controls, studying the effect of an ADS on the likelihood of supporting each proposal. Thick bars represent 90% CI; thin bars represent 95% CI. The full results, reported in A-16 based on responses collected from the first scenario.

The bottom panel of Figure 5 shows a similar negative effect in the context of child welfare– involved decisions designed to sanction individuals rather than collectives. Respondents were 12 percentage points less likely to support the proposal for choosing which child abuse allegation to investigate when an algorithm assesses the risk of child abuse or neglect in the family (p<0.05).

In the other two proposals that involved assisting rather than sanctioning decisions– prioritizing public housing and allocating funds to education programs–we see very different trends. The results show that respondents were almost indifferent to ADS in the context of assisting individuals. If anything, using a predictive algorithm instead of public housing officials increases the probability of supporting the proposal to prioritize housing based on disability rather than the time spent on the waiting list.

I find even stronger treatment effects in the context of assisting collectives with regard to the proposal to choose specific schools to receive funding for drug and alcohol education programs. The percentage of citizens expressing at least some support for this policy significantly increases by almost 14 percentage points when a predictive algorithm, rather than members of the school board, assesses the risk of juvenile crime in the area ($p<0.05$). To get a better sense of the substantive size of this effect, Table A-18 reports the effect of ADS on support for the policy, adjusting for sociodemographic factors. The table shows that the treatment effect is equal to the partisan difference in policy support between Democrats and Republicans. Taken together, the results suggest that citizens are especially susceptible to the use of ADS in decisions about sanctions which often entail irreversible consequences. Yet, they are more open to accept ADS when used to assist individuals and even are more supportive in decisions that assist collectives.[24]

I also assess the possibility that using algorithmic systems to assist rather than replace human decision-makers might have a different effect on public support. Table A-17 estimates the average treatment effect (ATE) while comparing the ADS and the hybrid conditions, showing little difference between these two conditions across policy domains. The sole exception I find to this pattern is policing. Interestingly, while the use of a predictive algorithm alone has a significant negative effect on support relative to the human decision-maker condition, support for this policy significantly increases when the predictive algorithm is used as a support tool ($p<0.01$). This is consistent with evidence of a trade-off that people face in using ADS in decisions that sanction collectives, which are considered less fair but relatively

[24]Table A-19 results are of a similar magnitude when using the alternative outcome measures.

accurate. It seems that in decisions of this type, using algorithms as a supportive tool, while keeping the "human in the loop," appears to be an attractive solution, as it provides more accurate assessments, without sacrificing the human element that is important in decisions with irreversible consequences.

Finally, to test whether the policy context moderates the effect of ADS, Table A-20 in the Appendix examines the interaction effects of ADS and the policy proposal (presented first to the respondent) on the probability of supporting the policy. The effect of ADS is negative and significant, suggesting that overall, respondents are less likely to support policies implemented by ADS. However, the negative effect of the decision maker is offset and even reversed in policy proposals involving decisions about assisting collectives. Overall, the experimental results provide support for the theory, suggesting that people are particularly sensitive to human presence in sanctioning decisions, which have less reversible consequences for the lives of either individuals or collectives. Adopting ADS in these contexts can significantly reduce the overall support for the policy decisions and actions they implement.

Conclusion and Implications

This article puts forward a theoretical framework and leverages a set of survey experiments to explain public attitudes and preferences on the use of ADS in government. The theory calls for distinguishing between four types of decisions when contemplating ADS uses. The experimental results provide strong support for this theory. Using evidence from a broad range of decisions and policy issues, I show that citizens resist the use of ADS in decisions that sanction, especially individuals, but are more willing to accept the use of these systems in decisions that assist especially collectives. Returning to the California referendum example, the analysis suggests that the public rejection of replacing cash bail with a risk assessment algorithm reflects citizens' sensitivity to the specific use of ADS in sanctioning decisions that have less reversible consequences for individuals' lives. As this study suggests, the use of ADS in this context is perceived both less fair and less accurate compared to human decision-makers.

The finding that the same algorithmic systems can be accepted as legitimate in certain decision areas but rejected in others underscores the limitations of recent public and private attempts to articulate a single prescription for a fair and accurate algorithmic system. Efforts to define the ethical principles that should guide the development and regulation of AI often do not consider the public views. Yet, as this study shows, even if engineers and ethicists were to agree on how ADS should be operated, it would have limited value if they are rejected by the citizens-—those who will have to live with and accept the superiority of ADS. Specifically, the fact that citizens have no freedom to choose whether and when to rely on the algorithm's output may generate substantial backlash. Indeed, the analysis reveals that public support for decisions designed to sanction individuals fell significantly when made by an algorithm rather than by a human decision-maker. This finding has important policy implications, as attested by recent high-profile examples of governments and municipalities backtracking or canceling initiatives that use ADS due to public opposition (e.g., Austen and Wakabayashi, 2020; Weale and Stewart, 2020).

Furthermore, the study provides insight into the feasibility of hybrid solutions that integrate human input and algorithms. The skepticism toward ADS, both in terms of fairness and accuracy, when used to sanction individuals suggests that the use of ADS as support tools may also be met with public scrutiny. The results indicate that citizens do not differentiate between *replacing* and *supporting* human decision-makers with an ADS when evaluating the policy of choosing which child abuse allegations to investigate. In contrast, in decisions that sanction collectives, where ADS accuracy outweighs fairness, using algorithmic output to support, rather than replace, human decision-makers may be an acceptable solution. These findings highlight the need for future research to delve deeper into the dynamics of human-algorithm collaboration in public policy.

This study adopted a broad definition of ADS, focusing on predictive software that relies on extensive data to make decisions without direct human instruction. This simplification aligns with current public understanding of AI-based algorithms and allows for a clear focus on the contextual factors influencing public attitudes. However, the algorithmic systems used in the public sector vary significantly in design and technical features, such as the size and source of training data and the number of factors considered. This raises the question of how these technical features interact with contextual factors. For example, while previous research suggests that people perceive algorithms trained on larger datasets as more reliable (Waggoner et al., 2019), the findings indicate this may not hold true for all types of decisions. In decisions involving sanctions on individuals, technical features ensuring accountability may be prioritized over data size. Therefore, further research is needed to explore the interplay between ADS technical features and the specific contexts in which they are used.

Moreover, variation in public views across contexts is more nuanced than the 2-by-2 framework introduced in this paper. The two dimensions are not all-encompassing, but they provide a useful starting point for further investigation of other relevant contextual factors shaping public preference for using ADS. For example, while this study focuses on whether the algorithmic decision targets individuals or collective cases, another factor that might be relevant is the way the target population is perceived-—whether they are seen as deserving or undeserving of assistance, or perceived as threatening or non-threatening when it comes to sanctioning decisions (Schneider and Ingram, 1993).

Finally, this study documented mass preferences at a relatively early stage of public debate, at a time when most citizens are just becoming aware of ADS and their increasing role in informing high-stakes decisions. As the use of algorithmic tools in government continues to grow, more stakeholders - including technology companies, politicians, and civil society organizations - will seek to inform the public about the potential impact of this technology. Whether and how citizens' views shift in response to new information and the extent to which they rely on cues from elite actors is a promising avenue for future study to understand the evolving politics of using AI and data-driven algorithms in government.

Overall, as this study makes clear, the growing use of AI and ADS in government touches on the very core of democracy—how we make public decisions. As such, it raises questions regarding the legitimacy and accountability of these decisions, inspiring a research agenda in political science on the political repercussions of this major technological change.

References

Alkhatib, Ali and Michael Bernstein (2019). "Street-level algorithms: A theory at the gaps between policy and decisions". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–13.

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.

Austen, Ian and Daisuke Wakabayashi (2020). "Google sibling abandons ambitious city of the future in Toronto". In: *The New York Times*.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan (2017). "Fairness in machine learning". In: *Nips tutorial* 1, p. 2017.

Berlin, Isaiah (1969). "Two concepts of liberty". In: *Four essays on liberty* 118, p. 172.

Binns, Reuben (2019). "Human Judgment in algorithmic loops: Individual justice and automated decision-making". In: *Regulation & Governance*.

Brauneis, Robert and Ellen P Goodman (2018). "Algorithmic transparency for the smart city". In: *Yale JL & Tech.* 20, p. 103.

Brayne, Sarah and Angèle Christin (2021). "Technologies of crime prediction: The reception of algorithms in policing and criminal courts". In: *Social Problems* 68.3, pp. 608–624.

Cuccaro-Alamin, Stephanie, Regan Foust, Rhema Vaithianathan, and Emily Putnam-Hornstein (2017). "Risk assessment and decision making in child protective services: Predictive risk modeling in context". In: *Children and Youth Services Review* 79, pp. 291–298.

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso (2011). "Extraneous factors in judicial decisions". In: *Proceedings of the National Academy of Sciences* 108.17.

Dietvorst, Berkeley J, Joseph P Simmons, and Cade Massey (2018). “Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them”. In: *Management Science* 64.3, pp. 1155–1170.

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

Ferguson, Andrew Guthrie (2017). *The rise of big data policing*. New York University Press.

Gallego, Aina and Thomas Kurer (2022). “Automation, Digitalization, and AI in the workplace: Implications for Political Behavior”. In.

Green, Ben and Yiling Chen (2019). “The principles and limits of algorithm-in-the-loop decision making”. In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW, pp. 1–24.

Helberger, Natali, Theo Araujo, and Claes H de Vreese (2020). “Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making”. In: *Computer Law & Security Review* 39, p. 105456.

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

Horowitz, Michael C and Lauren Kahn (2024). “Bending the Automation Bias Curve: A Study of Human and AI-Based Decision Making in National Security Contexts”. In: *International Studies Quarterly* 68.2, sqae020.

Horowitz, Michael C, Lauren Kahn, Julia Macdonald, and Jacquelyn Schneider (2023). “Adopting AI: how familiarity breeds both trust and contempt”. In: *AI & society*.

International, Amnesty (2021). “Dutch Childcare Benefit Scandal an Urgent Wake-up Call to Ban Racist Algorithms”. In.

Kennedy, Ryan P, Philip D Waggoner, and Matthew M Ward (2022). “Trust in public policy algorithms”. In: *The Journal of Politics* 84.2, pp. 000–000.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan (2016). “Inherent trade-offs in the fair determination of risk scores”. In: *arXiv preprint arXiv:1609.05807*.

Lee, Min Kyung (2018). “Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management”. In: *Big Data & Society* 5.1.

Lepri, Bruno, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck (2018). “Fair, transparent, and accountable algorithmic decision-making processes”. In: *Philosophy & Technology* 31.4, pp. 611–627.

Lipsky, Michael (2010). *Street-level bureaucracy: Dilemmas of the individual in public service*. Russell Sage Foundation.

Logg, Jennifer M, Julia A Minson, and Don A Moore (2019). "Algorithm appreciation: People prefer algorithmic to human judgment". In: *Organizational Behavior and Human Decision Processes* 151, pp. 90–103.

Lowi, Theodore J (1964). "American business, public policy, case-studies, and political theory". In: *World politics* 16.4, pp. 677–715.

Malhotra, Neil, Benoît Monin, and Michael Tomz (2019). "Does private regulation preempt public regulation?" In: *American Political Science Review* 113.1, pp. 19–37.

Meijer, Albert, Lukas Lorenz, and Martijn Wessels (2021). "Algorithmization of bureaucratic organizations: Using a practice lens to study how context shapes predictive policing systems". In: *Public Administration Review* 81.5, pp. 837–846.

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.

O'Shaughnessy, Matthew R, Daniel S Schiff, Lav R Varshney, Christopher J Rozell, and Mark A Davenport (2023). "What governs attitudes toward artificial intelligence adoption and governance?" In: *Science and Public Policy* 50.2, pp. 161–176.

Pasquale, Frank (2015). *The black box society.* Harvard University Press.

Pierson, Paul (1993). "When effect becomes cause: Policy feedback and political change". In: *World politics* 45.4, pp. 595–628.

Pislar, Yevgeniy P and Rachel Puleo (2020). "Proposition 25: Replace Cash Bail with Risk Assessment Referendum". In: *California Initiative Review (CIR)* 2020.1, p. 13.

Read, Blair, Lukas Wolters, and Adam J Berinsky (2021). "Racing the clock: Using response time as a proxy for attentiveness on self-administered surveys". In: *Political Analysis*, pp. 1–20.

Reich, Rob, Mehran Sahami, and Jeremy M Weinstein (2020). *System Error - Where big tech went wrong and how we cam reboot.* Harper.

Robertson, Samantha, Tonya Nguyen, and Niloufar Salehi (2021). "Modeling Assumptions Clash with the Real World: Transparency, Equity, and Community Challenges for Student Assignment Algorithms". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–14.

Sainato, Michael and Vivian Chiu (2021). "LAPD's predictive policing experiments were supposed to reform the system. Did they?" In: *The Guardian.*

Schiff, Daniel S, Kaylyn Jackson Schiff, and Patrick Pierson (2021). "Assessing public value failure in government adoption of artificial intelligence". In: *Public Administration.*

Schiff, Daniel, Justin Biddle, Jason Borenstein, and Kelly Laas (2020). “What’s next for ai ethics, policy, and governance? a global overview”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 153–158.

Schiff, Kaylyn Jackson, Daniel S Schiff, Ian T Adams, Joshua McCrain, and Scott M Mourtgos (2023). “Institutional factors driving citizen perceptions of AI in government: Evidence from a survey experiment on policing”. In: *Public Administration Review*.

Schneider, Anne and Helen Ingram (1993). “Social construction of target populations: Implications for politics and policy”. In: *American political science review* 87.2.

Sunstein, Cass R (2019). “Algorithms, correcting biases”. In: *Social Research: An International Quarterly* 86.2, pp. 499–511.

Tyler, Tom R (2006). “Psychological perspectives on legitimacy and legitimation”. In: *Annu. Rev. Psychol.* 57, pp. 375–400.

Waggoner, Philip D, Ryan Kennedy, Hayden Le, and Myriam Shiran (2019). “Big Data and Trust in Public Policy Automation”. In: *Statistics, Politics and Policy* 10.2, pp. 115–136.

Waggoner, Philip and Ryan Kennedy (2022). “The Role of Personality in Trust in Public Policy Automation”. In: *Journal of Behavioral Data Science* 2.1, pp. 106–123.

Weale, S and H Stewart (2020). “A-level and GCSE results in England to be based on teacher assessments in U-turn”. In: *The Guardian*.

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

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

Young, Matthew M, Justin B Bullock, and Jesse D Lecy (2019). “Artificial discretion as a tool of governance: a framework for understanding the impact of artificial intelligence on public administration”. In: *Perspectives on Public Management and Governance* 2.4, pp. 301–313.

Zhang, Baobao and Allan Dafoe (2019). “Artificial intelligence: American attitudes and trends”. In: *Available at SSRN 3312874*.

Online Appendix for:
"When Do Citizens Resist The Use of AI Algorithms in Public Policy
Theory and Evidence"

Contents

A Data Description

A.1 Descriptive Statistics

As mentioned in the main text, I conducted two original experiments embedded in a national U.S. survey, using samples of 1,590 adults collected by Dynata an Internet survey company (formerly Survey Sampling International - SSI). I imposed quotas on gender, age, education, and race/ethnicity. Table A-1 reports the characteristics of the sample compared with those of the overall US population. Data come from the 2020 U.S. Census Bureau. As the table shows, the sample is representative along the quota dimensions.

Table (A-1) Summary statistics for the survey sample

Quotas	Population Percent	Sample N	Sample Percent
Age			
18-24	13	206	13.0
25-34	19	295	18.6
35-44	22	357	22.6
45-54	18	271	17.1
55+	28	453	28.6
Race			
White	60.7	976	63.5
Hispanic	18.1	252	16.4
Black	13.4	219	14.2
Asian	7.8	90	5.9
Gender			
Male	48	761	48.1
Female	52	821	51.9
Education			
High school diploma or less	40	648	41.0
Some college	19	301	19.0
Associate's degree or	29.7	460	29.1
Bachelor's or Graduate degree	11.3	173	10.9

A.2 Survey Questionnaire

This section gives the exact wording of the policy scenarios, the experimental conditions, and the survey questions included in the survey. As mentioned, to minimize the potential concern of social desirability, respondents were not informed about the study's focus on using ADSs. Instead, they were asked about their views on four policy proposals.

Definitions: Before beginning, please read these definitions that are relevant to the policies: (1) The pretrial stage in the criminal justice system is the time between arrest and trial. (2) A predictive algorithm is computer software that makes decisions without human instruction, relying on a massive amount of data. (3) Homelessness is defined as living somewhere that is below a minimum quality standard or that you can be evicted from with little or no warning.

Attention check 1 (screener): On the previous page, you were presented with three definitions that are relevant to the policies. Please select the definition that did not appear among the previous three definitions: (4) Screeners are workers in child welfare who respond to the hotline calls reporting child abuse allegations.

Attention check 2 (screener): People are very busy these days and many do not have time to follow what goes on in the government. We are testing whether people read questions. To show that you've read this much, answer both extremely interested and very interested.

Outcome questions: Decision-context experiment)
Perceived Appropriateness: Next, we ask that you read the descriptions of several policy decisions. For each, please indicate how appropriate it is to have that decision made by an algorithm rather than by a human being. Extremely appropriate 1 to Extremely inappropriate 7.

Perceived Accuracy and Fairness: People think differently about the extent to which algorithms would be accurate compared to fair. In some decisions, an algorithm may be considered accurate but unfair; in other decisions fair but inaccurate; and in some other decisions unfair and inaccurate; or both fair and accurate. For the same decisions you have just evaluated. Please indicate how FAIR/ACCURATE and ACCURATE/FAIR you think an algorithm would be in...

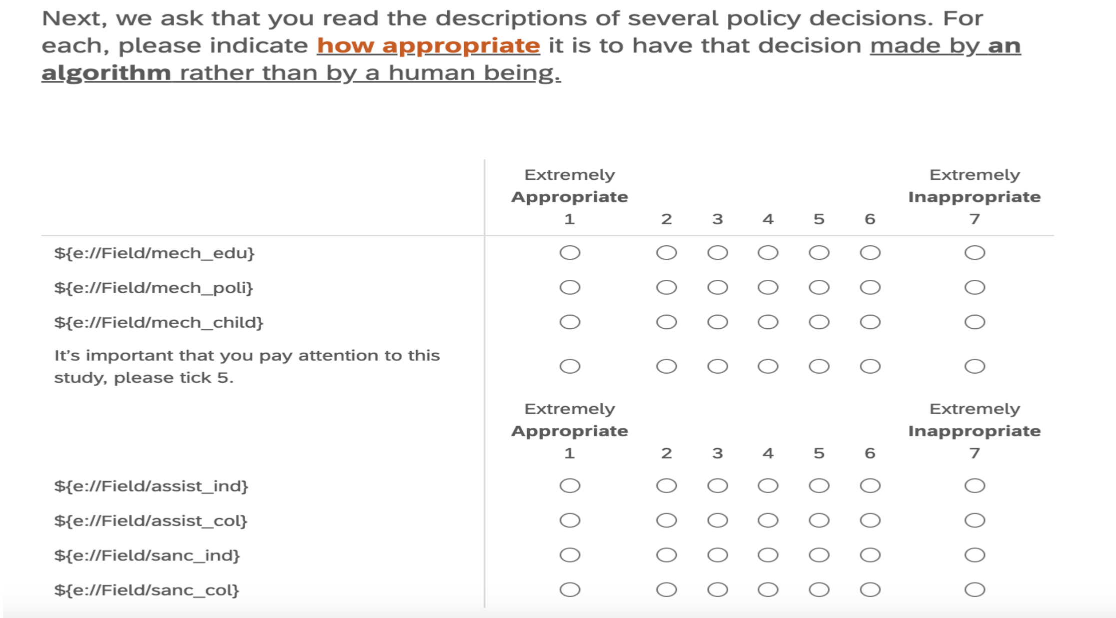
Technological Literacy: How familiar are you with the following computer and Internet-related items? Please choose a number between 1 and 5 where 1 represents "Totally unfamiliar," and 5 represents "Very familiar". Phishing; Cache; PDF; Tagging; JPEG; Malware; RSS; HTTP cookie; Fitibly.

Figures A-1 and A-2 provide a screenshot of the matrix presented to the respondent in the decision context experiment.

A.3 Decision-Context Experiment: Treatment Wording

Table 2 in the main text details the wording of the treatment conditions randomized in the first (between-subjects) component of the decision-context experiment: four types of decisions within each of three policy domains—policing, child welfare, and education. Table A-2 below details the treatment conditions randomized in the second (within-subjects) component: two different issue areas for each of the four decision types.

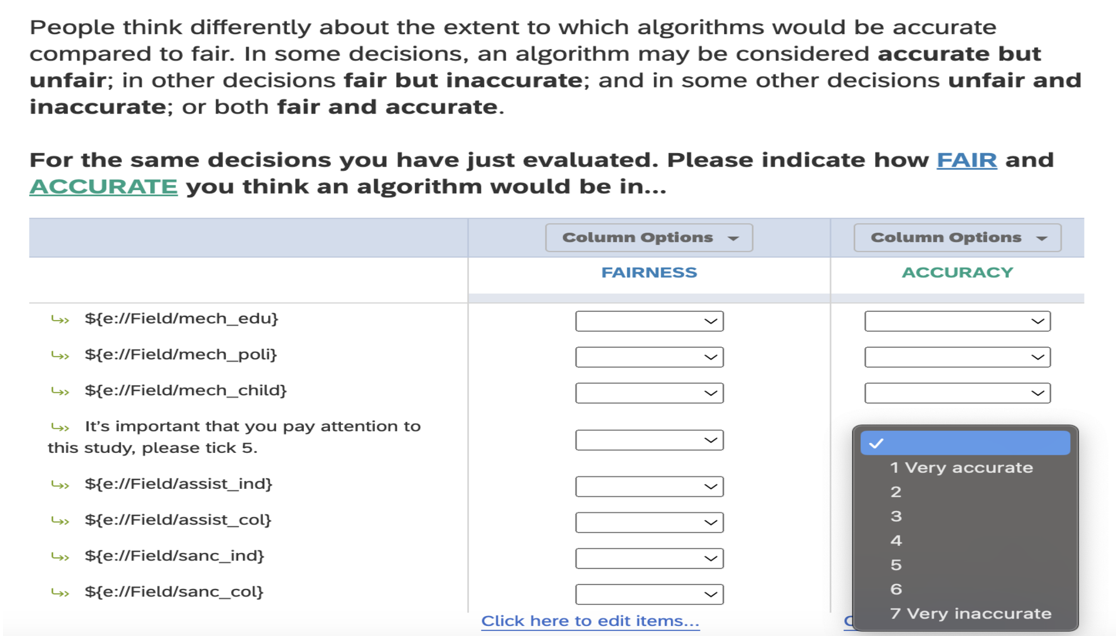
Figure (A-1) Screenshot of the Decision-Context Experiment: Perceived Appropriateness



Next, we ask that you read the descriptions of several policy decisions. For each, please indicate **how appropriate** it is to have that decision **made by an algorithm** rather than by a human being.

	Extremely **Appropriate** 1	2	3	4	5	6	Extremely **Inappropriate** 7
${e://Field/mech_edu}	○	○	○	○	○	○	○
${e://Field/mech_poli}	○	○	○	○	○	○	○
${e://Field/mech_child}	○	○	○	○	○	○	○
It's important that you pay attention to this study, please tick 5.	○	○	○	○	○	○	○
	Extremely **Appropriate** 1	2	3	4	5	6	Extremely **Inappropriate** 7
${e://Field/assist_ind}	○	○	○	○	○	○	○
${e://Field/assist_col}	○	○	○	○	○	○	○
${e://Field/sanc_ind}	○	○	○	○	○	○	○
${e://Field/sanc_col}	○	○	○	○	○	○	○

Figure (A-2) Screenshot of the Decision-Context Experiment: Perceived Accuracy vs. Fairness



People think differently about the extent to which algorithms would be accurate compared to fair. In some decisions, an algorithm may be considered **accurate but unfair**; in other decisions **fair but inaccurate**; and in some other decisions **unfair and inaccurate**; or both **fair and accurate**.

For the same decisions you have just evaluated. Please indicate how FAIR and ACCURATE you think an algorithm would be in...

	Column Options FAIRNESS	Column Options ACCURACY
${e://Field/mech_edu}		
${e://Field/mech_poli}		
${e://Field/mech_child}		
It's important that you pay attention to this study, please tick 5.		✓ / 1 Very accurate / 2 / 3 / 4 / 5 / 6 / 7 Very inaccurate
${e://Field/assist_ind}		
${e://Field/assist_col}		
${e://Field/sanc_ind}		
${e://Field/sanc_col}		

Click here to edit items...

Table (A-2) Decision Wordings Randomized in the Within-Subjects Component

	Decisions
Assisting individuals	Deciding which individual should receive food stamps based on an assessment of the neediness of the requester. Deciding which pupils should be offered study assistance based on an assessment of early learning problems in school.
Assisting collectives	Deciding where to build shelters for homeless people based on an assessment of the risk of homelessness in the area. Deciding where to place fire stations based on a prediction of the risk of fire outbreaks nearby.
Sanctioning individuals	Determining a sentence based on an assessment of the defendant's risk of committing another crime. Determining whether a restraining order should be issued based on a prediction of the individual's risk of assaulting their intimate partner.
Sanctioning collectives	Deciding where to increase police enforcement based on an assessment of the likelihood of illegal work in the area. Deciding where to increase police enforcement based on an assessment of the likelihood of illegal building in the area.

Notes: This table details the treatment conditions included in the within-subject experiment. All respondents were presented with the four types of decisions, where the manipulation is in the issue area.

B Decision-Context Experiment

B.1 Descriptive Statistics

Table A-5 below presents descriptive statistics for the main dependent variables: the perceived appropriateness, fairness, and accuracy of using algorithmic decision systems (ADS) across the three policy domains, by the four types of decisions randomly assigned to respondents.

B.2 Decision-Context Experiment: Between-Subjects Component Results

This section provides the demographic balance tables for the between-subjects experiment. The tables below show the results of t-tests of each treatment condition for each policy domain. Results confirm that the randomization of treatment assignment makes the four groups essentially identical to one another on average.

B.3 Between-Subjects Results

Table A-6 below reports estimates from linear probability models studying the effect of (1) the subject of the decision and (2) its objective on the probability of viewing the use of ADS as appropriate across three policy areas: education, policing, and child welfare. The analysis limits the sample to responses to the first item, which provides the cleanest comparison. In the first model for each policy area (columns 1, 4, and 7), reports minimal specifications. The second model for each area (columns 2, 5, and 8) includes demographic controls: age, Table (A-3) Summary of statistics of Perceived Accuracy, Fairness, and Appropriateness

Consideration	**Context**	**Type**	**Mean**	**n**	**SD**	**SE**
Appropriateness	Education	Assisting Collectives	3.66	397	2.16	0.11
		Assisting Individuals	3.51	396	2.10	0.11
		Sanctioning Collectives	3.51	391	2.13	0.11
		Sanctioning Individuals	2.78	398	1.82	0.09
	Policing	Assisting Collectives	4.48	386	2.20	0.11
		Assisting Individuals	2.90	399	1.89	0.09
		Sanctioning Collectives	3.64	403	2.12	0.11
		Sanctioning Individuals	2.97	394	1.92	0.10
	Child Welfare	Assisting Collectives	0.48	397	0.50	0.03
		Assisting Individuals	0.35	395	0.48	0.02
		Sanctioning Collectives	0.30	401	0.46	0.02
		Sanctioning Individuals	0.18	389	0.38	0.02
Fairness	Education	Assisting Collectives	4.37	397	1.85	0.09
		Assisting Individuals	4.12	396	1.95	0.10
		Sanctioning Collectives	3.96	391	1.91	0.10
		Sanctioning Individuals	3.25	398	1.76	0.09
	Policing	Assisting Collectives	4.87	386	1.80	0.09
		Assisting Individuals	3.57	399	1.84	0.09
		Sanctioning Collectives	4.28	403	1.84	0.09
		Sanctioning Individuals	3.49	394	1.91	0.10
	Child Welfare	Assisting Collectives	4.38	397	1.83	0.09
		Assisting Individuals	3.79	395	1.84	0.09
		Sanctioning Collectives	3.89	401	1.73	0.09
		Sanctioning Individuals	3.15	389	1.91	0.10
Accuracy	Education	Assisting Collectives	4.51	397	1.81	0.09
		Assisting Individuals	4.15	396	1.78	0.09
		Sanctioning Collectives	4.53	391	1.70	0.09
		Sanctioning Individuals	3.86	398	1.78	0.09
	Policing	Assisting Collectives	4.89	386	1.73	0.09
		Assisting Individuals	3.45	399	1.80	0.09
		Sanctioning Collectives	4.62	403	1.74	0.09
		Sanctioning Individuals	3.76	394	1.86	0.09
	Child Welfare	Assisting Collectives	4.32	397	1.79	0.09
		Assisting Individuals	3.66	395	1.77	0.09
		Sanctioning Collectives	4.08	401	1.68	0.08
		Sanctioning Individuals	3.26	389	1.87	0.09

Table (A-4) Summary of statistics of binary outcomes

Decision Type	Policy Domain	Consideration		
		Appropriateness	**Fairness**	**Accuracy**
Sanction Individuals	Education	0.166 (0.018)	0.234 (0.021)	0.367 (0.024)
	Child welfare	0.177 (0.019)	0.247 (0.021)	0.247 (0.021)
	Policing	0.218 (0.020)	0.289 (0.022)	0.340 (0.023)
Assist Individuals	Policing	0.246 (0.021)	0.333 (0.023)	0.288 (0.022)
	Child welfare	0.347 (0.024)	0.387 (0.024)	0.291 (0.022)
	Education	0.384 (0.024)	0.447 (0.025)	0.409 (0.024)
Sanction Collectives	Child welfare	0.302 (0.023)	0.362 (0.024)	0.406 (0.024)
	Education	0.335 (0.023)	0.373 (0.024)	0.527 (0.025)
	Policing	0.367 (0.024)	0.444 (0.024)	0.558 (0.024)
Assist Collectives	Education	0.378 (0.024)	0.474 (0.025)	0.509 (0.025)
	Child welfare	0.479 (0.025)	0.496 (0.025)	0.491 (0.025)
	Policing	0.560 (0.025)	0.598 (0.025)	0.593 (0.025)

gender, political identification, education level, race, and digital literacy. The third model for each area (columns 3, 6, and 9) replaces digital literacy with prior knowledge of AI. This measure captures respondents' familiarity with the increasing use of algorithmic systems in public decision-making, aligning more closely with the knowledge domain relevant to this research.

Table A-7 shows that the results are robust when using the full sample, controlling for the order of presentation of the three policy domains.

Table A-8 below presents robustness checks of the main findings, demonstrating consistent results when controlling for respondent attentiveness. Two measures of attentiveness are employed: (1) whether respondents passed an attention check embedded within the decision-context experiment matrix (Models 2, 5, 8), and (2) response time per question, accounting for both those who rushed through the survey and those who may have been distracted (Models 1, 4, 7). Models 3, 6, and 9 replicate the results using logistic regression instead of LPMs for further robustness.

Table A-9 replicates the results using mixed-effects linear regressions. I estimate three separate models for alternative measures of the outcomes: Model 1 estimates the treatment effects on the main outcome of perceived appropriateness using the top three categories on the 7-point scale; As preregistered, I replicate the results, using the following alternative measures of the outcome variables: (1) a binary measure with a value of 1 for the last three options indicating "appropriate" and 0 otherwise (see Model 2); (2) continuous outcomes of a seven-point scale, with higher values indicating very appropriate (Model 3). The results are very much consistent with the main findings.

Table (A-5) Balance tests

Variable	Assisting Collectives		Assisting Individuals		Sanctioning Collectives		Sanctioning Individuals		Test
	N	**Percent**	**N**	**Percent**	**N**	**Percent**	**N**	**Percent**	
				Domain: Policing					
Gender	118		126		140		151		**X2=1.897**
Male	59	50%	69	54.80%	79	56.40%	75	49.70%	
Female	59	50%	57	45.20%	61	43.60%	76	50.30%	
Age	118		126		140		151		**X2=16.88**
18 - 24	11	9.30%	13	10.30%	22	15.70%	15	9.90%	
25 - 34	25	21.20%	23	18.30%	22	15.70%	36	23.80%	
35 - 44	25	21.20%	30	23.80%	32	22.90%	29	19.20%	
45 - 54	24	20.30%	22	17.50%	32	22.90%	25	16.60%	
55 - 64	7	5.90%	18	14.30%	12	8.60%	21	13.90%	
65 +	26	22%	20	15.90%	20	14.30%	25	16.60%	
Education	118		126		140		151		**X2=0.153**
Associate's or higher degree	48	40.70%	52	41.30%	56	40%	59	39.10%	
Some college or less	70	59.30%	74	58.70%	84	60%	92	60.90%	
Race	118		126		140		151		**X2=2.373**
Non White	41	34.70%	53	42.10%	61	43.60%	63	41.70%	
White	77	65.30%	73	57.90%	79	56.40%	88	58.30%	
Tech Literacy	118		126		140		151		**X2=2.615**
Low Literacy	72	61%	83	65.90%	81	57.90%	99	65.60%	
High Literacy	46	39%	43	34.10%	59	42.10%	52	34.40%	
				Domain: Education					
Gender	123		137		124		124		
Male	59	48%	58	42.30%	48	38.70%	65	52.40%	
Female	64	52%	79	57.70%	76	61.30%	59	47.60%	
Age	123		137		124		124		**X2=9.557**
18 - 24	16	13%	20	14.60%	21	16.90%	20	16.10%	
25 - 34	25	20.30%	23	16.80%	21	16.90%	23	18.50%	
35 - 44	24	19.50%	29	21.20%	30	24.20%	23	18.50%	
45 - 54	19	15.40%	19	13.90%	19	15.30%	24	19.40%	
55 - 64	10	8.10%	19	13.90%	16	12.90%	12	9.70%	
65 +	29	23.60%	27	19.70%	17	13.70%	22	17.70%	
Education	123		137		124		124		**X2=2.246**
Associate's or higher degree	48	39%	56	40.90%	40	32.30%	47	37.90%	
Some college or less	75	61%	81	59.10%	84	67.70%	77	62.10%	
Race	123		137		124		124		**X2=0.218**
Non White	45	36.60%	47	34.30%	43	34.70%	45	36.30%	
White	78	63.40%	90	65.70%	81	65.30%	79	63.70%	
Tech Literacy	123		137		124		124		**X2=1.155**
Low Literacy	88	71.50%	91	66.40%	86	69.40%	82	66.10%	
High Literacy	35	28.50%	46	33.60%	38	30.60%	42	33.90%	
				Domain: Child Welfare					
Gender	129		143		138		129		
Male	60	46.50%	66	46.20%	65	47.10%	58	45%	
Female	69	53.50%	77	53.80%	73	52.90%	71	55%	
Age	129		143		138		129		**X2=16.018**
18 - 24	16	12.40%	27	18.90%	16	11.60%	9	7%	
25 - 34	27	20.90%	26	18.20%	23	16.70%	21	16.30%	
35 - 44	28	21.70%	32	22.40%	36	26.10%	39	30.20%	
45 - 54	20	15.50%	17	11.90%	27	19.60%	23	17.80%	
55 - 64	16	12.40%	14	9.80%	16	11.60%	12	9.30%	
65 +	22	17.10%	27	18.90%	20	14.50%	25	19.40%	
Education	129		143		138		129		**X2=2.357**
Associate's or higher degree	49	38%	59	41.30%	65	47.10%	54	41.90%	
Some college or less	80	62%	84	58.70%	73	52.90%	75	58.10%	
Race	129		143		138		129		**X2=0.174**
Non White	51	39.50%	56	39.20%	53	38.40%	48	37.20%	
White	78	60.50%	87	60.80%	85	61.60%	81	62.80%	
Tech Literacy	129		143		138		129		**X2=2.364**
Low Literacy	85	65.90%	88	61.50%	84	60.90%	73	56.60%	
High Literacy	44	34.10%	55	38.50%	54	39.10%	56	43.40%	

Table (A-6) Decision context and perceived appropriateness, by policy domain

	Dependent variable:								
	Education			Policing			Child welfare		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Assisting	0.139***	0.138***	0.142***	0.114**	0.103**	0.095*	0.188***	0.187***	0.182***
	(0.040)	(0.040)	(0.039)	(0.039)	(0.038)	(0.037)	(0.039)	(0.038)	(0.038)
Individuals	−0.099*	−0.095*	−0.104**	−0.279***	−0.280***	−0.272***	−0.099*	−0.095*	−0.096*
	(0.040)	(0.040)	(0.039)	(0.039)	(0.038)	(0.037)	(0.039)	(0.038)	(0.038)
Age		−0.006	−0.007		−0.001	−0.007		0.004	0.002
		(0.013)	(0.013)		(0.013)	(0.013)		(0.013)	(0.013)
Female		−0.022	−0.032		−0.040	−0.054		0.007	−0.013
		(0.040)	(0.040)		(0.038)	(0.038)		(0.038)	(0.038)
Some college or less		−0.012	−0.013		−0.018	−0.022		−0.042	−0.034
		(0.043)	(0.042)		(0.040)	(0.039)		(0.041)	(0.040)
White		0.060	−0.017		0.226***	0.154***		0.114**	0.025
		(0.045)	(0.048)		(0.043)	(0.046)		(0.043)	(0.048)
High tech literacy		−0.114*			−0.107**			−0.098*	
		(0.045)			(0.041)			(0.042)	
Prior Knowledge			−0.263***			−0.236***			−0.247***
			(0.055)			(0.051)			(0.053)
Constant	0.271***	0.307***	0.388***	0.444***	0.389***	0.480***	0.258***	0.231**	0.325***
	(0.035)	(0.079)	(0.079)	(0.034)	(0.074)	(0.076)	(0.033)	(0.077)	(0.078)
Observations	508	508	508	535	535	535	539	539	539
R^2	0.034	0.056	0.086	0.100	0.186	0.207	0.052	0.088	0.116

Notes: This table reports results from LPMs estimated separately for each policy domain. The dependent variable is binary, taking the value of 1 if the respondent deems the use of ADS appropriate in a given policy area and 0 otherwise. The independent variables are indicators for the theoretical dimensions: the target of the decision (assisting or sanctioning) and the object of the decision (individuals or collectives). The reference categories are: collectives and sanctioning. Digital literacy is a binary variable that takes the value 1 if the respondent indicates familiarity with more than 5 items on the matrix of 8 technological-related items. †p<0.1; *p<0.05; **p<0.01; ***p<0.001

Table (A-7) Decision context and appropriateness, full sample controlling for order

	Dependent variable:								
	Education	Policing	Child welfare	Education	Policing	Child welfare	Education	Policing	Child welfare
	Binary (3 last cat)	Binary (2 last cat)	Seven-point scale						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Assisting	0.136^{***}	0.068^{***}	0.490^{***}	0.112^{***}	0.081^{***}	0.392^{***}	0.173^{***}	0.106^{***}	0.715^{***}
	(0.023)	(0.020)	(0.097)	(0.022)	(0.020)	(0.094)	(0.022)	(0.020)	(0.094)
Individuals	-0.084^{***}	-0.098^{***}	-0.466^{***}	-0.227^{***}	-0.214^{***}	-1.096^{***}	-0.126^{***}	-0.097^{***}	-0.699^{***}
	(0.023)	(0.020)	(0.097)	(0.022)	(0.020)	(0.094)	(0.022)	(0.020)	(0.094)
Order	0.011	0.001	0.004	−0.018	0.004	−0.056	0.017	0.005	0.075
	(0.014)	(0.012)	(0.059)	(0.014)	(0.012)	(0.058)	(0.014)	(0.012)	(0.058)
Age	−0.011	-0.016^{*}	-0.092^{**}	0.005	0.006	0.010	0.002	−0.004	−0.021
	(0.008)	(0.007)	(0.033)	(0.007)	(0.007)	(0.032)	(0.008)	(0.007)	(0.031)
Female	−0.032	−0.033	−0.099	-0.049^{*}	-0.047^{*}	−0.098	−0.004	−0.004	−0.025
	(0.023)	(0.020)	(0.098)	(0.022)	(0.020)	(0.095)	(0.023)	(0.020)	(0.094)
Some college or less	-0.050^{*}	−0.031	$-0.193^{\dagger}$	$-0.043^{\dagger}$	−0.033	-0.219^{*}	−0.027	−0.001	−0.140
	(0.024)	(0.021)	(0.104)	(0.024)	(0.022)	(0.101)	(0.024)	(0.021)	(0.101)
White	0.125^{***}	0.099^{***}	0.981^{***}	0.186^{***}	0.111^{***}	1.129^{***}	0.165^{***}	0.102^{***}	0.936^{***}
	(0.026)	(0.023)	(0.110)	(0.025)	(0.023)	(0.107)	(0.025)	(0.022)	(0.106)
Tech literacy	-0.084^{***}	$-0.042^{\dagger}$	-0.981^{***}	-0.098^{***}	-0.063^{**}	-0.969^{***}	-0.086^{***}	-0.049^{*}	-1.041^{***}
	(0.025)	(0.022)	(0.108)	(0.025)	(0.022)	(0.105)	(0.025)	(0.022)	(0.104)
Constant	0.307^{***}	0.263^{***}	3.580^{***}	0.394^{***}	0.267^{***}	3.760^{***}	0.207^{***}	0.154^{***}	3.157^{***}
	(0.054)	(0.047)	(0.230)	(0.052)	(0.047)	(0.218)	(0.053)	(0.046)	(0.219)
Observations	1,582	1,582	1,582	1,582	1,582	1,582	1,582	1,582	1,582
R^2	0.060	0.042	0.149	0.140	0.112	0.232	0.103	0.055	0.198

Notes: This table reports results from LPM estimated separately for each policy area for the full sample, controlling for the presentation order of the item in the matrix. †p<0.1; *p<0.05; **p<0.01; ***p<0.001

Table (A-8) Decision Context and Appropriateness, controlling for Inattentiveness

	Dependent variable:								
	Education			Policing			Child welfare		
	OLS		*logistic*	*OLS*		*logistic*	*OLS*		*logistic*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Constant	0.322^{***}	0.316^{***}	-0.873^{***}	0.403^{***}	0.401^{***}	$-0.502^{\dagger}$	0.215^{***}	0.214^{***}	-1.408^{***}
	(0.054)	(0.054)	(0.265)	(0.052)	(0.052)	(0.266)	(0.053)	(0.052)	(0.269)
Assisting	0.137^{***}	0.137^{***}	0.663^{***}	0.113^{***}	0.113^{***}	0.572^{***}	0.174^{***}	0.172^{***}	0.866^{***}
	(0.023)	(0.023)	(0.113)	(0.022)	(0.022)	(0.115)	(0.022)	(0.022)	(0.115)
Individuals	-0.083^{***}	-0.086^{***}	-0.413^{***}	-0.226^{***}	-0.226^{***}	-1.125^{***}	-0.127^{***}	-0.125^{***}	-0.634^{***}
	(0.023)	(0.023)	(0.112)	(0.022)	(0.022)	(0.117)	(0.022)	(0.022)	(0.114)
Order	0.010	0.011	0.054	−0.018	−0.018	−0.091	0.018	0.018	0.085
	(0.014)	(0.014)	(0.068)	(0.014)	(0.014)	(0.070)	(0.014)	(0.014)	(0.070)
Age	−0.012	−0.011	−0.055	0.004	0.005	0.021	0.002	0.003	0.010
	(0.008)	(0.008)	(0.037)	(0.007)	(0.007)	(0.038)	(0.008)	(0.008)	(0.037)
Female	−0.030	−0.030	−0.165	-0.048^{*}	-0.047^{*}	-0.258^{*}	−0.003	−0.003	−0.026
	(0.023)	(0.023)	(0.113)	(0.022)	(0.022)	(0.115)	(0.023)	(0.023)	(0.114)
Some college or less	$-0.047^{\dagger}$	$-0.043^{\dagger}$	-0.259^{*}	$-0.041^{\dagger}$	−0.038	-0.241^{*}	−0.025	−0.021	−0.146
	(0.024)	(0.024)	(0.120)	(0.024)	(0.024)	(0.123)	(0.024)	(0.024)	(0.122)
White	0.112^{***}	0.114^{***}	0.632^{***}	0.179^{***}	0.179^{***}	0.975^{***}	0.158^{***}	0.157^{***}	0.860^{***}
	(0.026)	(0.026)	(0.130)	(0.026)	(0.025)	(0.135)	(0.026)	(0.026)	(0.134)
Tech literacy	-0.065^{*}	-0.070^{**}	-0.425^{***}	-0.087^{***}	-0.089^{***}	-0.521^{***}	-0.075^{**}	-0.076^{**}	-0.454^{***}
	(0.026)	(0.025)	(0.127)	(0.026)	(0.025)	(0.131)	(0.026)	(0.025)	(0.130)
Inattentive (time)	-0.084^{*}			−0.048			−0.050		
	(0.033)			(0.033)			(0.033)		
Inattentive (matrix)		-0.142^{***}			-0.093^{*}			-0.102^{**}	
		(0.038)			(0.037)			(0.037)	
Observations	1,582	1,582	1,582	1,582	1,582	1,582	1,582	1,582	1,582
R^2	0.064	0.068		0.141	0.143		0.104	0.107	
Log Likelihood			−936.744			−901.647			−913.383
Akaike Inf. Crit.			1,891.488			1,821.295			1,844.766

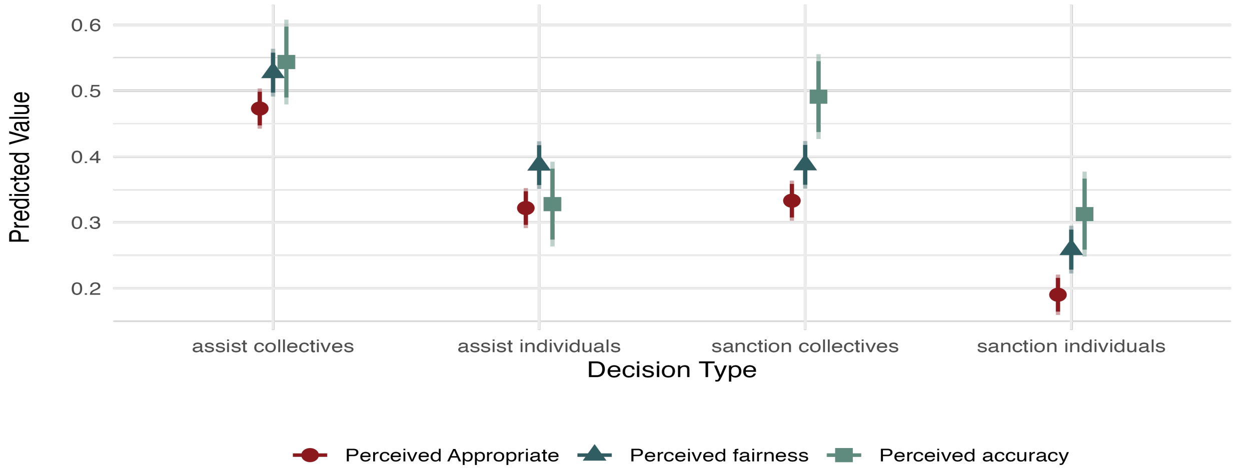
Notes: This table reports results from LPM (models 1-5) or Logistic regression (models 7-9) estimated separately for each policy domain. Models 1, 3, and 5 control for respondents who passed the attention check. Models 1, 4, and 7 control for response-time attentiveness, with respondents who completed the survey quickly as the reference category. Models 2, 5, and 8 control for respondents who failed the attention check. †p<0.1; *p<0.05; **p<0.01; ***p<0.001

Table (A-9) Mixed-Effects Linear Regressions for Perceived Appropriateness

	Dependent variable:		
	Perceived appropriate (3 last categories)	Perceived appropriate (2 last categories)	Perceived appropriate (7-point scale)
	(1)	(2)	(3)
Assisting	0.136^{***}	0.078^{***}	0.521^{***}
	(0.012)	(0.010)	(0.045)
Individuals	-0.147^{***}	-0.136^{***}	-0.746^{***}
	(0.012)	(0.010)	(0.045)
Domain: Education	−0.010	0.008	0.022
	(0.013)	(0.011)	(0.048)
Domain: Policing	$0.021^{\dagger}$	0.035^{**}	0.154^{**}
	(0.013)	(0.011)	(0.048)
Age	0.0002	−0.0002	−0.001
	(0.001)	(0.0005)	(0.003)
Female	$-0.029^{\dagger}$	$-0.028^{\dagger}$	−0.075
	(0.017)	(0.015)	(0.078)
Some college or less	-0.037^{*}	−0.019	-0.164^{*}
	(0.018)	(0.016)	(0.083)
White	0.156^{***}	0.103^{***}	1.005^{***}
	(0.019)	(0.017)	(0.088)
Tech Literacy	-0.088^{***}	-0.050^{**}	-0.991^{***}
	(0.019)	(0.017)	(0.086)
Order: second	0.028^{*}	$0.019^{\dagger}$	0.075
	(0.013)	(0.011)	(0.048)
Order: third	0.006	0.006	0.010
	(0.013)	(0.011)	(0.048)
Constant	0.283^{***}	0.207^{***}	3.369^{***}
	(0.038)	(0.033)	(0.167)
Observations	4,746	4,746	4,746
Log Likelihood	−2,682.765	−2,087.068	−9,260.676
Akaike Inf. Crit.	5,393.531	4,202.135	18,549.350
Bayesian Inf. Crit.	5,484.042	4,292.646	18,639.860

Notes: This table presents results from mixed-effects linear regressions analyzing the effect of treatment assignments on the perceived appropriateness of ADS. All models include domain fixed effects (education or policing) and control variables for respondent age, gender, education, race, technology literacy, and question order. Standard errors are clustered at the respondent level. †p<0.1; *p<0.05; **p<0.01; ***p<0.001

Figure (A-3) Decision context and predicted views on ADS, Between Subjects Component



Notes: Each dependent variable takes the value of 1 when a respondent indicates that ADS would be appropriate/fair/accurate in a given context. Thick bars represent 90% confidence intervals; thin bars represent 95% confidence intervals. Thick bars represent 90% confidence intervals; thin bars represent 95% confidence intervals.

B.3.1 Fairness vs Accuracy considerations

Figure A-3 plots the predicted values of each of the three outcomes (appropriateness, fairness, accuracy) from the mixed-effects models that regress these binary outcomes on indicators for the decision-type treatments using random intercepts for the policy domain and the respondent. Table A-10 reports the full regression results.

Table (A-10) Full results from Figure A-3

	Dependent variable:		
	Perceived Appropriate	Perceived Fair	Perceived accurate
	(1)	(2)	(3)
T2: Aassisting individuals	-0.151^{***}	-0.140^{***}	-0.216^{***}
	(0.017)	(0.018)	(0.018)
T3: Sanctioning collectives	-0.140^{***}	-0.140^{***}	-0.053^{**}
	(0.017)	(0.018)	(0.018)
T4: Sanctioning individuals	-0.283^{***}	-0.269^{***}	-0.231^{***}
	(0.017)	(0.018)	(0.018)
Constant	0.473^{***}	0.528^{***}	0.544^{***}
	(0.016)	(0.018)	(0.033)
Observations	4,746	4,746	4,746
Log Likelihood	−2,727.375	−3,000.481	−2,979.453
Akaike Inf. Crit.	5,468.749	6,014.962	5,972.905
Bayesian Inf. Crit.	5,514.005	6,060.217	6,018.160
Note:			†p<0.1; *p<0.05; **p<0.01; ***p<0.001

B.4 Decision-Context Experiment: Within-Subjects Component

B.4.1 Summary Statistics

Table A-11 presents descriptive statistics for the dependent variables: perceived appropriateness, fairness, and accuracy of using ADS across issue areas randomized within the four types of decisions.

Table (A-11) Summary of statistics of Perceived Accuracy, Fairness, and Appropriateness

Consideration	Decision Type	Issue Area	Mean	n	SD	SE
Accuracy	Assisting Individuals	Food stamps	3.7671	790	1.8206	0.0648
		Study assistance	4.1465	792	1.7746	0.0631
	Assisting Collectives	Fire stations	4.6907	792	1.7938	0.0637
		Shelters for homeless	4.7266	790	1.7517	0.0623
	Sanctioning Individuals	Restraining order	3.1897	780	1.7978	0.0644
		Sentence	3.6135	802	1.8711	0.0661
	Sanctioning Collectives	Illegal building	4.2286	796	1.7614	0.0624
		Illegal work	4.3435	786	1.7482	0.0624
Fairness	Assisting Individuals	Food stamps	3.9038	790	1.9021	0.0677
		Study assistance	4.3056	792	1.8407	0.0654
	Assisting Collectives	Fire stations	4.6806	792	1.8558	0.0659
		Shelters for homeless	4.6772	790	1.8141	0.0645
	Sanctioning Individuals	Restraining order	3.1346	780	1.8785	0.0673
		Sentence	3.4152	802	1.8885	0.0667
	Sanctioning Collectives	Illegal building	3.8706	796	1.8591	0.0659
		Illegal work	3.9987	786	1.8224	0.0650
Appropriateness	Assisting Individuals	Food stamps	0.3354	790	0.4724	0.0168
		Study assistance	0.4104	792	0.4922	0.0175
	Assisting Collectives	Fire stations	0.5265	792	0.4996	0.0178
		Shelters for homeless	0.5557	790	0.4972	0.0177
	Sanctioning Individuals	Restraining order	0.1769	780	0.3818	0.0137
		Sentence	0.2531	802	0.4351	0.0154
	Sanctioning Collectives	Illegal building	0.3116	796	0.4634	0.0164
		Illegal work	0.3511	786	0.4776	0.0170

Table (A-12) Summary of statistics of binary outcomes

Decision Type	Decision	Consideration		
		Appropriateness	Fairness	Accuracy
Sanction Individuals	Restraining order	0.177 (0.013)	0.232 (0.015)	0.217 (0.014)
	Criminal Sentencing	0.253 (0.015)	0.281 (0.015)	0.318 (0.016)
Sanction Collectives	Construction	0.312 (0.016)	0.358 (0.017)	0.437 (0.017)
	Immigration	0.351 (0.017)	0.388 (0.017)	0.491 (0.017)
Assist Individuals	Food stamps	0.335 (0.016)	0.396 (0.017)	0.327 (0.016)
	Study assistance	0.410 (0.017)	0.468 (0.017)	0.424 (0.017)
Assist Collectives	Fire stations	0.527 (0.017)	0.571 (0.017)	0.566 (0.017)
	Homeless Shelters	0.556 (0.017)	0.557 (0.017)	0.580 (0.017)

B.5 Additional results

To ensure the findings are not sensitive to the specific items and decisions used in the between-subject component, I analyze data from the within-subject component. Table A-13 reports results of LPM regressing alternative measures of the main outcome: perceived appropriateness on indicator variables for the two theoretical dimensions—the subject and the objective of the decision—and their interaction while controlling for the issue area randomized for each decision and using fixed effects for respondent. The results are highly consistent with the main findings. Results are of a similar magnitude when using the alternative outcome measure (columns 2-3), and when using linear mixed models (columns 4-6).

Table (A-13) Decision context and appropriateness, controlling for issue area (within-subjects component)

	Dependent variable:					
	Appropriate (3 cat)	Appropriate (2 cat)	Appropriateness (7 cat)	Appropriate (3 cat)	Appropriate (2 cat)	Appropriateness (7 cat)
	OLS	*OLS*	*OLS*	*linear mixed-effects*	*linear mixed-effects*	*linear mixed-effects*
	(1)	(2)	(3)	(4)	(5)	(6)
Individuals	-0.115^{***}	-0.063^{***}	-0.683^{***}	-0.115^{***}	-0.063^{***}	-0.683^{***}
	(0.013)	(0.012)	(0.052)	(0.013)	(0.012)	(0.052)
Assisting	0.210^{***}	0.188^{***}	0.893^{***}	0.210^{***}	0.188^{***}	0.893^{***}
	(0.013)	(0.012)	(0.052)	(0.013)	(0.012)	(0.052)
Issue area: Study assistance	-0.250	0.000	0.000	0.043^{*}	0.018	$0.144^{\dagger}$
	(0.266)	(0.236)	(1.025)	(0.017)	(0.015)	(0.082)
Issue area: Shelters for homeless	-0.750^{**}	-0.500^{*}	-3.000^{**}	0.010	0.003	0.063
	(0.266)	(0.236)	(1.025)	(0.017)	(0.015)	(0.082)
Issue area: Sentence	0.750^{**}	0.750^{**}	5.250^{***}	0.033^{*}	0.018	0.196^{*}
	(0.266)	(0.236)	(1.025)	(0.017)	(0.015)	(0.082)
Issue area: Illegal work	0.000	0.250	0.250	0.011	0.005	0.076
	(0.376)	(0.333)	(1.449)	(0.017)	(0.015)	(0.082)
Order 2	-0.011	-0.002	0.020	-0.011	-0.002	0.020
	(0.013)	(0.012)	(0.052)	(0.013)	(0.012)	(0.052)
Order 3	0.001	0.001	0.033	0.001	0.001	0.033
	(0.013)	(0.012)	(0.052)	(0.013)	(0.012)	(0.052)
Order 4	-0.028^{*}	$-0.021^{\dagger}$	-0.068	-0.028^{*}	$-0.021^{\dagger}$	-0.068
	(0.013)	(0.012)	(0.052)	(0.013)	(0.012)	(0.052)
Individuals X Assisting	-0.053^{**}	-0.106^{***}	$-0.128^{\dagger}$	-0.053^{**}	-0.106^{***}	$-0.128^{\dagger}$
	(0.019)	(0.017)	(0.073)	(0.019)	(0.017)	(0.073)
Constant	0.225	-0.280	0.681	0.292^{***}	0.177^{***}	3.245^{***}
	(0.624)	(0.553)	(2.404)	(0.022)	(0.019)	(0.102)
Observations	6,328	6,328	6,328	6,328	6,328	6,328
R^2	0.543	0.529	0.655			
Log Likelihood				−3,733.977	−2,947.396	−12,628.810
Akaike Inf. Crit.				7,493.954	5,920.792	25,283.620
Bayesian Inf. Crit.				7,581.739	6,008.578	25,371.400

Notes: Standard errors are clustered at the respondent level. †p<0.1; *p<0.05; **p<0.01; ***p<0.001

B.5.1 Perceived Fairness and Accuracy

Table (A-14) Pairwise Comparisons of Fairness and Accuracy by decision type and issue area

Decision Type	Decision	Fairness	Accuracy	p-value
Assist Collectives	Both	0.564	0.573	0.370
	Fire stations	0.571	0.566	0.717
	Homeless Shelters	0.557	0.580	0.103
Assist Individuals	Both	0.432	0.375	0.000
	Food stamps	0.396	0.327	0.000
	Study assistance	0.468	0.424	0.013
Sanction Collectives	Both	0.373	0.464	0.000
	Construction	0.358	0.437	0.000
	Immigration	0.388	0.491	0.000
Sanction Individuals	Both	0.373	0.464	0.264
	Restraining order	0.232	0.217	0.293
	Criminal Sentencing	0.281	0.318	0.008

C Decision-maker Experiment

C.1 Balance Tables

This section provides the demographic balance tables for the first experiment. I run t-tests of each treatment condition and find no statistically significant differences.

C.2 Average Treatment Effects

Table A-17 assesses the possibility that using algorithmic systems to assist, rather than human decision makers, has a different effect on attitudes. The table below reports the results of a linear probability model, estimating the effects of the two conditions relative to the control condition of the human decision maker. The results show that there are no significant differences between these two conditions across all policy domains, except policing.

Table (A-15) Balance Tables

Domain	**Public Housing**				**Public Education**			
	ADS		**HDS**		**ADS**		**HDS**	
Gender	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Male	66	50.4	65	48.9	61	46.6	65	48.9
Female	65	49.6	68	51.1	70	53.4	68	51.1
Age	**X2=2.588**				**X2=5.808**			
18-24	13	9.9	17	12.8	12	9.2	18	13.5
25-34	21	16.0	24	18.0	31	23.7	19	14.3
35-44	33	25.2	29	21.8	25	19.1	30	22.6
45-54	25	19.1	25	18.8	24	18.3	20	15.0
55-64	12	9.2	17	12.8	16	12.2	16	12.0
65+	27	20.6	21	15.8	23	17.6	30	22.6
Education	**X2=0**				**X2=2.189**			
Associate's or higher	51	38.9	51	38.3	60	45.8	48	36.1
Some college or less	80	61.1	82	61.7	71	54.2	85	63.9
Race	**X2=0.96**				**X2=0.121**			
Non-White	57	43.5	49	36.8	48	36.6	45	33.8
White	74	56.5	84	63.2	83	63.4	88	66.2
Tech Literacy	**X2=0.026**				**X2=0.374**			
Low	84	64.1	83	62.4	86	65.6	93	69.9
High	47	35.9	50	37.6	45	34.4	40	30.1
Total	131		133		131		133	

Domain	**Child Welfare**				**Policing**			
	ADS		**HDS**		**ADS**		**HDS**	
Gender	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Male	62	43.7	63	45.3	68	52.7	51	36.4
Female	80	56.3	76	54.7	61	47.3	89	63.6
Age	**X2=3.784**				**X2=2.526**			
18-24	28	19.7	30	21.6	11	8.5	18	12.9
25-34	26	18.3	28	20.1	28	21.7	25	17.9
35-44	35	24.6	30	21.6	29	22.5	26	18.6
45-54	14	9.9	22	15.8	21	16.3	22	15.7
55-64	13	9.2	9	6.5	15	11.6	19	13.6
65+	26	18.3	20	14.4	25	19.4	30	21.4
Education	**X2=0.009**				**X2=0**			
Associate's or higher	58	40.8	55	39.6	55	42.6	59	42.1
Some college or less	84	59.2	84	60.4	74	57.4	81	57.9
Race	**X2=0.022**				**X2=0.019**			
Non-White	57	40.1	58	41.7	45	34.9	51	36.4
White	85	59.9	81	58.3	84	65.1	89	63.6
Tech Literacy	**X2=0.752**				**X2=0.276**			
Low	88	62.0	94	67.6	86	66.7	88	62.9
High	54	38.0	45	32.4	43	33.3	52	37.1
Total	142		139		129		140	

Table (A-16) Effects of ADS on the Support for Policy Proposals Note:

Domain	Estimate	Std.Error	Statistic	P.Value	Conf.low	Conf.high
Public Housing	0.107	0.061	1.758	0.080	-0.013	0.227
Education	0.137	0.061	2.247	0.025	0.017	0.257
Child Welfare	-0.121	0.056	-2.150	0.032	-0.232	-0.010
Policing	-0.143	0.061	-2.356	0.019	-0.262	-0.023

Table (A-17) ADS versus HDM assisted by ADS

Domain	Estimate	Std.error	Statistic	P.value	Conf.low	Conf.high
Public Housing	0.011	0.061	0.176	0.861	-0.109	0.130
Education	-0.012	0.060	-0.205	0.838	-0.131	0.107
Child Welfare	-0.031	0.056	-0.551	0.582	-0.141	0.079
Policing	-0.250	0.061	-4.080	0.000	-0.370	-0.129

C.3 Robustness Checks

C.3.1 Full Sample Analyses

To learn about the respondents' initial reactions and to address potential priming effects, the main analysis limits the sample to include responses collected by the first scenario. The results are reported in columns 1,5,9,15 of Table A-18. Columns 3,7,11,17 report estimates based on the full sample, controlling for the presenting order of the vignettes. The results remain in the same direction. Furthermore, in the main text, I presented the basic average treatment effect estimate, leveraging only the random assignment for identification. As pre-registered for the secondary analysis, Columns 2,4,6,8,10,12,14,16 report results, including the covariates. The addition of covariates makes almost no difference in the estimate of the treatment effects. The models show that the findings are robust when controlling for both fast, inattentive respondents who rush through surveys and slow, inattentive respondents who may be distracted and exhibit longer response times.

Table (A-18) Additional Results

	Dependent variable:															
	Public Housing				Child Welfare				Public Education				Police Patrolling			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
ADS	0.107†	0.116†	0.048	0.049	−0.121*	−0.126*	−0.078**	−0.073**	0.137*	0.133*	0.116***	0.115***	−0.143*	−0.134*	−0.098**	−0.098**
	(0.061)	(0.061)	(0.031)	(0.030)	(0.056)	(0.055)	(0.029)	(0.028)	(0.061)	(0.061)	(0.031)	(0.030)	(0.059)	(0.060)	(0.030)	(0.030)
HDM assisted by ADS	0.096	0.107†	0.057†	0.058†	−0.090	−0.063	−0.054†	−0.043	0.149*	0.145*	0.112***	0.111***	0.107†	0.109†	0.020	0.021
	(0.061)	(0.061)	(0.030)	(0.030)	(0.058)	(0.057)	(0.028)	(0.027)	(0.061)	(0.061)	(0.030)	(0.030)	(0.061)	(0.062)	(0.030)	(0.030)
Female		0.083		0.025		−0.022		−0.020		0.064		0.032		−0.011		−0.018
		(0.051)		(0.025)		(0.046)		(0.023)		(0.050)		(0.025)		(0.051)		(0.025)
Age		−0.002		−0.001		−0.004**		−0.001		−0.002		−0.001		0.002		0.001
		(0.002)		(0.001)		(0.001)		(0.001)		(0.002)		(0.001)		(0.002)		(0.001)
Some college or less		0.115*		0.105***		0.071		0.045†		0.022		0.029		0.029		0.014
		(0.054)		(0.027)		(0.049)		(0.024)		(0.054)		(0.027)		(0.053)		(0.026)
White		0.052		−0.019		0.045		−0.117***		0.063		0.005		−0.097		−0.006
		(0.061)		(0.030)		(0.054)		(0.027)		(0.061)		(0.030)		(0.060)		(0.030)
Independent		−0.079		−0.047		0.035		−0.013		−0.016		0.023		−0.014		−0.024
		(0.068)		(0.033)		(0.059)		(0.029)		(0.066)		(0.033)		(0.064)		(0.032)
Republican		−0.021		−0.051		0.068		0.003		−0.135*		−0.087**		−0.034		−0.044
		(0.063)		(0.031)		(0.057)		(0.028)		(0.062)		(0.031)		(0.062)		(0.031)
High Tech Literach		0.051		0.063*		0.159**		0.167***		0.003		0.064*		0.115*		0.123***
		(0.057)		(0.029)		(0.053)		(0.026)		(0.059)		(0.029)		(0.057)		(0.029)
Inattentive		−0.057		−0.052		0.080		0.119***		0.013		−0.026		−0.008		0.063†
		(0.073)		(0.036)		(0.065)		(0.033)		(0.077)		(0.037)		(0.071)		(0.036)
Order			−0.008	−0.008			−0.010	−0.010			0.006	0.004			−0.022*	−0.021*
			(0.010)	(0.010)			(0.010)	(0.009)			(0.010)	(0.010)			(0.010)	(0.010)
Constant	0.511***	0.457***	0.543***	0.544***	0.403***	0.410***	0.393***	0.401***	0.474***	0.548***	0.445***	0.465***	0.600***	0.538***	0.653***	0.568***
	(0.043)	(0.119)	(0.034)	(0.063)	(0.040)	(0.097)	(0.032)	(0.057)	(0.043)	(0.112)	(0.035)	(0.063)	(0.041)	(0.121)	(0.034)	(0.062)
Observations	394	394	1,582	1,582	409	409	1,582	1,582	394	394	1,582	1,582	385	385	1,582	1,582
R^2	0.010	0.039	0.003	0.024	0.012	0.083	0.006	0.100	0.019	0.046	0.012	0.028	0.041	0.068	0.014	0.034

Notes: This table reports estimates from LPM. The base category of the independent variable is the human decision-maker (HDM).†p<0.1; *p<0.05; **p<0.01; ***p<0.001

C.3.2 Alternative Measures of Outcomes

To easily interpret the ATE as the percentage change in public support for a policy proposal as a result of the usage of ADS, I coded the outcome as a binary variable that takes the value "1" if the respondent "strongly" or "somewhat supports" the policy and 0 otherwise. As preregistered for the secondary analysis, I replicate the results, using an alternative outcome that measures support as a five-point scale (Columns 1-4) and as binary variable for strongly support' (Columns 5-8. The table below shows that all conclusions remained the same when I measured support for policy proposal as a scale with five values.

Table (A-19) Alternative Measures of Outcomes

	Dependent variable:							
	Strongly Support				Support (5-point)			
	(Housing)	(Child)	(Education)	(Policing)	(Housing)	(Child)	(Education)	(Policing)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ADS	$0.258^{\dagger}$	-0.350^{*}	0.309^{*}	-0.368^{**}	0.102^{*}	-0.053	0.011	-0.131^{*}
	(0.144)	(0.166)	(0.153)	(0.138)	(0.051)	(0.040)	(0.052)	(0.052)
HDM assisted by ADS	$0.247^{\dagger}$	$-0.296^{\dagger}$	0.352^{*}	0.142	0.073	-0.041	0.020	0.016
	(0.144)	(0.171)	(0.154)	(0.142)	(0.051)	(0.041)	(0.052)	(0.054)
Constant	3.376^{***}	2.914^{***}	3.256^{***}	3.686^{***}	0.165^{***}	0.158^{***}	0.218^{***}	0.286^{***}
	(0.101)	(0.118)	(0.108)	(0.096)	(0.036)	(0.028)	(0.037)	(0.036)
Observations	394	409	394	385	394	409	394	385
R^2	0.010	0.012	0.016	0.034	0.011	0.005	0.0004	0.023

Note: †p<0.1; *p<0.05; **p<0.01; ***p<0.001

C.3.3 Interaction Between Decision-maker and Context

To assess whether the decision context moderates the effect of ADS on the evaluation of policy proposals, I examine the interaction effect of the decision-maker and the decision context treatments on support. The table below reports the results of logistic regression models in which the probability of supporting the policy proposal is regressed on the decision context (4-category variable capturing the policy proposal presented first), the decision-maker (3-category variable capturing HDM, ADS, and HDM assisted by ADS), and their interaction. The base categories of the key variables are the policy of child-abuse allegations and HDM. Thus, the last row reports the baseline probabilities of support for the proposal to prioritize child abuse investigations by human decision-makers (child welfare workers). The analysis is based on data collected from all scenarios presented first for the respondents.

Table (A-20) Interaction - Policy context and Decision maker

	Dependent variable:			
	Policy Support			
	(1)	(2)	(3)	(4)
ADS X Education	0.258**	0.257**	0.660**	0.651**
	(0.083)	(0.083)	(0.213)	(0.212)
ADS X Policing	−0.021	−0.011	−0.018	0.012
	(0.083)	(0.083)	(0.212)	(0.211)
ADS X Public Housing	0.228**	0.235**	0.608**	0.616**
	(0.083)	(0.083)	(0.213)	(0.212)
ADS	−0.121*	−0.122*	−0.350*	−0.351*
	(0.058)	(0.058)	(0.148)	(0.147)
Proposal: Education	0.071	0.082	0.342*	0.378*
	(0.059)	(0.059)	(0.151)	(0.150)
Proposal: Policing	0.197***	0.200***	0.772***	0.781***
	(0.058)	(0.058)	(0.149)	(0.148)
Proposal: Public Housing	0.108†	0.111†	0.462**	0.476**
	(0.059)	(0.059)	(0.151)	(0.150)
Age		−0.017*		−0.041†
		(0.008)		(0.021)
Female		0.030		0.101
		(0.025)		(0.063)
Independent		−0.021		−0.035
		(0.032)		(0.082)
Republican		−0.032		−0.073
		(0.030)		(0.078)
Some college or less		0.053*		0.104
		(0.026)		(0.067)
White		0.018		−0.073
		(0.029)		(0.073)
Tech Literacy		0.084**		0.178*
		(0.027)		(0.070)
Constant	0.403***	0.388***	2.914***	2.943***
	(0.041)	(0.060)	(0.105)	(0.153)
Observations	1,582	1,582	1,582	1,582
R^2	0.066	0.080	0.097	0.113

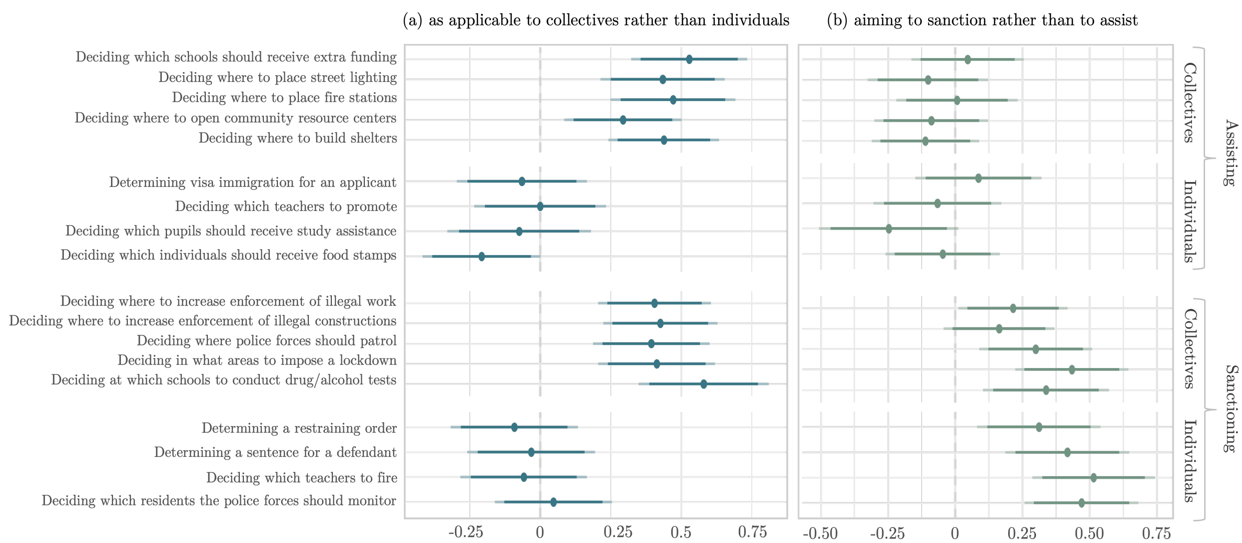
Note:
†p<0.1; *p<0.05; **p<0.01; ***p<0.001

D Validating the Theory Using MTurk Data

I tested the validity of the theoretical classification using survey data from MTurk (N=150).

Respondents were presented with a matrix of randomly selected decisions (six out of 19 decisions) and asked to classify each of the decisions into one of four decision types derived from the theory: assisting individuals, assisting collectives, sanctioning individuals, or sanctioning collectives. Notable, I provided no information about the identity of the decision-maker—whether decisions were made by humans or algorithms, as the main goal was to confirm that people do indeed agree with the theoretical classification of policy decisions into two be two dimensions.[25]

Figure (A-4) Empirical Validation of the 2x2 Classification



Figures A-4 show estimates of linear probability models that include fixed effects for each respondent. Figure A-4 (a) shows the probability of classifying each decision as aim to sanction rather than assist while Figure A-4 (b) displays the probability of classifying each decision as target collectives rather than individuals. Thick bars represent 90% CI; thin bars represent 95% CI. The original classification of the questions, as defined by the theory, is indicated on the right side of the figure. Since each respondent evaluated a random subset of six decisions from the pool of decisions, the baseline in the model is not a specific decision category but rather each respondent's own average response across the six decisions they evaluated. Therefore, each decision's coefficient compares its likelihood of being classified as targeting collectives rather than individuals in Model (a) or as sanctioning rather than

[25] The exact wording of the question was as follows: "Next, we will present you with several decisions. These decisions differ by: (1) What they aim to do: decisions that assist by providing social services or goods; and decisions that sanction by limiting lives or opportunities. (2) To whom they apply: decisions that apply to individuals and decisions imposed on collectives such as communities or areas. Please indicate which category best describes each of the following decisions."

assisting in Model (b) relative to the average classification across the decisions evaluated by each respondent.

The results are consistent with the theoretical classification. The figure shows that all (and only) decisions originally classified by the theory as targeting collectives have positive, statistically significant, and substantively large estimated coefficients. On average, these coefficients are statistically significantly different from the coefficients of decisions originally classified as targeting individuals, suggesting that respondents do distinguish between these decisions. Similar findings are observed in the model regressing the probability of defining decisions as sanctioning rather than assisting.

Furthermore, examining the proportion of respondents who classified each decision according to the theoretical classification reveals that respondents generally align with the two theoretical dimensions when categorizing policy decisions, with agreement levels ranging from 53% to 93%. While some variation exists in the level of agreement with the theoretical classification, this variation is not heavily biased towards any particular type of decision. Therefore, the 2x2 framework offers a useful initial structure for understanding contextual variation in preferences. These distinctions could be further studied as a continuum.

E Research Ethics

The study is based on a survey administered by the survey company Dynata (previously known as SSI). The survey was reviewed and approved by IRB before the study was initiated (protocol numbers: 0004542-1). It was complied using the current standards for research transparency and ethics, including the American Political Science Association's "Principles and Guidance for Human Subjects Research" as approved by the APSA Council in April, 2020. Informed consent was obtained from each participant at the beginning of the survey. Specifically, respondents were informed that (1) the survey was voluntary, (2) they could exit it at any time without penalty, and (3) they were free to decline to answer any particular question. Respondents were reimbursed by the survey firm with standard compensation. Moreover, the survey companies did not provide any identifying data, such as names or email addresses, so the data used in the analysis and provided for the replication would be anonymous. Finally, the policy proposals that respondents were asked to evaluate were based on real initiatives to incorporate AI technologies. This means that the experiment did not include false information.

F Pre-registration

This study was pre-registered on OSF (EGAP Registration ID: 20220323AA) on March 23, 2022 in a non-anonymous version. This section includes a blind version of the pre-registration. Note that the pre-registration report includes 3 experiments, but only 2 of them are relevant for this paper.

Is this Registration Prospective or Retrospective? Registration prior to any research activities.

Is this an experimental study? Yes.

Date of start of study 23/04/2002.

Was this design presented at an EGAP meeting? No.

Background and explanation of rationale. In this project, I seek to understand how people respond to the growing use of algorithmic decision systems (ADSs) in public policy and to explain variations in preferences across policy domains and decision contexts. I will do so by (1) developing a theoretical framework to account for variations in views on the fairness, accuracy, and legitimacy of ADSs across decision contexts and (2) subjecting the theory and its implications to empirical tests using novel data from three experiments embedded in a national representative survey of the US population.

What are the hypotheses to be tested/quantities of interest to be estimated? If people are, as previous studies have suggested, algorithm averse, we would expect that, when asked directly, citizens will prefer that a human being rather than an algorithm make high-stake decisions in the public sector. However, people's preferences over human decision-makers (HDM) would not translate uniformly into less support for policy decisions that rely on algorithmic assessment. The sensitivities people have toward the use of ADS differ depending on the decision context in which it is deployed. I propose classifying decisions in the public sector along two dimensions that I consider relevant to the way we define a correct or incorrect decision and the consequences of that decision. The first dimension relates to the population directly affected by the decision (individuals vs collectives). The second dimension relates to the broader objective of the decision (assisting vs sanctioning). In decisions about collectives, the reliance on big data to draw predictions about aggregate cases will be perceived as highly accurate when compared with predictions on individuals, wherein the dependence upon bigdata may be perceived as less accurate because it is more vulnerable to errors in the context of certain cases (more specifically, borderline or marginal cases). Due to the perceived lack of subjectivity surrounding algorithms, they may be perceived to be fairer in assisting decisions that provide public goods and services. However, this exact lack of subjectivity makes ADSs be perceived less appropriate when they are used in the context of sanctioning decisions, which may have irreversible consequences that limit individual or collective lives.

Based on this theory, I put forth the following expectations about variations in people's views across these dimensions: (1) People will be less sensitive to the use of ADS in decisions on collective cases versus individuals. (2) People will be more supportive of the use of ADS in decisions that assist collectives versus punish collectives. (3) When there is a tradeoff between fairness and accuracy, people are likely to be less tolerant to ADS in decisions that entail irreversible consequences.

How will these hypotheses be tested? To test the hypotheses outlined above, I designed a survey that compromises three experiments (A flow diagram of the survey experiment is provided below).

To assess whether and how algorithmic decision-making, compared to human decision-making, affects the evaluation of policy decision-making, the first experiment manipulates the decision-maker (ADS), a human decision-maker (HDM), and an HDM assisted by an ADS).

The third experiment is designed to assess the theory I developed to explain variation in attitudes across the two dimensions: (1) the broader objective of the decision (assisting versus sanctioning decisions) and (2) the subject who will be affected by the decision (individuals versus collectives). Respondents will be asked to express their opinion on the use of ADS in 7 randomly selected decision contexts presented in a matrix. The section is composed of two designs: between and within-subjects.

A between-subjects design: To assess variation in people's sensitivities to ADS within policy domains, the first three decisions in the matrix ask on three fixed policy domains: policing, child welfare, and education, where the type of decision in each policy domain is randomly assigned into 1 our of 4 types of decisions derived from the theory I developed: assisting individuals, assisting collectives, sanctioning individuals or sanctioning collectives.

A within-subjects design: The matrix includes 4 additional decisions (1 out of 2) in each of the four decision types: assisting individuals, assisting collectives, sanctioning individuals or sanctioning collectives, where the policy domain is randomly assigned.

Before the treatments are allocated, I will collect demographic information (birth year, race/ethnicity, and education). I will also ask two attention check questions (one before the first experiment and one before the second experiment). If respondents fail those attention checks, they are removed from the survey. All randomizations of the survey elements (listed below) will take place at the level of the individual respondent. Conditions will be randomly assigned with equal probability using random number generation within Qualtrics survey software.

<u>How will these hypotheses be tested?</u> The target sample size is 1,500 in the United States. A sample size of 1,500 respondents will allow us to detect an effect size of approximately 0.2 standard deviations at the standard 0.05 alpha error probability in the experiment with the largest number of conditions (using the conventional 80% power level).

<u>Country</u> United States.

<u>Sample Size (of Units)</u> The target sample size is 1,500 in the United States. A sample size of 1,500 respondents will allow us to detect an effect size of approximately 0.2 standard deviations at the standard 0.05 alpha error probability in the experiment with the largest number of conditions (using the conventional 80% power level).

<u>Was a power analysis conducted prior to data collection?</u> Yes.

<u>Has this research received Institutional Review Board?</u> Yes.

<u>IRB Number</u> 0004542-1.

<u>Date of IRB Approval</u> 13-02-2022.

G Pre Analysis Plan

<u>**1. Decision-Maker Experiment**</u> Assessing the hypothesis that people's preferences over human decision-makers would not translate uniformly into lower support for policy decisions that rely on algorithmic assessment.

1.1. Primary analyses: (a) For each of the four policy proposals, I will compute the average support (the main outcome) and standard deviations across the two key experimental groups (HDM and ADS). (b) For each of the four policy proposals, I will compute the average treatment effect (ATE) of ADS versus HDM and its standard error.

1.2. Secondary analyses: (a) I will compute the ATEs of ADS vs. HDM and vs. HDM assisted by ADS (HDM+ADS). (b) I will report estimates from OLS regression models adjusting for respondents' socio-demographic characteristics: gender, age, race, and education, in order to improve the precision of estimates. I do not expect the inclusion of these covariates to meaningfully change the size of estimated effects – just the size of the standard errors. (c) I will report estimates from OLS regression models using alternative measures of the outcome variable (see secondary outcomes).

1.3 Explanatory analyses: I will report the conditional marginal effects of decision-maker and policy context, using OLS models regressing the outcome on dummies for each treatment -decision-maker and policy context – and their interaction. The analysis will be based on the data collected from the first policy proposal randomly presented to the respondent. Note that the key aim here is only to provide suggestive evidence for the variation in people's sensitivities to the use of AI across contexts. A more nuanced examination of this variation is provided in the third part of the survey, which is designed to assess the theory I developed.

2. Decision-Context Experiment Analyzing in more depth the theory of variation across decision contexts.

2.1. Between-subjects analysis

2.1.1. Primary analyses (a) For each of the three policy domains, I will compute the average of the three primary outcomes and standard deviations across the four types of decision: assisting individuals, assisting collectives, sanctioning individuals, sanctioning collectives. (b) For each of the three policy domains, I will calculate the ATEs of the four types of decisions. Specifically, I will report estimates from OLS regression models studying the effect of the type of the decision on the probability to view the use of ADS as (1) appropriate, (2) fair, (3) accurate, leaving sanctioning collective as the reference category.

2.1.2 Secondary analyses (a) I will report estimates from OLS regression models adjusting for respondents' socio-demographic characteristics: gender, age, race, education, party affiliation, technological orientation. I do not expect the inclusion of these covariates to meaningfully change the size of estimated effects – just the size of the standard errors. (b) I will report estimates from OLS regression models using alternative measures of the outcome variables (see secondary outcomes). (c) For each of the three policy domains, I will report estimates from OLS regression models studying the interaction between the two theoretical dimensions. I will report estimates from OLS models regressing the three outcomes (appropriate, fair, accurate) on dummies for each of the two theoretical dimensions – the objective of the decision (assisting or sanctioning) and the subjects of the decision (individuals or collectives) – and the interaction between them. The analysis will be based on the data collected from the first policy proposal randomly presented to the respondent.

2.3. Within-subjects analysis

2.3.1. Primary analyses (a) I will compute the average of the three primary outcomes and standard deviations across the four types of decisions. (b) I will report estimates from OLS regression models studying the independent predictive role of each theoretical dimension – the objective of the decision (assisting or sanctioning) and the subjects of the decision (individuals or collectives) on the three outcomes of interest. The model will include fixed effects for respondents.

2.3.2. Secondary analysis: I will report estimates from OLS regression models adjusting for respondents' socio-demographic characteristics: gender, age, race, education, party affiliation, technological orientation, and using alternative measures of the outcome variables (see secondary outcomes). I will also check for heterogeneity in people's attitudes based on demographics and general trust in traditional DM and AI technology.