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

Shir Raviv[1]

Abstract

Government agencies increasingly use algorithmic decision systems (ADS) to assist or replace human judgment across various policy areas such as criminal justice, welfare, and education. 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 data from original experiments embedded in a national U.S. survey. I show that across a wide range of policy domains, citizens strongly oppose using 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 using ADS in these contexts can significantly undermine the legitimacy of the policy interventions they inform. The study offers a framework to identify where AI-based tools will be deemed appropriate and where they might trigger a backlash, underscoring the importance of accounting for citizens' values and concerns in governing AI.

[1] Data Science Institute, Columbia University; New York, NY 10027, US. sr4093@columbia.edu

[2] Supplementary material for this article is available in the Appendix in the online edition. Replication files are available in the JOP Dataverse (https://dataverse.harvard.edu/dataverse/jop). The empirical analysis has been successfully replicated by the JOP replication analyst. The study was conducted in compliance with relevant laws and was approved by the Institutional Review Board at Tel Aviv University (approval number: 0004542-1).

Introduction

In November 2020, Californians voted on a referendum to replace cash bail for pretrial release with algorithmic risk assessment. Under the proposed law, local courts would use algorithms to decide whether defendants should remain in custody or be released before trial, based on their likelihood to appear in court, the seriousness of their offense, and their probability of reoffending (Pislar and Puleo, 2020). Despite evidence that such systems could lower crime rates among released defendants without increasing incarceration (Kleinberg, Mullainathan, and Raghavan, 2016), voters rejected the proposition by a wide margin (56% to 44%). What explains this rejection? Does this opposition reflect general skepticism toward the use of algorithms in governance, or does it stem from specific concerns about applying them in criminal justice?

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, government agencies and public authorities are increasingly relying on AI-based algorithms–software that autonomously makes assessments and predictions based on inferences from big data without explicit human instructions–to make determinations on 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).

This paper introduces a theory that explains when citizens accept algorithmic governance by focusing on considerations of fairness and accuracy. I argue that these views are context-dependent and vary as a function of (1) the objective of the decision at stake, specifically whether it is seen as assisting or sanctioning, and (2) the population directly affected by the decision: individuals versus collectives. I test this theory and its observable implications using novel data from two original, pre-registered experiments embedded in a nationally representative survey of the U.S. population.

The first experiment systematically examines individuals' perceptions of the appropriateness, fairness, and accuracy of ADS in governance by randomizing both the decision type and the policy domain in which the algorithm is employed. The results provide strong support for the theory: people exhibit aversion to ADS, particularly in decisions perceived as designed to sanction rather than assist, as well as when they require inferences about individuals rather than collectives. These findings are generalizable across a wide range of decisions and policy domains, including public education, immigration, social welfare, and criminal justice. The analysis also highlights the trade-off 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 follows the pattern predicted by the theory: respondents were less tolerant of ADS in sanctioning decisions, even when such systems were perceived to improve accuracy in decision-making.

While public opinion does not always shape policy, public approval of ADS in governance is crucial for maintaining legitimacy and democratic accountability. This is evident in recent high-profile cases where governments and municipalities have reversed or abandoned policies implemented by ADS due to public backlash. For example, both New Orleans and Los Angeles terminated their predictive policing programs following public outcry over racial bias and lack of transparency (Winston, 2018; Sainato and Chiu, 2021). Similarly, the UK's Department for Education withdrew its grade prediction algorithm amid protests over unfair treatment of disadvantaged students (Walsh, 2020), while in the Netherlands, public backlash against an algorithmic system for welfare fraud detection led to the government's resignation (International, 2021).

To empirically assess the political implications of public attitudes toward ADS in governance, I present results from a second experiment that examines how algorithmic implementation affects overall policy support. By asking respondents to evaluate identical policy proposals while randomizing the decision-maker implementing the policy, I explore whether citizens actually care about the use of ADS and whether they consider it when evaluating policy issues. The results suggest that using ADS in contexts where citizens view them as inappropriate can undermine the legitimacy of the policy decisions and interventions they inform. Policy proposals involving sanctioning decisions, such as prioritizing child abuse investigations, received significantly less support when implemented algorithmically rather than by human officers. In contrast, policies that involve decisions assisting collectives–such as allocating additional school funding–gained more support when implemented algorithmically. The findings also indicate that when there is a tradeoff between fairness and accuracy, using algorithms as a supportive tool while keeping “humans in the loop” is an attractive option. Overall, the study provides a useful framework to assess where AI-based tools will be deemed appropriate, might trigger backlash, and where combining algorithmic assessment with human judgment is most appealing.

Beyond their practical implications, these findings contribute to the growing literature on the determinants of public opinion toward AI and data-driven decision-making. Most experimental research on this topic focuses on the views and reactions of AI users or operators who interact directly with AI algorithms and can choose whether and how to use their output (Lee, 2018; Waggoner and Kennedy, 2022). More recently, studies have shifted their focus to the general public, who are subjected to algorithmic decisions without the ability to opt out (Zhang and Dafoe, 2019; O’Shaughnessy et al., 2023). The findings presented in this paper add to the limited but rapidly growing research highlighting 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 perceptions of fairness and accuracy regarding the same algorithmic system can vary depending on the type of decision it informs, this study provides more nuanced and systematic insights that apply across policy areas.

More broadly, the study contributes to the growing body of research on the political ramifications of recent advancements in AI and digitization, which has primarily focused on labor market disruptions (e.g., Gallego and Kurer, 2022). It provides insights into an important yet underexplored domain where AI-based technology increasingly shapes citizens' lives and their interactions with government agencies, with significant implications for democratic governance. It therefore underscores the need for a broader research agenda in political science that examines citizens' values, expectations, and concerns regarding the use of AI in governance and explores ways to incorporate these perspectives into AI governance frameworks.

Contextual Attitudes Toward Using AI Algorithms in Governance

The integration of ADS in high-stakes policy domains has sparked a debate about the potential benefits and risks (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, lack of transparency and accountability, and privacy violations (e.g., Barocas, Hardt, and Narayanan, 2017).

Much of this debate centers on whether ADS can enhance or diminish accuracy and fairness in decision-making. Accuracy, in this context, refers to the extent to which the algorithm achieves its intended outcomes, such as correctly identifying individuals likely to recidivate 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 overlap with accuracy when reducing bias leads to decisions that are both fairer and more accurate. However, it also involves more substantive aspects that go beyond accuracy, such as ensuring equal opportunities and accountability (Reich, Sahami, and Weinstein, 2020). The latter concerns the consequences of such decisions, specifically the extent to which they affect or constrain citizens' lives.

How do citizens evaluate the fairness and accuracy of ADS? Most empirical work assumes that people's views of algorithms are relatively fixed, determined either by their predispositions toward the technology (Dietvorst, Simmons, and Massey, 2018; Zhang and Dafoe, 2019) or by their prior knowledge about AI (Horowitz and Kahn, 2024). Other research emphasizes the design features of the technology, such as the quality and quantity 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 people's evaluations of ADS vary depending on the context in which 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 regarding the accuracy and fairness of using ADS in governance depend significantly 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 that target *collectives* (i.e., groups or areas), such as which neighborhoods to patrol or which schools should receive additional funding.

The second dimension relates to the decision's objective, specifically whether it seems designed to sanction or assist. *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 undocumented immigration or removing a child from their parent's care.

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, in some cases, be viewed as sanctioning rather than assisting—for example, when individuals are deemed *ineligible* or do not *qualify* for support. However, the theory assumes a fundamental difference between decisions that "do not give" (assisting) and those that "take away" (sanctioning). This difference stems from the potential change to the status quo, which has implications for the decision's consequences, particularly the extent to which the decision outcome is reversible.[2] Drawing on Berlin (1969)'s distinction between negative and positive liberty, sanctioning decisions directly constrain an individual's choices and behaviors, thus impacting their negative liberty—their freedom from external constraints. In contrast, assisting decisions shape the conditions and resources that enable individuals or groups to pursue their goals, relating to their positive liberty—that is, their capacity to act. This distinction has important implications for the reversibility and lasting consequences of such decisions.

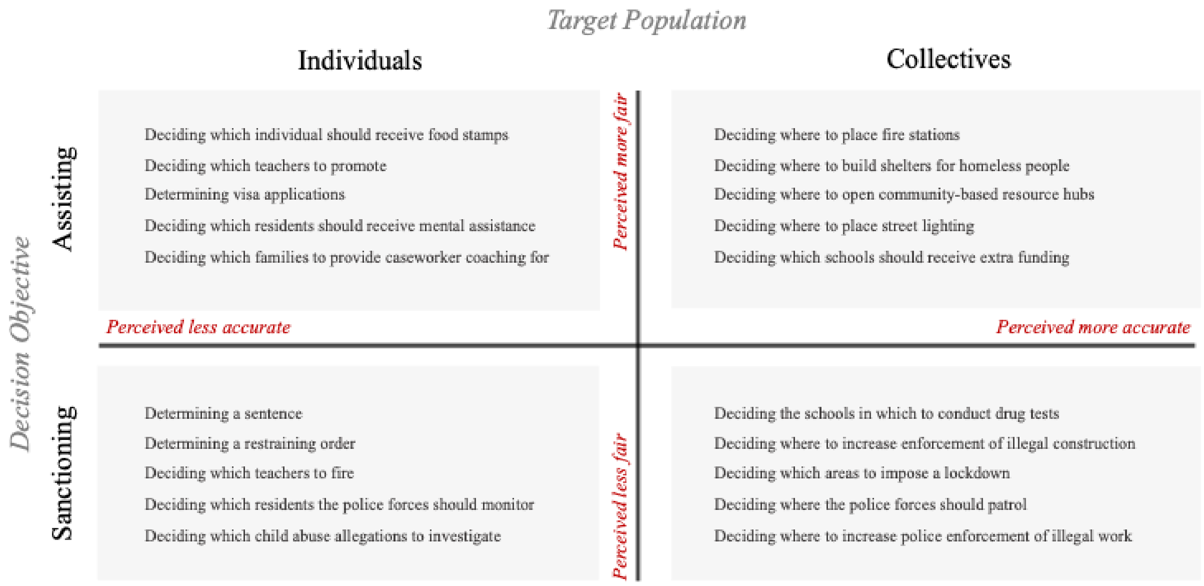
To validate this theoretical framework, I conducted a survey on MTurk, asking 150 respondents to categorize six randomly selected decisions into the four types derived from the framework, without providing any information about the identity of the decision maker. The results, reported in Figure A-4, show that respondents' answers are highly consistent with this two-by-two classification.

While the two dimensions are not exhaustive, they provide a useful starting point for understanding contextual variation in preferences. As Figure 1 illustrates, many real-world decisions in the public sector can be categorized within this two-by-two framework.

I contend that ADS are more likely to be seen as enhancing accuracy when applied to *collectives* rather than *individuals*, as they excel at processing large volumes of data but may overlook individual nuances and exceptional circumstances that human judgment and discretion are better suited to address. In terms of fairness, the impersonal nature of ADS may be viewed as an asset in *assisting* decisions that allocate benefits, as it reduces the risk of favoritism and corruption often associated with human decision-makers. However, this same impersonality can make algorithms appear less fair when used in *sanctioning* decisions, which tend to have less reversible consequences. In such contexts, human accountability plays a key role in ensuring fairness. In what follows, I characterize each of the four decision types in terms of their implications for perceived accuracy and fairness and derive observable implications for the perceived legitimacy of using ADS in each case.

[2] Building on this framework, future studies should examine these distinctions as a spectrum, as some decisions may be perceived as more assisting than sanctioning. Another useful direction is to explore heterogeneity across individuals in how they classify policy decisions, as this could influence their views on using ADS in these contexts.

Figure (1) Four Types of Decisions in Public Policy



	Individuals	Collectives
Assisting	Deciding which individual should receive food stamps Deciding which teachers to promote Determining visa applications Deciding which residents should receive mental assistance Deciding which families to provide caseworker coaching for	Deciding where to place fire stations Deciding where to build shelters for homeless people Deciding where to open community-based resource hubs Deciding where to place street lighting Deciding which schools should receive extra funding
Sanctioning	Determining a sentence Determining a restraining order Deciding which teachers to fire Deciding which residents the police forces should monitor Deciding which child abuse allegations to investigate	Deciding the schools in which to conduct drug tests Deciding where to increase enforcement of illegal construction Deciding which areas to impose a lockdown Deciding where the police forces should patrol Deciding where to increase police enforcement of illegal work

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

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 when compared to the limited ability of humans to capture, aggregate, and process such vast amounts of information (Green and Chen, 2019). Research suggests that people view algorithms that use big data as inherently trustworthy, treating "big data" as a heuristic to judge the algorithm's quality (Waggoner et al., 2019).

Since algorithms make decisions based on rules applied consistently across time, parties, and situations, several studies indicate that such technology can improve not only accuracy but also fairness in decision-making (e.g., Sunstein, 2019; Helberger, Araujo, and Vreese, 2020). This consistency is particularly valuable in distributive decisions, which involve allocating benefits to specific groups or communities. According to Lowi (1964), distributive policies typically operate on a case-by-case basis, allocating benefits to specific constituencies without clear overarching principles. This fragmented decision-making process creates opportunities for favoritism as decision-makers can make ad hoc allocations that benefit particular groups while spreading costs diffusely across society. The impartial, rule-based nature of algorithmic systems may therefore be seen as particularly desirable in such contexts, as it potentially limits discretionary allocation based on political or personal considerations.

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

2. Sanctioning collectives

For the same reasons discussed in the context of assisting decisions, data-driven algorithms appear highly accurate when identifying areas or communities likely to face significant challenges or risks. However, using these assessments and predictions—however accurate they may be—to sanction and punish targeted communities rather than provide needed resources can be perceived as unfair in substantive ways.

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 could reflect and therefore perpetuate 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).[3]

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 place more streetlights or where to open a community-based resource center) but significantly unfair in decisions that sanction collectives (e.g., deciding which schools should be subject to drug and alcohol testing).

The observable implication is that using ADS for decisions that sanction collectives 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 government, including judges, teachers, social workers, and police officers, that directly interacts with citizens and makes day-to-day decisions (Lipsky, 1980). These decisions often involve nuances or extenuating circumstances, making it impossible to prescribe (and thus code) a correct response in advance for all cases.

Human bureaucrats can flexibly adjust their decision boundaries when confronted with novel or marginal cases. Algorithms, by contrast, rely on patterns in existing data and can only refine their judgments after receiving feedback or additional training data—typically after an error has occurred (Binns, 2019). By design, data-driven systems simplify complexity: they cannot account for all relevant contextual details and tend to treat people as members of categories rather than as individuals (Brauneis and Goodman, 2018). As a result, algorithms may struggle to identify borderline or exceptional cases and may be perceived as less

[3] 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).

accurate than humans in individual-level decisions.[4]

At the same time, the discretion that allows human decision-makers to tailor responses can also lead to misuse—whether intentional or not—due to personal biases, favoritism, or reliance on irrelevant factors (Danziger, Levav, and Avnaim-Pesso, 2011; Alkhatib and Bernstein, 2019). In contrast, ADS apply consistent, rule-based criteria, which may lead people to perceive them as fairer from a procedural standpoint.

Taken together, people are expected to weigh a trade-off between accuracy and fairness when evaluating the use of ADS in decisions assist individuals. 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 assisting decisions, algorithms' limited ability to account for novel or borderline circumstances may lead people to see 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 make it difficult not only for programmers to explain their outputs but also for ordinary citizens to understand or challenge the decisions these systems produce (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 an innocent person) and for society's safety (e.g., a false negative that exonerates a guilty individual).

[4] This concern reflects public intuitions rather than statistical or expert perspectives. People tend to be more attuned to individual-level variance when decisions are granular, whereas they assume such variance is averaged out at the collective level.

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

		Target Population	
		Individuals	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	

Returning to the example that opened this paper–the proposal to replace California's cash bail system with an algorithmic risk assessment tool. 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.

In summary, Table 1 outlines the characteristics of each decision type, focusing on accuracy, fairness, and the potential trade-offs between them. Citizens generally view ADS as both fair and accurate when used to assist collectives, but see them as less fair and less accurate when used to sanction individuals. In the remaining two types—assisting individuals and sanctioning collectives—they may face a trade-off between fairness and accuracy, preferring human decision-makers, especially when the decisions carry lasting, less-reversible consequences and demand clear accountability. The following sections test these theoretical expectations empirically.

Research Design

To test the theoretical predictions, I fielded two original experiments embedded within a nationally representative survey of U.S. adults. The sample included 1,590 respondents recruited by Dynata (formerly Survey Sampling International) in March–April 2022. Dynata is a widely used provider in social science research and employed quota sampling to approximate the U.S. adult population in terms of gender, age, education, race and ethnicity (Malhotra, Monin, and Tomz, 2019; Read, Wolters, and Berinsky, 2021). Table A-1 in the Appendix compares the sample's characteristics with those of the general U.S. population and shows that the sample is broadly representative along the quota dimensions. Additional details about the sample are provided in Appendix A.

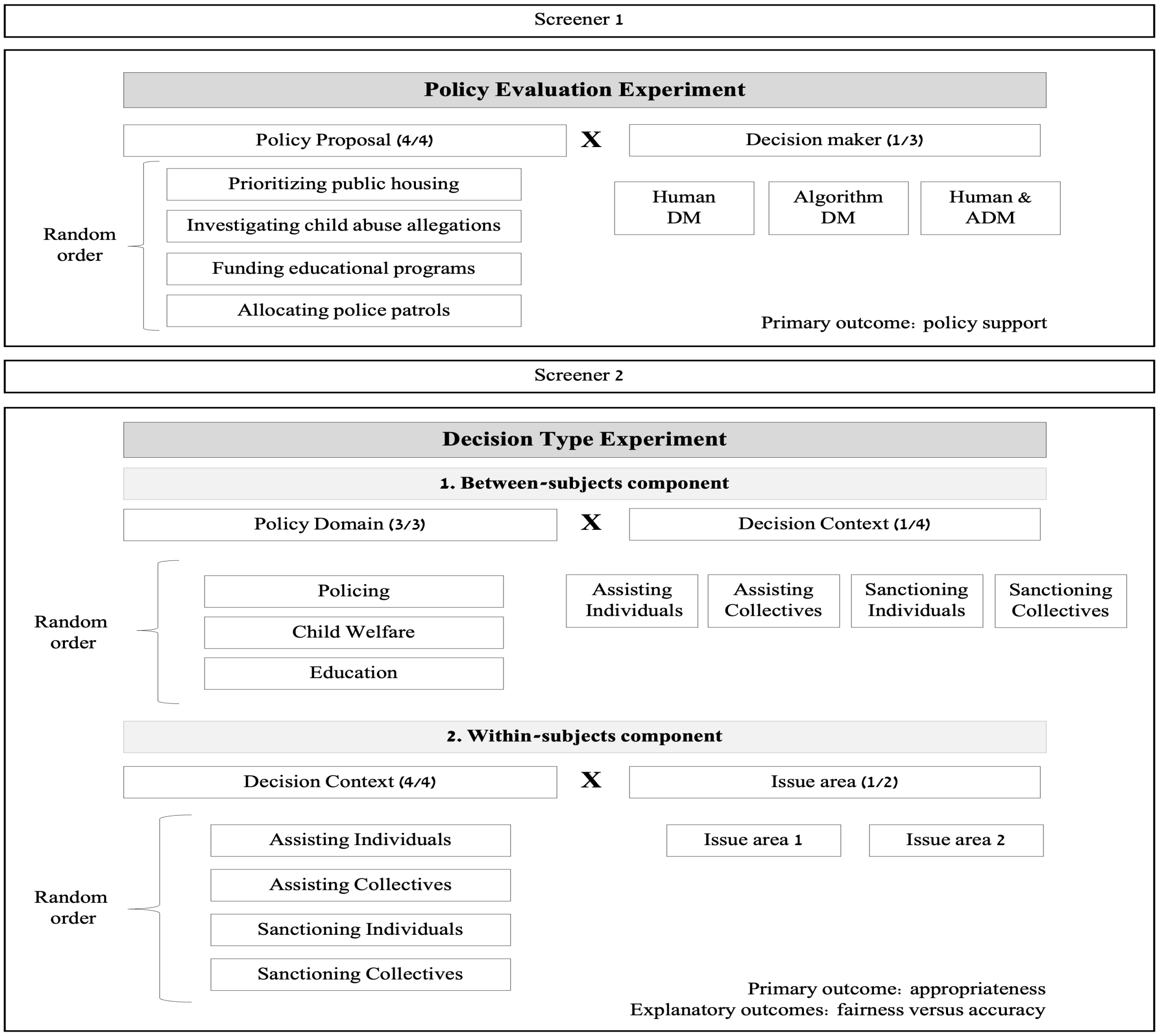
The survey includes two experiments. The *Decision Type Experiment* directly tests the theory by asking respondents to evaluate the appropriateness of ADS across different policy contexts, randomizing the type of decision along two theoretical dimensions. The *Policy Evaluation Experiment* examines the political implications of these views by asking respondents to evaluate identical policy proposals, randomizing the decision-maker who implements the policy. Figure 2 illustrates the structure of the survey design. Respondents participated in both experiments, but treatment assignment in each experiment was independent.[5]

To ensure that respondents shared a common understanding of what a predictive algorithm is, the survey began with the following meaning: "A predictive algorithm is computer software that makes decisions without human instruction, relying on large amounts of data." To minimize experimenter demand effects, I presented this definition indirectly by embedding it among two other definitions related to the survey topic.[6]

[5] Presenting the Policy Evaluation Experiment first ensures respondents evaluate policies on their merits without priming them to consider who implements the policy.

[6] This item also served as a screener. On the next page, respondents saw four definitions and were asked to identify the one that had not appeared earlier. Those who failed to answer correctly were removed from the study before the first randomization. I included a second screener before the second experiment. Only respondents who passed both pre-treatment screeners completed the full survey and were included in the analysis.

Figure (2) Survey Design



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

Decision Type Experiment

The *Decision Type Experiment* directly tests the theory by examining how people's views on the use of ADS in public policy vary across policy domains and decision types. 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 components.

Between-Subject Component. Respondents evaluated decisions from three high-stakes policy domains: policing, education, and child welfare, presented in a random order on the same matrix. I independently randomized each policy domain along the two theoretical dimensions: (1) whether the decision assists or sanctions and (2) whether the decision targets individuals or collectives.[7] Table 2 provides the wording of the decisions by policy domains.

***Within-subject Component*.** Respondents were presented with four additional items on the same matrix, each corresponding to one of four decision types: assisting individuals, assisting collectives, sanctioning individuals, and sanctioning collectives. For each decision type, the specific issue area was randomly assigned to one of two options. For example, all respondents evaluated the use of ADS in one of two sanctioning decisions targeting individuals: either determining a sentence based on a prediction of the individual's risk of committing a future crime, or issuing a restraining order based on a prediction of the individual's risk of assaulting their partner.[8] Rather than isolating the effect of specific policy issues, this component aimed to assess systematic variation within individuals across the four decision types, while covering a wider range of policy issues beyond those used in the first component. This approach provides complementary correlational evidence to support the causal findings from the between-subject design.

Table A-5 confirms that all conditions are balanced across key demographic covariates, including gender, race, age, educational attainment, and technological literacy. To account for potential spillover effects, I randomized the order of the items presented to respondents within each matrix component.[9] Table A-7 shows the results remain robust when controlling for order effects.

The primary outcome of interest is the perceived appropriateness of using algorithmic rather than human decision-making across various contexts. This measure captures citizens'

[7] While each respondent evaluated all three policy domains, the specific decision type was independently randomized for each domain.

[8] See Table A-2 for item wordings.

[9] By asking first about the three policy domains (i.e., the first component), the experiment incentivizes respondents to compare ADS across policy domains rather than to focus on differences in the type of decision as the theory predicts. This approach thus provides a hard test for the theory.

general acceptance or rejection of algorithmic governance. The wording for the question reads as follows: "For each [decision], 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 "extremely inappropriate" to "extremely appropriate." Following my preregistered plan, I dichotomized these responses to facilitate substantive interpretation, coding respondents as 1 if they rated the use of ADS as appropriate (values above the midpoint labeled "indifferent") and 0 otherwise.[10]

To test the specific mechanisms proposed by the theory, the survey includes follow-up questions about the perceived accuracy and fairness of using ADS in each decision previously presented. Respondents rated both dimensions on seven-point scales, presented side by side in randomized order to prevent sequence effects. I dichotomized these measures following the same approach as the main outcome, coding responses above the midpoint as 1 (perceived as accurate/fair) and at or below the midpoint as 0.[11]

Results: Effect of Decision Type on perceived appropriateness

To test the theoretical predictions, I first examine how the perceived appropriateness of using ADS varies according to decision objective (assisting versus sanctioning) and target population (individuals versus collectives), using data from the between-subjects component, which independently varies these two dimensions across three distinct policy domains: policing, education, and child welfare. For each domain, I estimate the average treatment effects (ATEs) of these two dimensions.[12]

Figure 3 presents estimates from linear three probability models (LPMs) studying the effect of the two theoretical dimensions on the probability of viewing ADS appropriate in each policy area: education, policing, and child welfare. To minimize order effects, the analysis

[10] Table A-7 confirms that main effects remain statistically significant and substantively similar using alternative thresholds and the full scale.

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

[12] The mean values of the three dependent variables and associated confidence intervals by decision type are reported in Table A-4.

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

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.

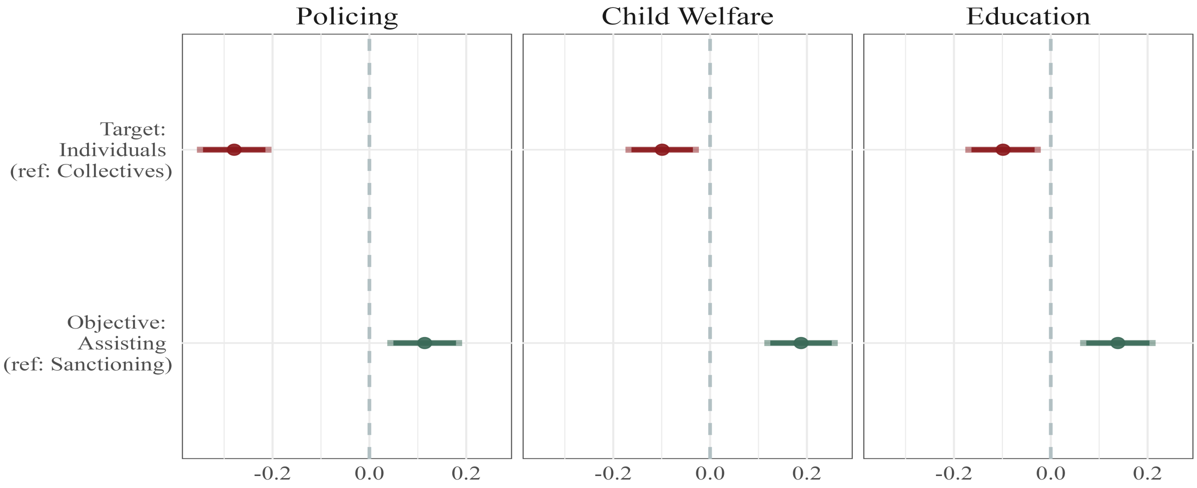
uses data from the first item randomly presented to respondents. Results are reported in columns 1, 4, and 7 of Table A-6.[13] Table A-9 confirms that all results remain substantively similar when using a multilevel analysis that accounts for both between and within-subject variation.

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 significant and substantively meaningful across all three policy domains ($p< 0.05$). For example, in child welfare, using an algorithmic system to assess the risk of child abuse in specific families rather than neighborhoods significantly decreases the probability of viewing such use as appropriate by 10 percentage points.

When looking at the objective of the decision, Figure 3 shows that ADS face significantly less opposition when used for assistance rather than sanctioning. Across all three

[13] To enhance statistical power, Table A-7 replicates the results using pooled data from all policy domains and controlling for the presentation order of the items.

Figure (3) Effects of decision type 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 target of the decision and its objective. 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.

policy domains, respondents were significantly more likely to view ADS as appropriate when implementing 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 results 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. [14] 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 component, 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. The analysis employs an LPM that regresses a binary outcome for the perceived appropriateness of using ADS on indicator variables for the two theoretical dimensions, as well as their interaction term, while controlling for the issue area and using fixed effects for respondent. The results, reported in Table A-13, strongly support the main findings and further validate the theory, showing a clear association between decision type and perceived ADS appropriateness across various policy domains.

Once again, respondents consider ADS significantly less appropriate in decisions involving sanctions rather than assistance ($p < 0.01$) and in decisions applying to individuals rather than collectives ($p < 0.01$). These findings remain consistent when using the alternative outcome measure (columns 2 and 4), and when using a linear mixed model with random intercepts for different policy issues and for each respondent (columns 4-6).[15]

[14] This analysis was not pre-registered.

[15] While findings demonstrate systematic variation in citizens' views of ADS based on target type and decision objective, as predicted by the theory, policy domains likely contribute independently to the remaining unexplained variation in attitudes. Theorizing and testing explanations for these domain-specific differences represents a promising direction for future research.

Additional Results: Fairness-Accuracy Trade-offs

The findings indicate that public opinion on ADS in government varies across and within policy domains, depending on (1) the decision objective and (2) its target population. Respondents are more likely to find ADS appropriate for assisting rather than sanctioning decisions and when targeting collectives rather than individuals.[16] 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 now turn to 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 share of respondents who deem ADS use fair with those who regard it as accurate, for each decision type and policy issue. Table A-14 formally tests these differences using paired t-tests.

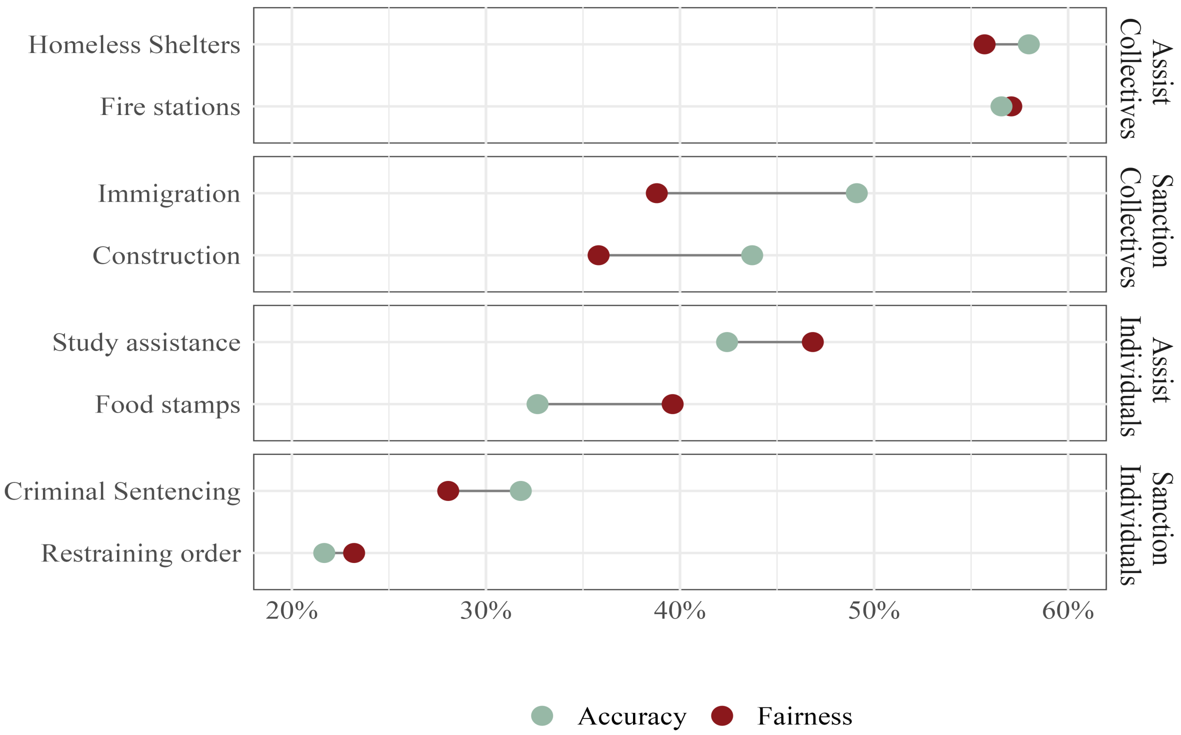
First, consistent with the main results, respondents view ADS most favorably in decisions that assist collectives, such as determining the location of a new fire station or homeless shelter. In this context, ADS receives the highest ratings on both fairness and accuracy, with no significant trade-off between the two. Similarly, the figure suggests minimal trade-off between fairness and accuracy in decisions that sanction individuals (as demonstrated by paired t-tests across both issue areas (p=0.264, Table A-14). The bottom panel shows that in both issue areas—criminal sentencing and restraining order—only 22% to 32% of respondents consider ADS both fair and accurate, approximately 26-34 percentage points lower than when ADS assist collectives ($p< 0.01$).[17] The absence of perceived trade-offs is particularly notable given that the question format explicitly asked respondents to compare these dimensions, potentially incentivizing them to identify differences.[18]

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

[17] The strong disapproval is also evident in the between-subject experiment. As Table A-4 shows, the percentage of respondents who view ADS in this context as appropriate is significantly low, ranging from 17% to 21%.

[18] One potential concern is that the results may reflect people's general aversion towards these decisions, regardless of the decision-maker. The Policy Evaluation experiment addresses this concern by isolating the effect of the decision-maker (human vs. algorithm) on policy support.

Figure (4) Perceived Fairness versus Accuracy, by Decision Type 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.

The remaining two decision types—assisting individuals and sanctioning collectives—elicit more ambivalent attitudes, with respondents weighing trade-offs between fairness and accuracy, consistent with theoretical expectations. For decisions assisting individuals (e.g., determining eligibility for public assistance programs or educational subsidies), the perceived fairness of ADS is significantly higher than its perceived accuracy (p<0.05 and p<0.001, respectively). Conversely, for decisions sanctioning collectives (e.g., enforcing regulations against illegal construction or work), perceived accuracy outweighs perceived fairness (p<0.001). This finding aligns with the theory, suggesting that although ADS may enhance accuracy in such contexts due to their capacity to process extensive data, deploying these systems to sanction rather than assist specific communities may be perceived as unfair.

The between-subjects analysis further supports these findings. Figure A-3 illustrates the predicted probabilities of respondents perceiving ADS as appropriate, fair, and accurate across decision types. The estimates are derived from a mixed-effects model that regresses these outcomes on indicators for the decision-type treatments using random intercepts for both the policy domain and respondent. Notably, when ADS are used to sanction collectives, respondents perceive them as significantly more accurate than fair. However, despite acknowledging this accuracy, respondents' overall perceptions of appropriateness remain closer to fairness than to accuracy, as reflected by the clearly overlapping confidence intervals between appropriateness and fairness measures, while confidence intervals for accuracy show no overlap with either. This pattern is consistent with the theoretical expectation that respondents may be reluctant to endorse algorithmic implementation for sanctioning decisions given their less reversible outcomes. The result also aligns with research showing that the public prior values of fairness when contemplating the use of ADS in government (Schiff, Schiff, and Pierson, 2021).[19]

Policy Evaluation Experiment

The results presented thus far reveal systematic variations in citizens' willingness to accept ADS across contexts, depending on the type of decision at stake. When asked directly, people particularly oppose AI tools that sanction individuals but accept them more readily when they inform assistance decisions, especially for collectives. What are the political implications of these views? Do citizens' views on ADS actually influence their support for the policy actions and interventions these systems inform?

To answer this question, I designed a second experiment where respondents evaluated identical policy proposals that differed only in who implemented them. Rather than explicitly asking about preferences for delegating decisions to ADS, this experiment tests whether

[19] While the current analysis provides suggestive evidence consistent with theory, it cannot determine the relative influence of accuracy versus fairness in shaping appropriateness perceptions, as the design treats these as separate outcome variables. Experimentally isolating these two considerations will be an important task for future research.

citizens actually care about the use of ADS in practice and consider the decision-making procedures when assessing concrete policy issues.

All respondents evaluated four randomly presented policy proposals: (1) prioritizing housing based on disability rather than waiting time; (2) selectively investigating allegations based on child abuse risk; (3) allocating police patrols to specific locations based on crime risk; and (4) providing extra funding for education programs for specific schools. Table 3 presents the wording of these policy proposals. As the table shows, each policy represents one of the four types of decisions in the two by two framework and builds on real-world ADS initiatives that government agencies currently promote or implement across the public sector.[20]

For each policy proposal, I independently randomized the identity of the decision-maker implementing the policy decisions while holding the policy content constant: a human officer in the control group and a predictive algorithm in the treatment group.[21] While the theoretical framework of this paper focuses on comparing algorithmic versus human decision-making approaches, many real-world applications involve a hybrid approach where algorithms assist, rather than fully replace, humans. To reflect these practices, the experiment includes a third condition where a human decision-maker uses algorithmic assistance. Appendix C.3.1 reports the full results, which I discuss in more detail later.

The key dependent variable measures support for the policy proposal. Respondents indicated their degree of support or opposition to a policy proposal on a five-point scale ranging from “strongly oppose” to “strongly support.” As preregistered, I recoded this scale to a binary measure with a value of 1 for positive answers (“strongly support” or “somewhat support”) and 0 otherwise.[22]

[20] I deliberately limited the experiment to four scenarios—one representative case from each decision type across different policy domains. This design choice balances experimental realism with practical constraints on respondent cognitive load while maintaining sufficient statistical power. Future research should expand this approach to include additional scenarios.

[21] Table A-15 reports summary statistics and balance tests across experimental conditions.

[22] This binary outcome allows me to capture potential shifts among initially indifferent respondents–a key segment that could determine political outcomes.

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.

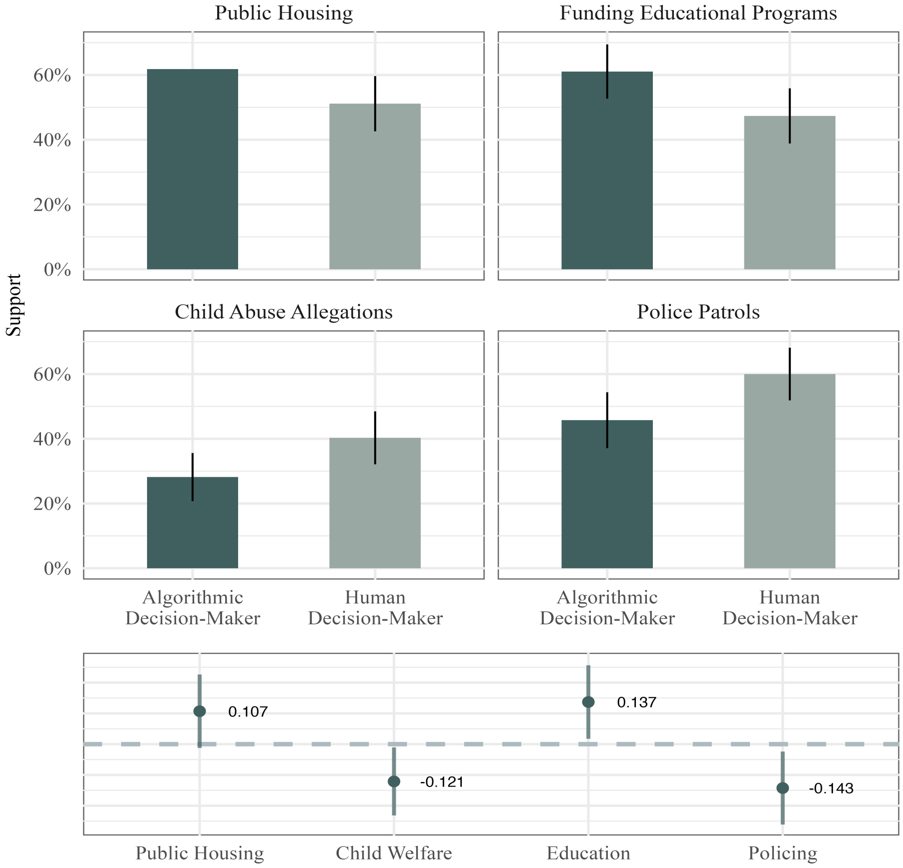
A central feature of the design directs respondents' attention to the policy reform itself rather than who implements it. Each scenario presents a clear policy dilemma that generates meaningful variation in responses even in the control group. For example, in the housing proposal, respondents evaluate whether to prioritize public housing based on disability status rather than waiting time. By having a clear policy tradeoff to evaluate, I prevent respondents in the treatment group from viewing the introduction of ADS itself as the main policy reform, which would cause treatment and control groups to interpret the question differently and potentially bias the results.

Results: Effect of decision-maker on policy support

Does using an ADS affect public support for policy proposals? I estimate the average treatment effect of the decision-maker on support for four policy proposals. To avoid spillover effects, the analysis relies on data gathered from the first scenario randomly presented to respondents. Table A-18 replicates the analysis using data pooled from all policy proposals, controlling for the presentation order.

Figure 5 shows the percentage of respondents who support each policy proposal as a

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 Table A-18 based on responses collected from the first scenario.

function of the decision-maker treatment: human versus algorithmic. Consistent with the theoretical expectations, the results suggest that citizens do not respond uniformly to the use of ADS in governance. The same proposal to allocate police patrols to specific areas received significantly less support when implemented by

On average, respondents were 14 percentage points less likely to support the proposal to allocate police patrols in specific areas when an algorithmic system, rather than a human officer, implemented it ($p<0.05$). This effect is both statistically and substantively significant, reducing support from 60% to below 46%. The bottom panel of Figure 5 shows a similarly negative effect in the child welfare context, which also involves sanctioning decisions—this time targeting individuals rather than collectives. Respondents were 12 percentage points less likely to support the proposal for selectively investigating child abuse allegations, when ADS assesses the risk of abuse or neglect in the family ($p<0.05$).

For the two proposals that involved assistance—prioritizing public housing and allocating funds for education programs—the figure reveals a different pattern. Respondents were almost indifferent to ADS in the context of assisting individuals. If anything, using a predictive algorithm instead of public housing officers increases overall support for prioritizing housing based on disability rather than waiting time.

I find even stronger effects in the context of assisting collectives, specifically when deciding which schools should receive funding for drug and alcohol education programs. The share of citizens expressing at least some support for this policy rises by nearly 14 percentage points when a predictive algorithm—rather than 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.[23]

I also examine whether deploying algorithmic systems to support (rather than replace) human decision-makers leads to a different effect on public support. Table A-17 compares the average treatment effects of the pure ADS and hybrid conditions, showing little difference across policy domains. The only exception to this overall pattern is in the policing domain. While using a predictive algorithm on its own reduces support compared to human decision-makers, support for the policy increases once the algorithm is used as a complementary tool ($p<0.01$). This pattern aligns with prior evidence that, in decisions sanctioning collectives, people perceive a trade-off: ADS can be relatively accurate yet less fair. This pattern aligns with earlier evidence of a trade-off in using ADS for decisions that sanction collectives,

[23] Table A-19 shows similar results when using alternative outcome measures.

which respondents tend to consider as less fair yet relatively accurate. In such decisions, using algorithms as a supportive tool while keeping “humans in the loop” appears to be an attractive solution. It provides more accurate assessments without sacrificing the human oversight that is crucial in decisions with less-reversible outcomes.

Finally, to test whether the decision 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 overall effect of ADS is negative and statistically significant, suggesting respondents are less likely to support policies when ADS implement them. However, the negative effect of the decision maker is offset and even reversed in policy proposals involving decisions about assisting collectives.[24]

Taken together, the two experiments support the theory that people are especially sensitive to human presence in sanctioning decisions, which often carry less reversible consequences for both individuals and collectives. Adopting ADS in these contexts can significantly reduce the overall support for policy decisions and actions.

Conclusion and Implications

This article puts forward a theoretical framework and leverages a set of survey experiments to explain when and why citizens resist or accept the use of AI-based algorithmic decision systems in governance. The theory calls for distinguishing between four types of decisions when evaluating such uses. The experimental results provide strong support for this theory. Using evidence from a broad range of policy domains and 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 that opened this article, the public’s rejection of algorithmic risk assessment for pretrial detention likely reflects citizens’ resistance to using ADS in

[24] A potential concern is that variation across contexts might stem from policy domain differences rather than decision type. However, as shown in Table A-9, when controlling for decision type in the Decision Type experiment, differences between domains are minimal and not statistically significant.

sanctioning decisions with potentially irreversible consequences for individuals—contexts where algorithms are perceived as both less fair and less accurate than human decision-makers.

These findings offer important practical implications for the responsible governance of AI development and implementation. The framework provides a systematic approach for identifying ex-ante where AI-based tools will likely be considered appropriate and where they might provoke resistance. Current governance efforts largely overlook citizens' perspectives—those who ultimately bear the consequences of AI-based decisions without the ability to opt out. As this study demonstrates, even technically and ethically sound AI applications may not be politically feasible without public acceptance. The analysis reveals that public support for policy decisions among individuals fell significantly when implemented by an algorithm rather than by a human decision-maker. This dynamic has already manifested in recent high-profile cases where public opposition forced governments and municipalities to abandon ADS initiatives (e.g., Austen and Wakabayashi, 2020; Weale and Stewart, 2020).

Furthermore, the finding that identical algorithmic systems can be accepted as legitimate in certain decisions yet rejected in other–theoretically predictable—contexts, underscores the limitations of current efforts to develop one-size-fits-all standards for algorithmic fairness and accuracy. Design efforts and implementation strategies may benefit from context-specific approaches that better address citizens' concerns, values, and expectations regarding algorithmic governance.

The findings also highlight the potential appeal of hybrid approaches, where algorithmic systems support rather than fully replace human decision-makers. When there is a tradeoff between fairness and accuracy, particularly in decisions that sanction collectives, combining algorithmic accuracy with human oversight can enhance public acceptance. Future research should further explore how different configurations of human-algorithm interactions influence trust and legitimacy across various contexts.

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.

While the 2-by-2 framework provides a useful starting point for understanding contextual variation, additional factors likely shape public preferences. For instance, while this study focuses on whether the algorithmic decision targets individuals or collective cases, another factor that might be relevant is how 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, 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 expand, more stakeholders 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 research to understand the evolving politics of using AI and data-driven algorithms in government.

Overall, as this study makes clear, the growing integration of AI-based tools in governance touches on the very core of democracy—how we make public decisions. It raises fundamental questions regarding legitimacy and accountability, inspiring a research agenda in political science on the political repercussions of this major technological change.

Acknowledgments

I would like to thank Yotam Margalit, Neil Malhotra, Alex Kuo, Omer Solodoch, Roni Porat, Roi Levi, and Patrick Egan for their helpful comments and suggestions. I also thank five anonymous reviewers for their feedback.

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.

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.

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 (1980). “Street-Level Bureaucracy (New York: Russell Sage)”. In: *Moving Toward Mixed Service Delivery* 31.

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.

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

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*.

Biographical Statement

SHIR RAVIV is a Post Doctoral Researcher at the Data Science Institute, Columbia University, New York, NY 10027.