What Do People Want from Algorithms? Public Perceptions of Algorithms in Government

Amit Haim[1] and Dvir Yogev[2]

[1]Tel Aviv University

[2]UC Berkeley

Author Note

Amit Haim https://orcid.org/0000-0002-0832-6924 Dvir Yogev https://orcid.org/0000-0003-0466-9804

Authors are listed alphabetically and acknowledge no conflict of interest. We thank the Center for American Democracy in the Institute for Research in the Social Sciences, Stanford University, for funding to conduct this project. We thank participants at the UMass Amherst Workshop on New Directions in Law and Society 2022 and the Conference of Empirical Legal Studies at the University of Virginia 2022 for helpful feedback. Many thanks to the anonymous reviewers and journal editors for helpful comments. The data used in this study is publicly available at https://osf.io/ax83q/.

Correspondence concerning this article should be addressed to Dvir Yogev. E-mail: dyo@berkeley.edu

Published in Law and Human Behavior, 49(3), 263–280. https://doi.org/10.1037/lhb0000614

©American Psychological Association, 2025. This paper is not the copy of record and may not exactly replicate the authoritative document published in the APA journal.

Abstract

Objectives: This study examines how specific attributes of Algorithmic Decision-Making Tools (ADTs), related to algorithm design and institutional governance, affect the public's perceptions of implementing ADTs in government programs.

Hypotheses: We hypothesized that acceptability varies systematically by policy domain. Regarding algorithm design, we predicted that higher accuracy, transparency, and government in-house development will enhance acceptability. Institutional features were also expected to shape perceptions: explanations, stakeholder engagement, oversight mechanisms, and human involvement are anticipated to increase public perceptions.

Method: This study employed a conjoint experimental design with 1,213 U.S. adults. Participants evaluated five policy proposals, each featuring a proposal to implement an ADT. Each proposal included randomly generated attributes across nine dimensions. Participants decided on the ADT's acceptability, fairness, and efficiency for each proposal. The analysis focused on the average marginal conditional effects (AMCE) of ADT attributes.

Results: A combination of attributes related to process individualization significantly enhanced the perceived acceptability of using algorithms by government. Participants preferred ADTs that elevate the agency of the stakeholder (decision explanations, hearing options, notice, and human involvement in the decision-making process). The policy domain mattered most for fairness and acceptability, while accuracy mattered most for efficiency perceptions.

Conclusions: Explaining decisions made using an algorithm, giving appropriate notice, a hearing option, and maintaining the supervision of a human agent are key components for public support when algorithmic systems are being implemented.

Public Significance Statement

Artificial intelligence and algorithmic tools are increasingly used in government

institutions, reshaping relationships with the public. However, their impact on public perception remains unclear. Government institutions must consider key attributes of algorithmic tools that influence public acceptance. Features enhancing human agency enhance acceptance, while costly programs reduce it. Public support is stronger in less politically controversial domains. This study informs practitioners and researchers on AI acceptability in government, showing that human communication around AI decisions improves public perception.

Keywords: Algorithms, public decision-making, government, experimental

What Do People Want from Algorithms? Public Perceptions of Algorithms in Government

Algorithmic decision-making tools (ADT), including machine learning models and other forms of automation, are becoming increasingly common in public government programs (Engstrom and Ho, 2020; Levy et al., 2021). This trend was enhanced when artificial intelligence (AI) swept over public discourse (Pew, 2023), leading to a widespread interest in utilizing AI for governance. Proponents of the use of ADTs in government argue they can reduce errors, speed up processing, and eliminate human biases in public bureaucracies (Kleinberg et al., 2016; Kolkman, 2020; van der Voort et al., 2019). Critics, however, point out potential concerns and risks, including heightened racial and socioeconomic biases, errors, and opacity (Medaglia et al., 2023; Pasquale, 2015; Sætra, 2020). They highlight that hasty or flawed adoption of ADTs may culminate in public policy fiascos, such as occurred when New York City launched a bot as a "one-stop shop" for business owners in need of licensing support which provided false legal information (Associated Press, 2023).

Opinion polls show the American public is concerned and suspicious of implementing algorithms in government operations (Pew, 2023). When asked, "How *acceptable* do you think using people's information for the purpose of using computer programs, or algorithms, to determine who should be eligible for public assistance?" 55% responded it was somewhat or very unacceptable, with only 28% holding the opinion it was somewhat or very acceptable. Past public surveys have also indicated negative views on AI and institutions managing the use of AI (Cave et al., 2019; Smith, 2019; Zhang and Dafoe, 2020). There is thus an imminent need to understand better how the design of governmental ADTs may mitigate some of the skepticism.

Given these potential promises and problems, there has been a surge of scholarly interest in the public's attitudes toward implementing ADTs in government. However, currently most studies have focused on juxtaposing and comparing human and algorithms to assess the public's reaction (Bambauer and Risch, 2021; Chen et al., 2022; Hermstrüwer and Langenbach, 2022; Starke and Lünich, 2020; Waldman and Martin, 2022; Yalcin et al., 2023); while in reality, the more common and realistic scenario is a mix of human bureaucrats and ADTs (K. J. Schiff et al., 2024).

AI tools for government use can be designed, implemented, and utilized in many ways, and a range of attributes and characteristics may influence how willing the public is to accept the integration of ADTs into government programs. Such attributes pertain both to the algorithm design (e.g. what kinds of inputs are provided) (Grimmelikhuijsen, 2023; R. P. Kennedy et al., 2022; D. S. Schiff et al., 2022) and to the institutional design (e.g. what kind of human overview is in place) (Aoki, 2020; Chatterjee et al., 2022; Ingrams et al., 2022; Miller and Keiser, 2021). Only several studies have focused more closely on aspects of governmental use of ADT that may influence public attitudes, such as effectiveness (König et al., 2024), transparency (König et al., 2024), and domain or context (Wenzelburger et al., 2024). To date, no comprehensive study has taken stock of the whole range of technical and institutional design attributes. Against this backdrop, we inquire what makes using ADTs acceptable in the public's view. Put otherwise, which attributes or features of a governmental program proposal that utilizes an ADT will make it more likely to be perceived as favorable in government use?

The present study leverages a conjoint design, which has become popular in policy preference research (Bansak et al., 2021), and more recently in studies of public attitude toward governmental ADTs (Gaozhao et al., 2024; R. Kennedy et al., 2024; R. P. Kennedy et al., 2022; König et al., 2024; Waggoner et al., 2019). A conjoint design allows to experimentally identify the effects of different attributes on participants' preferences, and allows measuring potential trade-offs between multifarious policy dimensions. We construct five policy proposal scenarios varying the attributes that compose the proposals. We identify the attributes of ADTs that matter most for public acceptability.

Public Attitudes toward Algorithms in Government

Understanding public attitudes toward using ADTs in governmental programs is becoming increasingly important. It is, however, challenging to untangle attitudes towards institutions, programs, specific policy instruments, and general views about the government. Additionally, public views about a particular technology are often dynamic and may interact with other perceptions of government.

To a large extent, public agencies rely on public support and trust in the government to implement policy successfully (Bitektine, 2011; Dacin et al., 2007). Legal scholarship is rife with debates over the legitimacy of the "administrative state" (Havasy, 2023; Mashaw, 2018; Rahman, 2017), a major topic for constitutional, administrative, and other legal scholars that are concerned with political accountability and democratic representation. However, this scholarship has largely not been grounded in empirical research (Feinstein, 2024; Johnson et al., 2014; Stiglitz, 2022). Social scientists, conversely, have tended to be more focused on the empirical mechanisms of public support and compliance with government policies (Tyler, 2006).

In this vein, a dominant strand of research in public administration has focused on public perceptions of acceptability (Hutchison and Johnson, 2011; Levi and Sacks, 2009; Levi et al., 2009), which has been described as the "sense of obligation or willingness to accept their authority" (Risse and Stollenwerk, 2018, p. 404) or "a favorable judgment on the acceptability of an organization's actions, based on their utility, justice, and appropriateness" (Díez-Martín et al., 2021, p. 3). The theoretical literature regarding public institutions' legitimacy is broader and encompasses more complex judgments and assertions on institutions, yet it is vast and often lacks consensus (Beetham, 2013; Díez-Martín et al., 2021; Schoon, 2022; Suddaby et al., 2017). Acceptability, conversely, is more focused on the ability to carry out policy with the public's support (Deephouse and Suchman, 2008; Jackson, 2018; Risse and Stollenwerk, 2018), and thus allows to gauge the legitimacy of government actions.

A central insight to this body of literature, epitomized in the procedural justice approach (Tyler, 2006), suggests that the proper design of procedures can positively influence acceptance, regardless of substantive outcomes. Studies have applied this notion to legal institutions, examining perceptions of procedures and people's willingness to accept government policies (Gibson et al., 2005; Gibson et al., 2014). Consequently, positive attitudes toward governmental programs' design may increase citizens' tendency to comply with governmental authority (Hibbing and Theiss-Morse, 2001; Meier and Bohte, 2007; Tyler and Lind, 2001).

Notably, introducing technologies in the interaction with government agencies may transform public perceptions and is therefore consequential for public administration and endeavors to implement algorithmic tools in government. In recent years, researchers have begun paying attention to the incorporation of Algorithmic Decision-making Tools (ADTs) in government, measuring public attitudes and perceptions towards these scenarios. Scholars have entertained the effects of replacing humans with AI algorithms in legal proceedings and other governmental functions, primarily finding that people perceive algorithmic decision-making as less fair and less acceptable (Chen et al., 2022; Yalcin et al., 2023).

Some studies have focused more closely on the design characteristics of government decision-making and their effects on perceptions of acceptability. Nevertheless, the vast majority of these studies have purported to measure the variation in perceptions between algorithms and humans performing the same task (Bansak and Paulson, 2023; Gaozhao et al., 2024; Hermstrüwer and Langenbach, 2022; Starke and Lünich, 2020; Waldman and Martin, 2022). Yet, in reality, the more likely scenario is for government programs to adopt a mix of humans and ADTs (K. J. Schiff et al., 2024), such that assessing displacement does not capture the full scope of the matter.

More recently, there has been a much-warranted surge in studies on the public perception of government use of ADTs, with a more granular focus on the features and attributes of a hybrid human-algorithm apparatus (Aoki, 2020; Grimmelikhuijsen, 2023; R. P. Kennedy et al., 2022; König et al., 2024; Miller and Keiser, 2021; Miller et al., 2022; D. S. Schiff et al., 2022; G. Wang et al., 2023; Willems et al., 2022). Kennedy et al. (2022), for example, find predominantly positive attitudes towards the use of algorithms in government, contrary to the notion of algorithm aversion that has gained traction (Dietvorst et al., 2015). They find that certain factors drive those attitudes, including having a human in the loop and a noticeably positive developer reputation.

In general, those studies have provided mixed evidence on the public's views: while some have found that the public is often not reactive to design factors that theoretically should have mattered, such as administrative capacity (Grimmelikhuijsen, 2023; König et al., 2022), others have found that factors such as the domain or context in which the algorithm is used significantly affect public views (Wenzelburger et al., 2024).

The Attributes of Algorithmic Decision-making Tools

A governmental program adopting an ADT faces many design attribute choices, both technical and institutional. Various regulations and legal requirements have been a focus of attention for policymakers worldwide over the last decade. For example, regulations may require an opt-out option from automated decision-making (GDPR Art. 22) or to ensure human oversight over automated decisions (EU AI ACT Art. 14). Several empirical studies have touched on these points in the context of public perceptions, such as how the kinds of inputs that are supplied to the algorithm shift public perceptions (Grimmelikhuijsen, 2023; R. P. Kennedy et al., 2022; D. S. Schiff et al., 2022).

Generally, studies have found that both the "openness" of algorithmic systems - as is expressed in transparency and public information - and their integration with human expertise is crucial for fostering trust. Based on previous literature – empirical and theoretical – we identify a list of attributes that may influence public perceptions and classify them into one of two broad categories: algorithm design and institutional design.

Algorithm Design

Algorithm design has been the focus of attention for the last decade and a half in legal, data science, and policy circles (Kroll et al., 2017; Wachter et al., 2017) and revolves around attributes that have to do with the development, design, and deployment of algorithmic systems.

Accuracy. An important attribute is an algorithm's accuracy, which shapes people's willingness to accept its use. König et al. (2024) find that subjects will prefer even small performance gains over transparency and stakeholder involvement, in the context of policing and healthcare decisions. Likewise, Shin (2020) finds that accuracy plays an important role in users' perceptions of an algorithm. However, some have found no significant connection between accuracy and favorable attitudes: Wenzelburger et al. (2024), for example, have found that the algorithm's performance is negligible compared to other factors such as the policy domain and reputation of the specific agency.

Public information. Many studies have pointed out that the availability of public information about the algorithm leads to more positive attitudes (Aoki, 2020; Grimmelikhuijsen, 2023; R. P. Kennedy et al., 2022; Miller and Keiser, 2021; D. S. Schiff et al., 2022). However, this attribute's importance may be overblown since other studies that considered trade-offs with other values, such as effectiveness, found it to matter less (König et al., 2024).

Development. Additionally, we identify the algorithm's source of development as an important attribute. Governments have much leeway in procuring algorithmic systems from third-party vendors or developing them in-house (Mulligan and Bamberger, 2019). Despite being highlighted in the theoretical literature (e.g., Engstrom and Haim, 2023), it has not been studied empirically.

Institutional Design

The institutional design choices have been especially prominent in the legal and theoretical literature. It is primarily concerned with the governance of algorithms within a public policy framework, highlighting their impact on social life and the conditions under which institutions and policy can regulate the use of algorithms and AI (Coglianese and Lehr, 2019; Katzenbach and Ulbricht, 2019). Scholars, advocates, and activists have proposed strategies to ensure and enhance administrative accountability (Joshi, 2021; Ranchordas, 2021). Some of the discussed and debated tools include independent audits, external evaluations, institutional transparency, and public outreach and engagement initiatives. In the empirical context of public perceptions, there has been some work in recent years to uncover the effects of such proposals on the public's perception, for instance, with regards to human review or overview of algorithmic decisions (Aoki, 2020; Chatterjee et al., 2022; Ingrams et al., 2022; Miller and Keiser, 2021; Starke and Lünich, 2020).

Human involvement. Human involvement in decision-making is often referred to as "human in the loop." Many studies have found that ensuring a human is involved in the algorithmic decision-making process in a reviewing role increases public support (Aoki, 2020; Busuioc, 2021; R. P. Kennedy et al., 2022; Keppeler, 2024; Waldman and Martin, 2022) and increases perceived fairness of decisions (Lee et al., 2019). It is noteworthy, however, that most have pitted human decision-makers versus algorithms and not considered them in tandem. Several aspects may be included in human involvement, which includes a variety of roles a human may play (such as a review of a decision by an algorithm or an algorithm providing recommendations), whether a subject may request an in-person hearing, and more.

Oversight. Studies, mostly theoretical, have pointed out that oversight mechanisms are important, such as audits or impact assessments by experts who have the know-how to assess the impact of the algorithm (Grimmelikhuijsen, 2022; Kaminski and Malgieri, 2020; Oetzel and Spiekermann, 2014). **Stakeholder engagement** is closely related to this aspect, which has been touted as garnering public support (König et al., 2024l).

Explanation. Providing reasons for an algorithmic prediction or recommendation has been widely discussed, especially given the challenge of explainability in machine-learning systems (Kaminski, 2019). Several studies have found that providing explanations enhances trust in algorithms (Glikson and Woolley, 2020; Grimmelikhuijsen, 2022; Shin, 2021). Recently, Henning and Langenbach (2024) found that providing reasons for the algorithmic output enhances the perceived fairness of decision-making, regardless of whether humans or machines make decisions and that sufficiently individualized reasoning largely mitigates the human-automation fairness gap.

These attributes - both related to the algorithm and the institutional design - are purported to affect public perceptions of government ADT implementation. Previous studies have focused on a subset of attributes, such as transparency or the presence of a human in the loop. To date, however, studies have mostly fallen short of comprehensively studying the range of attributes, while also accounting for potential tradeoffs. The current study was designed with this in mind, as we explore in the following section.

Overview of Current Study and Hypotheses

The current study examined which attributes increase the favorable public perception of using algorithmic decision-making tools (ADTs) in government programs. We employed a conjoint experimental design to gauge the effect of various attributes of ADTs on the perceptions of the acceptability of administrative decision-making. This design has recently gained popularity in policy preference research (Bansak et al., 2021). We constructed profiles of policy proposals for participants to evaluate, systematically varying both algorithm design and institutional design attributes. Based on our literature review, we developed several hypotheses concerning both categories of attributes:

Firstly, we expected policy domain-specific effects:

- **H1 (Policy Domain Effects):** We hypothesized that the importance of specific attributes for acceptability will vary systematically by policy domain (Wenzelburger et al., 2024), with human involvement having stronger positive effects in high-stakes domains and explanation provision being more important for decisions affecting individual rights or benefits.

Algorithm Design Hypotheses

- **H2 (Accuracy):** We hypothesized that profiles with higher algorithm accuracy levels will be rated more acceptable than those with lower accuracy levels. This is consistent with findings from König et al. (2024) and Shin (2020), though some studies found mixed results (Wenzelburger et al., 2024).

- **H3 (Public Information):** We hypothesized that profiles with greater transparency and public information about the algorithm will be rated as more acceptable than those with less information, as supported by multiple studies (Aoki, 2020; Grimmelikhuijsen, 2023; R. P. Kennedy et al., 2022; Miller and Keiser, 2021; D. S. Schiff et al., 2022).

- **H4 (Development Source):** We hypothesized that profiles describing algorithms developed in-house by government agencies will be rated as more acceptable than those developed by third-party vendors, drawing on theoretical literature (Engstrom and Haim, 2023; Mulligan and Bamberger, 2019). However, this has not been extensively tested empirically.

Institutional Design Hypotheses

- **H5 (Explanation Provision):** We hypothesized that profiles where algorithms provide explanations for their decisions will be rated as more acceptable than those without explanations, as found in several studies (Glikson and Woolley, 2020; Grimmelikhuijsen, 2022; Henning and Langenbach, 2024; Shin, 2021).

- **H6 (Hearing and Notice):** We hypothesized that profiles with greater stakeholder engagement in algorithm development and implementation will be rated as more acceptable than those with limited engagement (König et al., 2024).

- **H7 (Oversight Mechanisms):** We hypothesized that profiles featuring robust oversight mechanisms (e.g., independent audits, impact assessments) will be rated as more acceptable than those with limited oversight, drawing on theoretical literature (Grimmelikhuijsen, 2022; Kaminski and Malgieri, 2020; Oetzel and Spiekermann, 2014).

- **H8 (Human Agent Involvement):** We hypothesized that profiles with higher levels of human involvement in the algorithmic decision process will be rated as more acceptable than those with minimal involvement, consistent with numerous studies (Aoki, 2020; Busuioc, 2021; R. P. Kennedy et al., 2022; Keppeler, 2024; Waldman and Martin, 2022).

Through this conjoint analysis, we aimed to identify which attributes of ADTs most significantly influence public perceptions of acceptability and examine how these preferences may vary across different governmental contexts and policy domains.

Method

Participants

We sampled 1213 participants from the U.S. adult population, recruited through the Prolific online platform. Participants were paid according to a $10 hourly rate. [REDACTED FOR REVIEW] University IRB Protocol number 65006, exempted from review (May 9, 2021). OSF Pre-registration https://osf.io/gz8m2/?view_only=8ab9b5c695284187ba03a69798b188db. The data used in this study is publicly available at https://osf.io/ax83q/. We chose quota sampling to produce a representative sample matched to the US population on age, gender, and political views. Recent research demonstrates the suitability of online recruitment platforms for evaluating social scientific theories (Coppock, 2023). We gathered the sample using Prolific, which has been identified as a reliable and cost-effective platform for data collection. Recent studies have underscored the high data quality provided by Prolific, with participants demonstrating superior attention to study questions, comprehension of instructions, and honesty in responses (Peer et al., 2022; Stagnaro et al., 2024). Importantly, these findings have been independently verified, adding to their credibility (Douglas et al., 2023). Compared with other platforms, Prolific consistently outperforms in terms of data quality and cost (Douglas et al., 2023; Peer et al., 2022).

Materials and Measures

Main Outcome: Acceptability. The principal outcome in our design is the acceptability of a specific policy proposal ("Do you think this proposal is acceptable?" binary Yes or No question; see Figure 1). This concept of acceptability, often used in past studies (Gaozhao et al., 2024; König et al., 2024; Starke and Lünich, 2020) and public opinion surveys (Pew, 2023), is understood as a subjective evaluation that indicates a positive preference. This measure is aligned with research in social science that focuses on the willingness of the public to accept the authority wielded by a public institution (Levi and Sacks, 2009; Levi et al., 2009; Risse and Stollenwerk, 2018) and their ability to carry out policy with popular support (Deephouse and Suchman, 2008; Jackson, 2018; Risse and Stollenwerk, 2018). The study's pre-analysis plan originally proposed a Likert scale for the outcome; however, we were prompted to reconsider the measurement scale due to potential issues tied to Likert scales in a study like ours, such as response bias, the challenges associated with equidistance assumptions, and central tendency bias. The binary choice, a preferred method in the traditional conjoint design Knudsen and Johannesson, 2019, ensures a clear, bias-minimized measure of participants' attitudes without jeopardizing the integrity of the pre-analysis plan, which contains our analysis script created with simulated data pre-data collection.

Alongside acceptability, we also considered two additional outcomes which are discussed in Supplement C: Fairness ("Do you think this proposal is fair? Fair means treating everyone equally and justly") and Efficiency ("Do you think this proposal is efficient? Efficient means achieving results without wasting time or resources").

Attributes. We divided the ADT attributes into Institutional Design and Algorithm Design attributes. Table 1 presents the full scope of attributes and the breakdown of levels. We elaborate on the theoretical underpinnings of the attributes in the background section .

All attributes are variables with multiple levels (mostly four). Several of the attributes are categorical with no inherent order within levels, and others are ordinal, depending on the specific attribute. **Categorical** (no inherent order): Algorithm development; Oversight; Public engagement; Human-agent role. For this attribute, while the levels range from 'Review algorithmic decisions' to 'No agent involved', we treat this as categorical rather than ordinal since the intermediate options ('Make decisions using algorithmic recommendations' and 'Upon request, an agent reconsiders') represent qualitatively different approaches rather than points on a continuum of human involvement. **Ordinal** (clear ordering from less to more): Publicly available information (none → simple overview → technical information); Notice (none → basic notification → consent option); Explanation of decision (none → bottom line → detailed explanation → full algorithmic report); Cost ($2m → $5m → $10m); Accuracy (75% → 85%, with "Unknown" as a separate category). The level of "Not Shown" was not shown to participants in instances that were randomly chosen to have the attribute omitted, as distinct from presenting to participants the level "Unknown".

Policy Domains. The policy domain or context in which the algorithm is deployed may be important in determining public support. For instance, people tend to accept algorithms more if they carry out technical and routine tasks as opposed to tasks that

Table 1
Attribute Levels

Attribute	**Levels**
Publicly available information	A quick, simple overview; Technical information; None*
Accuracy	75%; 85%; Unknown; Not shown*
Algorithm development	In-house, no outside partners*; Purchased from a private company; Partnership with academics; Non-profit organization
Human-agent role	Review algorithmic decisions; Make decisions using algorithmic recommendations*; Upon request, an agent reconsiders the algorithmic decision; The algorithm makes an immediate decision, no agent involved
Explanation of the decision	Only the bottom line; Detailed explanation of reasons and process; Original algorithmic report; None*
Notice	That personal information was accessed; That there was an algorithmic assessment; The person will have an option to consent to the algorithmic assessment; None*
Hearing	No, only legal appeal; Video hearing by request; In-person hearing by request*; No, only written
Oversight	Evaluation by university researchers; The agency will publish public reports; The agency has an oversight board with public experts; The state accountability office will publish reports; A quality review team reviews cases at random*
Public engagement	Public town-hall meetings; Written public comments; Only through political representatives; None*
Cost	$2m*; $5m; $10m

Note: All attributes are variables with multiple levels. Categorical attributes include Algorithm development, Oversight, Public engagement, and Human-agent role. For the Human-agent role, while the levels range from 'Review algorithmic decisions' to 'No agent involved,' this attribute is treated as categorical rather than ordinal because the intermediate options ('Make decisions using algorithmic recommendations' and 'Upon request, an agent reconsiders') represent qualitatively different approaches rather than points on a continuum of human involvement. Ordinal attributes include Publicly available information, Notice, Explanation of decision, Cost, and Accuracy. The "Not Shown" level in Accuracy refers to instances where the attribute was randomly omitted from presentation to participants, distinct from the level "Unknown." The starred (*) levels are reference categories in the analysis.

involve human judgment (Aoki, 2020; Chatterjee et al., 2022; Ingrams et al., 2022; Miller and Keiser, 2021; Wenzelburger et al., 2024). Relatedly, Raviv (2023) found, using survey experiments in the U.S. population, that people are averse to the use of ADTs in decisions that sanction rather than assist and decisions that make inferences about individual people. Wenzelburger et al. (2024) argue that the domain, or context, of an algorithm, may be "more important for the acceptance of algorithms than the technical features of the systems themselves." (Wenzelburger et al., 2024, p. 42) They find that the personal importance and values at stake shape general support, and that trust in an organization predicts acceptance of its use of algorithms. Relatedly, Margalit and Raviv (2023) and Arnesen et al. (2024) find that familiarity with the policy domain increases favorable views of the use of ADTs.

We used policy domains because judging a program in the abstract is less informative and less resembling real-life opinions. A conjoint design is an effective method for eliciting preferences in specific scenarios. We chose five policy domains that reflect different levels of moral and political salience to mitigate idiosyncrasies, investigate heterogeneous effects, and substantiate generalizability. Criminal justice was chosen to reflect high moral salience; unemployment benefits were chosen because of economic ideology salience; refugee policy represents a politically salient arena; building permits are a more neutral policy domain, although we recognize no policy domain is devoid of political baggage.

Moderators. We presented participants with the following definition: "Artificial Intelligence (AI) algorithms refer to computer systems that perform tasks or make decisions that usually require human intelligence. Algorithms can perform these tasks or make these decisions without explicit human instructions."

Participants were then asked to indicate how much they had heard about AI before taking the survey. We also measured the participants' reported computer science education and experience in computer programming. Following the comprehensive work of Zhang and Dafoe (2020) and O'Shaughnessy et al. (2021), we asked about the perceived benefits of

Table 2
Policy Domains

Policy Domain	Presented Description
Building Permits	A building permit gives you legal permission to start a new construction or remodeling project. Most home improvement projects require a permit, but there are exemptions. The county is considering the use of a new AI to determine whether a project is exempt.
Unemployment Benefits	If you lost your job or had your hours reduced, you may be eligible for unemployment benefits. The state is considering the use of a new algorithm that will determine the eligibility of new claims.
Release on Parole	When a person is sentenced to ’life with the possibility of parole’ they are eligible for a parole hearing in which a board decides if they can be released. Some jurisdictions use AI, or “risk assessment tools,” to inform board decisions. The state is considering the use of such an algorithm to determine if one should be granted parole.
Refugee Resettlement	People who have been granted refugee asylum in the US are resettled in different places across the country. The government is considering the use of a new AI to match refugees and locations to yield positive integration outcomes.
Small Business License	If you wish to start a small business, you need to get a license from your local government. The city is considering the use of a new AI chatbot that will check eligibility qualifications.

AI, the benefits of new technologies, and the need to regulate AI according to the participant. We code responses according to the following variables: AI knowledge (Continuous scale, 0-1); CS education (Binary, 0 = no college-level CS education, 1 = college-level or higher)); CS experience (Binary, 0 = no/little experience, 1 = some/a lot); Technology/AI attitude measures (Continuous scales, 0-1).

We also collected demographic information known to moderate attitudes toward AI: sex, education, and computer science knowledge and experience.

Experimental Design

Our study employed a single-profile conjoint experimental design with a fully randomized factorial structure.

Independent Variables: We manipulated ten independent variables (attributes) across multiple levels as detailed in Table 1. These included five algorithm design attributes (Publicly available information, Accuracy, Algorithm development, Cost, Human-agent role) and five institutional design attributes (Explanation of the decision, Notice, Hearing, Oversight, Public engagement).

Factorial Structure: We utilized a $3 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 5 \times 4 \times 3$ factorial design, representing the number of levels for each attribute. For the Accuracy attribute, participants could see one of three levels (75%, 85%, Unknown) or the attribute could be omitted entirely.

Within-Subjects Design: Our design was fully within-subjects. Each participant evaluated five separate profiles, with each profile presented in one of the five policy domains described in Table 2. For each profile, all ten attributes were randomly and independently varied, creating a balanced distribution of attribute levels across the sample.

Profile Evaluation: Each participant viewed and evaluated a total of five profiles (one per policy domain). The order of policy domains was randomized across participants to control for order effects.

Outcome Measurement: For each profile, participants provided a binary acceptability judgment (Yes/No) as our primary dependent variable, with secondary measures of perceived fairness and efficiency (discussed in Supplement C).

Procedure

Participants completed the study online. Upon accessing the survey platform, they first read and acknowledged an informed consent statement. After providing consent, they completed a series of demographic questions, including age, sex, race, education, computer science knowledge, and political views and affiliation.

Following the demographic section, participants answered questions about their prior knowledge and attitudes toward artificial intelligence (AI). They indicated how much they had heard about AI and their agreement level with various statements related to AI's societal impact, regulation, and personal views.

Next, participants responded to items assessing their attitudes toward technology in general. They rated their agreement with statements about technological optimism, potential dangers, and perceived societal benefits of new technologies.

In the subsequent section, which was the main conjoint part of the study, participants evaluated five hypothetical government applications of AI. Each participant was shown, consecutively, five policy domains in a randomized order: parole decisions, refugee resettlement, unemployment benefits eligibility, building permit approvals, and small business licenses. The policies they read are presented verbatim in Table 2.

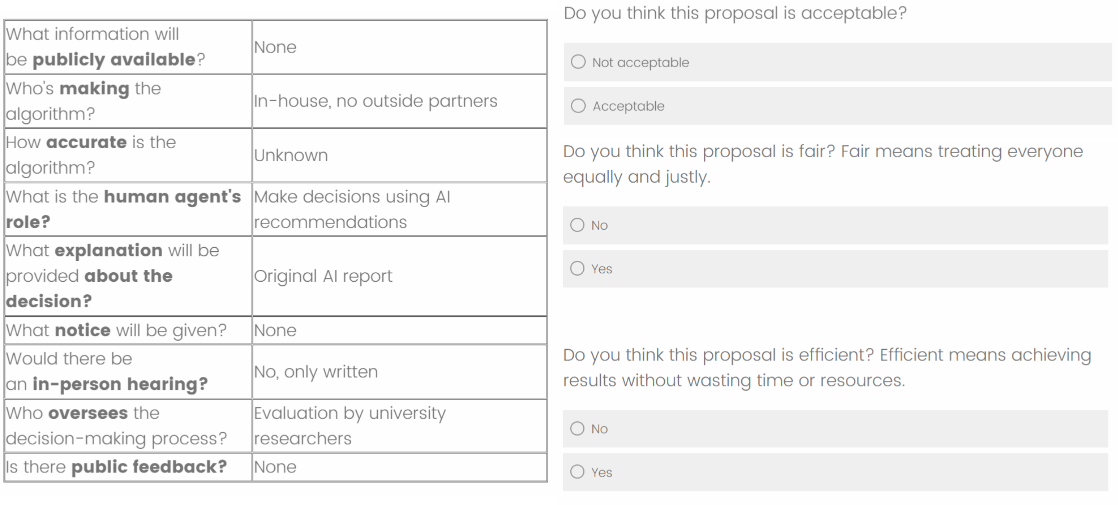
Participants were shown the hypothetical policy proposal presented in a table format. Each policy was presented as shown in Table 2 with a table that describes the attributes and levels under the policy description. The left side of the table indicated the category of attribute information, while the right side presented the particular value. Figure 1 shows an example of a proposal page. Each proposal had one of the values for each attribute randomly inserted with equal probability.

We used repeated measures, so each participant was asked on all five domains. However, the order of the policy proposals' domain was randomly assigned to address potential order effects, for instance, stemming from some domains being more acceptable (e.g., permitting) than others (e.g., refugee resettlement).

An additional randomization step was introduced for the attribute of accuracy. After piloting the design, a concern arose that accuracy was too salient and could be skewing results. Additionally, it is seldom the case that model accuracy is reported in such a succinct manner in public debate. Therefore, participants were assigned to a pure control

Figure 1

Proposal Table Example



What information will be **publicly available**?	None
Who's **making** the algorithm?	In-house, no outside partners
How **accurate** is the algorithm?	Unknown
What is the **human agent's role?**	Make decisions using AI recommendations
What **explanation** will be provided **about the decision?**	Original AI report
What **notice** will be given?	None
Would there be an **in-person hearing?**	No, only written
Who **oversees** the decision-making process?	Evaluation by university researchers
Is there **public feedback?**	None

Do you think this proposal is acceptable?

○ Not acceptable

○ Acceptable

Do you think this proposal is fair? Fair means treating everyone equally and justly.

○ No

○ Yes

Do you think this proposal is efficient? Efficient means achieving results without wasting time or resources.

○ No

○ Yes

group, where the table was presented without an accuracy attribute, and a treatment group, where accuracy was presented (with one of the three levels). This allowed us to ensure that the effects were robust and were not driven solely by the presence or absence of the accuracy feature.

We then asked participants for their support or opposition to the proposal, first on the main outcome question and then on secondary outcomes: fairness and efficiency. If they found the application acceptable, they explained their reasoning. If they found it unacceptable, they described their concerns. The study concluded with open-ended questions about factors influencing participants' views on AI acceptability and decision-making criteria, followed by a final debriefing message. The survey also included an attention check between scenarios, asking respondents to click "OK" "if you are a robot" leading to an error message, prompting them to pay attention before continuing.

Results

Overview of Analysis

A conjoint design allows one to experimentally identify the effects of different attributes on participants' preferences. It provides reliable measures of multidimensional preferences and estimates the causal effects of multiple attributes on hypothetical choices, as it allows researchers to measure any potential trade-offs between multifarious policy dimensions (Bansak et al., 2021). Such studies present participants with a policy "package" and elicit their responses.

Since multiple dimensions of algorithmic governance and procedural justice may affect how legitimate an application is perceived to be, an experimental design that focuses on a few treatment conditions cannot capture the whole picture. Thus, a conjoint design is more suitable for assessing perceptions of legitimacy. We use a single profile conjoint design for the sake of simplicity and to better simulate real policy proposals (as opposed to a choice between, for example, two candidates in an election run-off; Bansak et al., 2021).

We analyzed our conjoint experimental data using Average Marginal Component Effects (AMCEs) to estimate the causal effect of each attribute level on selection probability relative to the reference category (Hainmueller et al., 2014). AMCE represents the marginal effect of moving from the reference level to another level within the same attribute, averaged over the joint distribution of all other attributes. Thus, these are relative rather than absolute effects. We estimated these effects using logistic regression models with standard errors clustered at the respondent level to account for within-respondent correlations across multiple decisions.

Our primary specification takes the form:

$$Y_{ij} = \beta_0 + \beta_k X_{ijk} + \varepsilon_{ij} \tag{1}$$

where Y_{ij} represents the binary decision outcome for profile i evaluated by respondent j, and X_{ijk} represents the k attribute levels. To facilitate interpretation, we convert the logistic regression coefficients to changes in probability using the following approach: First, we estimate the baseline selection probability from an intercept-only model. Then, for each attribute level, we calculate the change in probability relative to this baseline using the logistic regression coefficients while accounting for the non-linear nature of the probability transformation. We report these effects in percentage points, along with cluster-robust standard errors transformed using the delta method to maintain appropriate uncertainty estimates in the probability scale. Our visualizations present the probability changes with corresponding 95% confidence intervals, enabling direct interpretation of the magnitude and precision of each attribute's effect on selection decisions.

Take a simple example where a decision maker must choose whether or not to hire a job candidate (yes=1, no=0). While the experimental profiles vary across multiple attributes, including age, sex, and other characteristics, we focus on interpreting the effect of hair color: black (reference level), grey, and blond. After running a logistic regression on the conjoint experimental data controlling for all other attributes, we obtain coefficients of -0.6 for grey hair and 0.3 for blond hair, with an intercept of 0.8. Converting from log odds to probabilities while holding all other attributes constant at their reference levels, we find that candidates with black hair (the reference level) have a 69% chance of being hired [exp(0.8)/(1 + exp(0.8))]. Relative to this baseline, grey-haired candidates face a 14 percentage point penalty, with only a 55% chance of being hired [exp(0.8 - 0.6)/(1 + exp(0.8 - 0.6))]. In contrast, blond-haired candidates enjoy a 6 percentage point advantage, with a 75% chance of being hired [exp(0.8 + 0.3)/(1 + exp(0.8 + 0.3))].

Descriptive Statistics

The sample contained about 51% female participants; the median age of the sample was 48 years. As per race and ethnicity, participants who identified as White comprised 77.2% of the sample, 6.4% Asian, 15.1% African-American, and 5.9% as Latino or Hispanic (participants could choose multiple categories). The use of weights in survey experiment analysis hinges on the researcher's intended generalization (external validity) (Egami and Hartman, 2022) and the ability to identify covariates that predict both treatment heterogeneity and selection into the sample (Miratrix et al., 2018). In this study, the design assumes treatment heterogeneity. There is no concern for systematic differences in the composition of units in the experimental sample and the target population (voting-age Americans); thus, treatment validity (T-validity, Egami and Hartman, 2022) is not compromised, which is the validity of interest in this study.

As per the extra randomization step, 51% of the sample was shown the attribute of accuracy, while 49% did not have the attribute included in the experiment table.

Table 3
Descriptive Statistics of Sample Demographics and Attitudes

Variable	**Value**	***N* (%)**
Demographics		
Age	48 [33, 60]	1213
Female		619 (51)
White		937 (77.2)
Black or African American		183 (15.1)
Asian		78 (6.4)
Hispanic or Latino Ethnicity		71 (5.9)
Bachelor's degree or higher		651 (53.7)
Liberal		475 (39.2)
Conservative		393 (32.4)
CS Knowledge and Experience		
No CS education		439 (36.2)
No CS experience		694 (57.2)
CS education: College-level or higher		518 (42.7)
CS experience: Some or a lot		519 (42.8)
AI Knowledge and Attitudes		
AI general knowledge	0.75 [0.64, 0.89]	1213
Used ChatGPT or similar		946 (78)
AI beneficial	0.62 [0.38, 0.75]	1213
Tech Attitudes		
AI regulated	0.75 [0.62, 0.88]	1213
Tech beneficial	0.55 [0.4, 0.7]	1213

Note: Values are presented as median [IQR] for continuous variables and *N* (%) for categorical variables. For the attitudes, the higher the score the more the participants agree with the category.

Table 4 shows that participants were more likely to indicate that using AI for public policy is fair if they believe that technological advancement and AI are beneficial, but less likely to think this is fair if they support regulating AI and if they have greater knowledge of AI (although this effect is small). Similarly, participants were more likely to imply that using AI in public policy is efficient if they believe technology and AI are beneficial. Again, participants who supported AI regulation were less likely to think that using AI for public policy is efficient. Unlike attitudes regarding fairness, which were somewhat negatively impacted by AI knowledge, attitudes regarding efficiency were negatively impacted by CS experience.

Table 4
Attitudes and Knowledge Relationship to Fairness and Efficiency Perceptions

	Efficiency	**Fairness**
Believes technology is beneficial	1.828	2.964
	p = .010	p <.001
Believes AI is beneficial	5.162	3.113
	p <.001	p <.001
Supports regulating AI	0.545	0.426
	p <.001	p <.001
Has CS knowledge	1.267	0.843
	p = 0.141	p = 0.264
Has CS experience	0.565	0.898
	p <.001	p = 0.414
Has AI knowledge	0.760	0.667
	p = .098	p = .012
Num. Obs.	4440	4440
AIC	5439.8	5826.9
BIC	5631.8	6018.9

Note: Results from logistic regression models predicting binary perceptions of algorithmic policy decisions as efficient or fair. From the full sample of 6,065 participants, analyses include 4,440 complete observations after list-wise deletion. Coefficients are presented as odds ratios, where values greater than 1 indicate increased odds of perceiving the algorithm as efficient/fair, and values less than 1 indicate decreased odds. For example, an odds ratio of 1.5 means 50% higher odds of a positive perception. Models control for demographic characteristics (sex, age, education), political views, and racial/ethnic identity. Standard errors are clustered at the respondent level.

We note that no participants failed the attention check, and none were disqualified on this or any other basis.

Perceptions of Acceptability

First, participants were sensitive to the policy domain. For example, programs proposing algorithmic parole decisions were markedly less acceptable—a marginal effect of

Table 5

Conjoint Experiment Results: Effects on Outcome Probability (pp)

Feature	Level	Effect (*pp*)	*SE* (*pp*)	*z*	*p*	CI Lower	CI Upper
Accuracy	not shown	0.0	–	–	–	–	–
Accuracy	75%	-8.3	2.1	-3.89	0.000	-12.5	-4.1
Accuracy	85%	2.6	2.2	1.15	0.252	-1.8	7.0
Accuracy	Unknown	-17.0	2.0	-8.55	0.000	-21.0	-13.1
Cost	$2m	0.0	–	–	–	–	–
Cost	$10m	0.2	1.7	0.13	0.894	-3.0	3.5
Cost	$5m	1.0	1.6	0.60	0.551	-2.2	4.1
Publicly available information	No_details	0.0	–	–	–	–	–
Publicly available information	A quick, simple overview	3.3	1.7	1.97	0.049	0.0	6.5
Publicly available information	Technical information	7.8	1.7	4.70	0.000	4.6	11.1
Algorithm development	In-house, no outside partners	0.0	–	–	–	–	–
Algorithm development	Non-profit organization	0.9	1.9	0.46	0.645	-2.8	4.6
Algorithm development	Partnership with academics	-0.1	2.0	-0.08	0.940	-4.0	3.7
Algorithm development	Purchased from a private company	-2.3	1.9	-1.18	0.239	-6.0	1.5
Explanation of the decision	No_explanation	0.0	–	–	–	–	–
Explanation of the decision	Detailed explanation of reasons and process	14.8	1.9	7.64	0.000	11.0	18.6
Explanation of the decision	Only the bottom line	9.3	1.9	4.86	0.000	5.6	13.1
Explanation of the decision	Original AI report	11.4	2.0	5.83	0.000	7.6	15.2
Public engagement	No_feedback	0.0	–	–	–	–	–
Public engagement	Only through political representatives	3.0	2.0	1.53	0.125	-0.8	6.9
Public engagement	Public town-hall meetings	8.7	1.9	4.49	0.000	4.9	12.6
Public engagement	Written public comments	7.1	1.9	3.71	0.000	3.4	10.9
Hearing	In-person hearing by request	0.0	–	–	–	–	–
Hearing	No, only legal appeal	-8.8	1.8	-4.86	0.000	-12.3	-5.2
Hearing	No, only written	-12.4	1.8	-7.04	0.000	-15.9	-9.0
Hearing	Video hearing by request	-2.7	1.9	-1.46	0.145	-6.4	1.0
Notice	No_notice	0.0	–	–	–	–	–
Notice	That personal information was accessed	-0.6	2.0	-0.33	0.744	-4.5	3.2
Notice	That there was an assessment by AI	0.4	1.9	0.22	0.827	-3.4	4.2
Notice	Option to consent to the AI assessment	4.7	1.9	2.40	0.016	0.9	8.5
Oversight	A quality review team reviews cases at random	0.0	–	–	–	–	–
Oversight	Evaluation by university researchers	0.5	2.2	0.23	0.820	-3.8	4.8
Oversight	Oversight board with public experts	1.6	2.2	0.76	0.448	-2.6	5.9
Oversight	The agency will publish public reports	2.0	2.1	0.97	0.333	-2.1	6.2
Oversight	State accountability office will publish reports	1.4	2.1	0.65	0.513	-2.8	5.6
Human agent role	Make decisions using AI recommendations	0.0	–	–	–	–	–
Human agent role	Review AI decisions	-0.7	1.9	-0.34	0.731	-4.4	3.1
Human agent role	The AI makes an immediate decision	-15.0	1.8	-8.53	0.000	-18.4	-11.6
Human agent role	Agent reconsiders the AI decision	0.3	2.0	0.14	0.890	-3.6	4.1
Policy domain	City	0.0	–	–	–	–	–
Policy domain	Parole	-24.9	1.4	-17.60	0.000	-27.7	-22.2
Policy domain	Permit	-2.3	1.8	-1.30	0.192	-5.7	1.1
Policy domain	Refugee	-12.2	1.7	-7.04	0.000	-15.6	-8.8
Policy domain	Unemployment	-9.3	1.7	-5.40	0.000	-12.7	-5.9

–24.9 percentage points, *SE*=1.4, *z*=–17.60, *p*<.001 relative to the reference category. In contrast, while the AMCE for building permit decisions was modest of –2.3 *pp*, *SE*=1.8, *p*=.192 (Figure 2), the marginal means analysis (Figure10) revealed that acceptability for city-based chatbot applications was significantly higher than the overall mean. In this analysis, unemployment and refugee-related decision-making did not differ significantly from the grand mean, p>.10, underscoring that the negative effects were concentrated in the parole context.

Accuracy is also vital for participants. Compared to the reference condition in which no accuracy information was provided, .0 *pp*), reporting a high accuracy of 85% had a negligible effect, 2.6 *pp*, *SE*=2.2, p=.252. In contrast, indicating a lower accuracy level (75%) reduced acceptability by 8.3 percentage points, *SE*=2.1, p<.001, and labeling accuracy as "Unknown" produced an even larger decrease, –17.0 *pp*, *SE*=2.0, p<.001.

Conversely, the *cost* of the government program was not a significant driver of acceptability. Neither the $10m, 0.2 *pp*, *SE*=1.7, p=.894, nor the $5m option, 1.0 *pp*, *SE*=1.6, p=.551, differed significantly from the baseline cost of $2m. In terms of transparency, however, the level of *publicly available information* mattered. Providing a brief overview increased acceptability by 3.3 percentage points, *SE*=1.7, p=.049, and sharing detailed technical information raised it even further, 7.8 *pp*, *SE*=1.7, p<.001.

Participants' judgments were largely unaffected by the identity of the *algorithm developer*. Whether the algorithm was developed in-house, by a non-profit, via an academic partnership, or purchased from a private company, none of the alternatives yielded statistically significant differences, p-values ranging from 0.239 to 0.940.

In contrast, the *explanation of the decision* emerged as a key attribute. Compared to providing no explanation at all, offering a detailed explanation increased acceptability by 14.8 percentage points, *SE*=1.9, p<.001, and even a brief "bottom line" explanation raised it by 9.3 *pp*, *SE*=1.9, p<.001. Similar patterns were observed for *public engagement*: allowing direct public input — via town-hall meetings, 8.7 *pp*, *SE*=1.9, p<.001, or written comments, 7.1 *pp*, *SE*=1.9, p<.001 — significantly boosted acceptability, whereas engagement only through political representatives did not, 3.0 *pp*, *SE*=2.0, p=.125. Regarding the option of a *hearing* in the decision process, participants were indifferent between in-person and video hearings. An in-person hearing (the reference condition) elicited neutral responses, but a "no hearing" option—where decisions could only be appealed legally or provided in writing—reduced acceptability by 8.8 pp, *SE*=1.8, *p*<.001, and 12.4 pp, *SE*=1.8, *p*<.001, respectively. Similarly, a *notice* that included an option to consent to the AI assessment raised acceptability by 4.7 pp, *SE*=1.9, *p*=.016, compared to providing no notice, whereas other notice formats did not yield significant effects.

Institutional design features related to *oversight* showed little differentiation. None of the oversight options (whether via a quality review team, university researchers, an oversight board, or state accountability offices) produced significant differences from the reference condition, with all *p*-values>.33.

Finally, having any *human agent involvement* in the process markedly increased acceptability. When decisions were made solely by the AI—with no human involvement—acceptability dropped by 15.0 pp, *SE*=1.8, *p*<.001. In contrast, conditions that involved human review or a reconsideration process did not differ significantly from each other, –0.7 *pp* and 0.3 *pp*, respectively, both with *p*-values>.73.

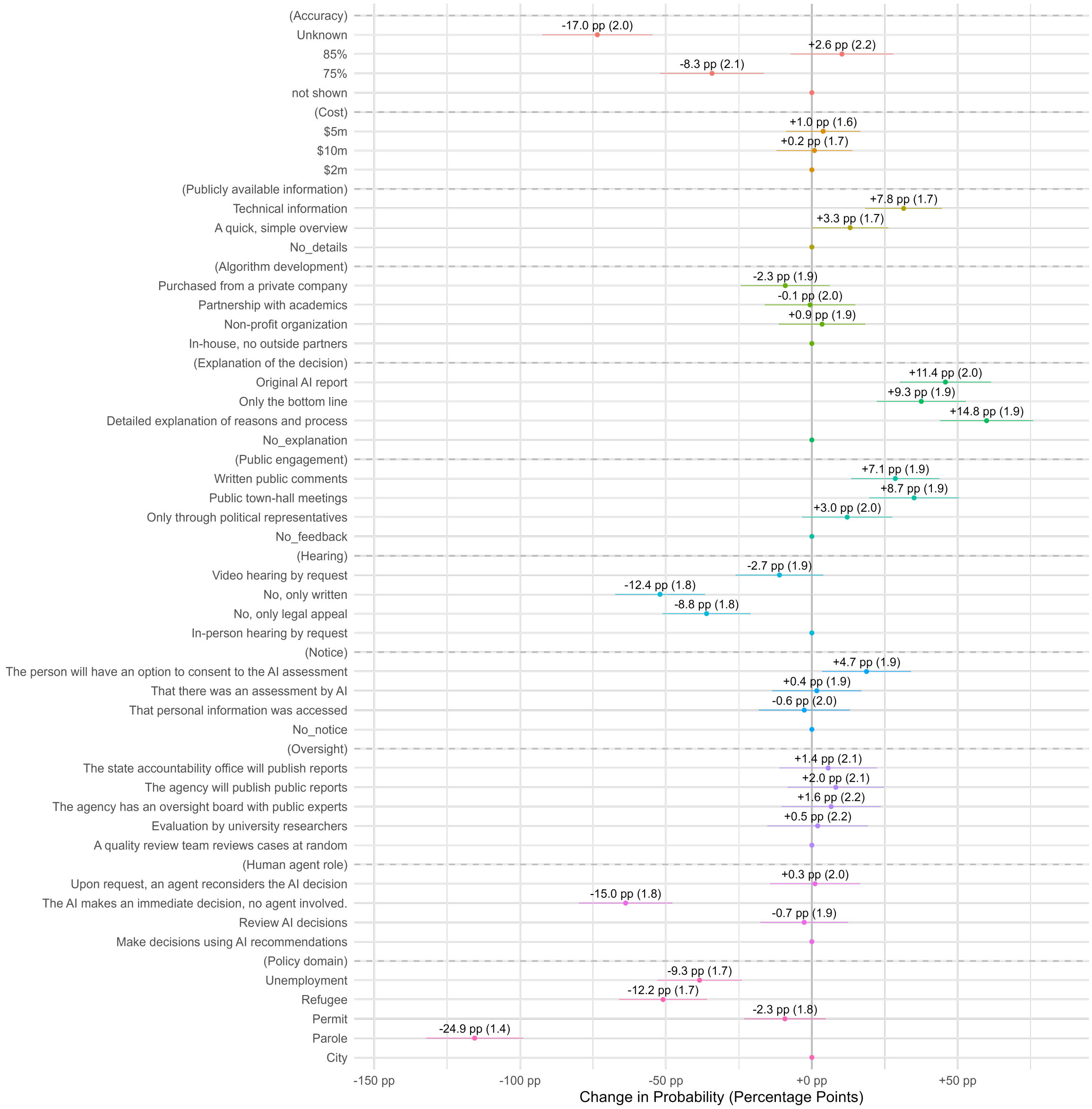
Table 6
Summary of Main Findings

	Attribute	**Findings**
Algorithmic Design	Accuracy	An unknown or low accuracy affects acceptability
	Cost	Unimportant
	Publicly available technical information	Preferred
	Whose Developing	Unimportant
Institutional Design	Post-decision Hearing	Important
	Notice and Consent	A preference for consent
	Explaining the AI's Decision	Important
	Human-in-the-loop	Any role is better than none
	Public engagement on AI use	Direct engagement preferred
	Oversight	Unimportant
Policy Domain	Release on Parole	Less acceptable
	Building and small business permits	More acceptable

We examine the *attribute importance* scores, a method that decomposes participants' choices into the relative contribution of each attribute level, in Figures,3 and 4. Figure,3, shows *Policy domain* standing out as the most influential driver of *Acceptability*, followed by *Accuracy*, *Human agent role*, and *Explanation of the decision*; attributes like *Cost*,

Figure 2

Average Marginal Effects on Probability of Acceptability



Note: This figure shows the Average Marginal Component Effects (AMCE) for the main outcome of acceptability, expressed as changes in probability. The analysis was conducted using a logistic regression model. The effects represent the change in the probability of acceptance when moving from the reference level to each attribute level, holding all other variables constant. The baseline probability of acceptance is derived from an intercept-only model. Points represent point estimates, and bars represent 95% confidence intervals. The numbers next to each point show the percentage point change in probability and the standard error (in percentage points) in parentheses.

Oversight, and *Notice* show comparatively more minor effects. In contrast, Figure,4 reveals that *Efficiency* judgments hinge overwhelmingly on *Accuracy*, which eclipses all other attributes, with *Explanation of the decision* ranking a distant second. Together, these findings indicate that while Acceptability depends strongly on contextual and procedural factors, Efficiency perceptions are driven primarily by the algorithm's perceived accuracy.

Discussion

Much of the extant literature is focused on critiquing the adoption of algorithmic tools in government or drawing direct comparisons between algorithmic and human decision-making (R. P. Kennedy et al., 2022; Waldman and Martin, 2022; Yalcin et al., 2023). The current study allows us to focus the discussion by grounding findings in specific attributes and suggesting an applied framework.

Policy domain. Notably, the single most important attribute was the policy domain (Figure 3). We present AMCE estimates by domain in Supplement A. However, these results should be interpreted with caution. Our design examines under which conditions an algorithm would be perceived as more or less acceptable within the context of an administrative agency. The hypothetical policies were used for mundane realism, as we could not test our hypotheses without concrete policy examples. As a result, from an external validity perspective, we cannot extrapolate to policies outside the pool of the five examples used, nor can we confidently account for policy variation. Nonetheless, with caution, we can infer that the public is more willing to accept algorithmic involvement in morally neutral domains, such as applications for building permits and small business licensing, than in morally loaded applications, such as refugee resettlement or release from prison (similar to Raviv, 2023; Waldman and Martin, 2022; Wenzelburger et al., 2024). In line with previous research on criminal justice (Simmons, 2018; A. J. Wang, 2018), parole decisions were the least acceptable among the domains studied, suggesting that people are more sensitive to areas that involve curtailing liberties. This is especially interesting because much of the innovation in algorithmic decision-making has occurred in criminal justice settings (Engel and Grgić-Hlača, 2021; Lima et al., 2021), such as pre-trial detention, sentencing, and release on parole – often justified as an instrument to ameliorate racial bias and other inconsistencies common in the present system. A possible alternative to this explanation is that the use of algorithms in criminal justice contexts has gained public attention and is thus much more salient than other domains. This does not explain, however, the variation between the other domains. Refugee resettlement and unemployment benefits turned out to be in the middle zone, perhaps as they are less related to the core of individual freedoms, yet are also less politically charged than parole. All of this suggests that policy domain and context matter when considering algorithmic decisions (Wenzelburger et al., 2024). Given the scholarly and regulatory interest in focusing on high-stakes applications, our findings indicate that there is a real difference in public opinion.

In general, we find that both algorithm design and institutional design attributes affect public perceptions of ADTs in government programs. Still, key attributes touted in the policy literature do not present significant effects. For instance, despite the emphasis on algorithmic transparency, development of the algorithm, and institution of oversight bodies, they all appear to have negligible impact on public perception.

Accuracy. Accuracy significantly influences perceptions of AI government decision-making, particularly regarding efficiency. This finding bolsters the face validity of the efficiency outcome compared to fairness and acceptability (in which accuracy had less relative importance than the policy domain). Notably, excluding Accuracy information yields substantial effects, possibly due to its specificity. People react negatively to low or no "accuracy" information, but do not significantly reward higher levels. Accuracy affects Acceptability (Figure 3), but comparing with its impact on efficiency reveals major differences (Figure 4), with nearly *three times* the relative weight of the second most important attribute. This suggests participants closely equate efficiency with accuracy (see the complete analysis of exploratory outcomes in the Online Supplemental Material).

Contrary to many critics, our results show that people care less about governance measures such as transparency and oversight bodies than issues of the administrative decision-making and its procedural components, which they associate with more acceptable government actions. The results suggest that the acceptability of a decision in the public's eyes hinges on effective *human communication* during the decision-making process, which goes beyond the observation that people perceive a decision to more fair when a human is "in the loop" (Hermstrüwer and Langenbach, 2022; R. P. Kennedy et al., 2022).

Decision explanation. Our study also shifts the focus from explaining the algorithmic output to elucidating the decision. Moreover, in the novel context of human-AI interactions, we find that notice plays a crucial role. Our data suggest that the awareness of, and even consent to, algorithmic assessments is a strong factor in improving public perceptions. Previous research has shown that procedural rights (such as hearings and reason-giving) enhance perceived acceptability, regardless of whether the decision-makers are human or algorithmic (Grimmelikhuijsen, 2022; Henning and Langenbach, 2024). We propose that the perceived value of human involvement is closely tied to the inherent need for human communication, a prominent aspect of the process-based procedural justice approach (Tyler and Lind, 2001).

Importantly, our findings show that the importance of procedural guarantees goes beyond the skepticism of automation: even the subset of participants most reserved toward AI and the role of technology in society were more likely to accept policy proposals based on the aforementioned attributes. Essentially, we suggest that people tend to be more accepting of algorithmic decision-making when they feel a greater sense of control and individuation over the process or an opportunity to influence the process.

Figure 3

Attribute Importance Weights of Acceptability Outcome

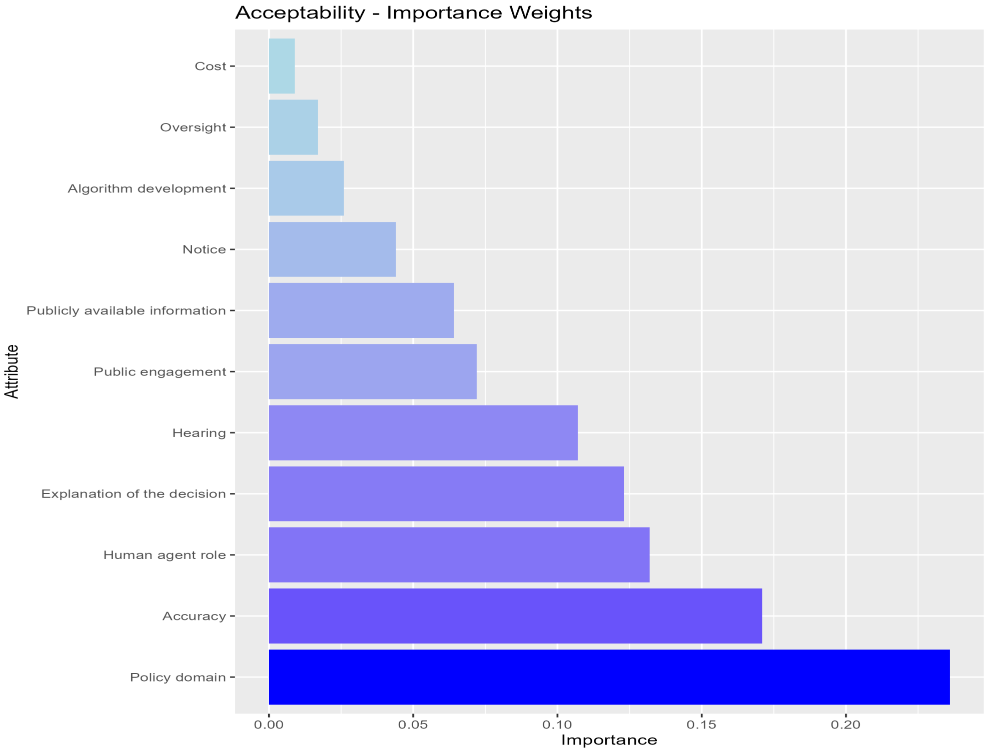
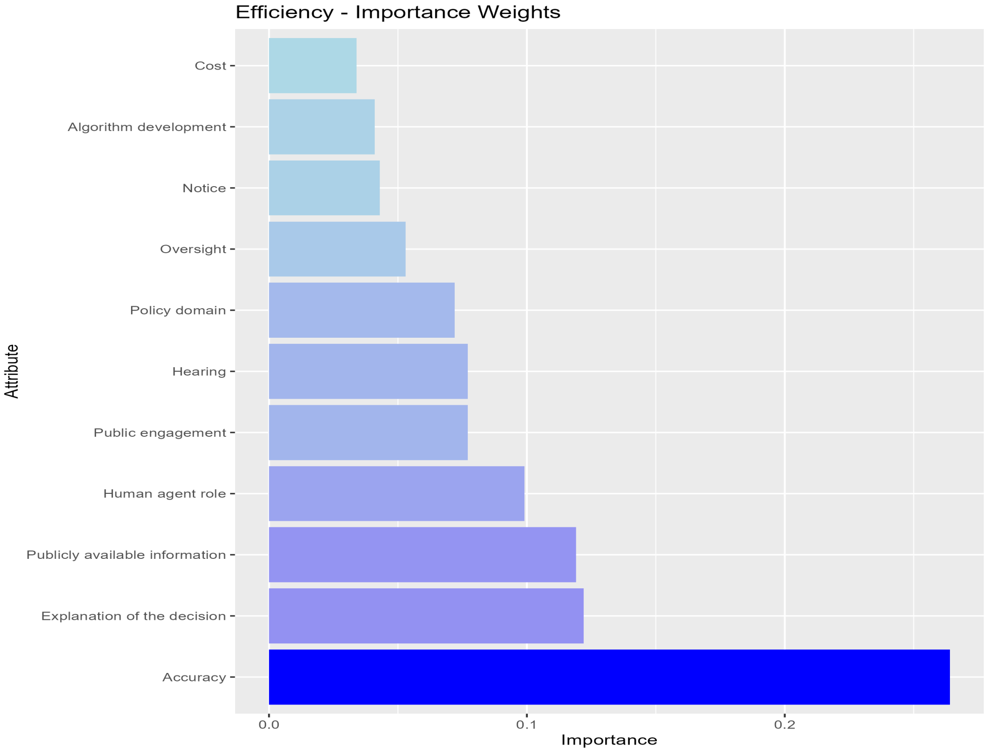


Figure 4

Attribute Importance Weights of Efficiency Outcome



Note: Importance Weights were calculated using the `radiant.multivariate` package in R. The weights represent the relative importance of each attribute in determining the respective outcomes.

Limitations and Future Directions

Limitations

While this study offers valuable insights into the perceived acceptance of government use of algorithms, several limitations should be noted. Despite the benefits of using a single-profile conjoint design—such as the ability to manipulate and test multiple factors within a controlled setting systematically—participants were exposed to different scenario levels across multiple hypothetical vignettes, potentially allowing them to compare these levels as they formed their judgments. Consequently, participants' responses may reflect a comparative process influenced by the variety of profiles presented in the study, rather than the more isolated, context-specific reasoning that would occur in actual practice.

Our research relied on hypothetical scenarios and policy domains, potentially limiting the generalizability of findings to real-world implementations of AI systems. Our experimental design may not capture the complexity of actual administrative processes and the nuances of human-AI interactions fully. Additionally, our focus on specific policy domains, while allowing for contextual analysis, may not represent the full spectrum of government services where algorithmic decision-making could be applied.

Our study primarily examined perceptions without directly measuring actual outcomes or long-term impacts of algorithmic decision-making in government. This focus on perceptions, while valuable, may not fully reflect the practical implications of implementing such systems.

From a methodological standpoint, our study used a binary choice measure for the outcomes. Despite several advantages associated with this design, including a clear, bias-minimized measure of participants' attitudes (Knudsen and Johannesson, 2019), it also has disadvantages. Primarily, a binary measure, as opposed to Likert scales, forces respondents to choose, and leads to a loss of potential variation in responses.

Additionally, several of the attributes we tested produced statistically insignificant results. Nevertheless, a null effect in an experiment does not imply that an attribute can have no influence, theoretically; rather, it indicates that when averaging over the distribution of all other attributes, the tested levels did not produce significant differences compared to each other. In our case, for example, when cost was found to be non-significant, this means that none of the cost levels we tested showed significant differences from one another when averaging over the distribution of all other attributes. Unlike other attributes with clear categorical distinctions or well-defined upper and lower bounds (e.g., accuracy), cost does not lend itself to such constraints.

Finally, this study was conducted within the American context, which represents both a limitation and a strength. The focus on the U.S. administrative state and its legitimacy challenges provides depth and specificity to our findings. However, it also potentially limits the direct applicability of our results to other national contexts with different regulatory environments, administrative structures, and public attitudes toward technology and governance. Moreover, our sample was 77% White, slightly higher than the US Census Bureau's 2022 estimate of 72.5% Americans identifying as White. This occurred because we utilized Prolific's representative sampling option, prioritizing stratification by political affiliation over racial/ethnic representation. This slight demographic overrepresentation likely has minimal impact on our findings, as our analyses control for demographic factors and our research primarily focuses on within-subject evaluations of policy attributes rather than between-group comparisons.

Future Directions

Future research should aim to address these limitations and expand upon our findings. One key area for investigation is the variability of administrative agencies and programs. Our study highlights the contextual nature of perceptions, suggesting that the attributes that matter most for public safety or rehabilitation may differ from those that matter for agencies regulating business. Further work could unpack these differences across a broader range of administrative domains.

Researchers should also consider conducting longitudinal studies to examine the relationship between perceptions and real-world outcomes of algorithmic decision-making in government. This could provide valuable insights into such systems' long-term impacts and effectiveness.

While our study utilized American institutions and scholars, the core themes and questions raised are globally relevant (*cf.* Wenzelburger et al., 2024). Future research could extend this analytical approach to similar democratic systems in other countries, allowing for comparative analyses. This would help understand how cultural, social, and political contexts influence perceptions of algorithms in government decision-making. Lastly, future studies might explore the potential gap between public perceptions of algorithmic systems and their actual performance or impact. This could involve parallel studies of public attitudes and system outcomes, providing a more comprehensive understanding of the role of algorithmic decision-making in public administration.

Conclusion

This study examines public perceptions of the use of algorithms in decision-making in government programs. Our framework suggests that explaining decisions made using an algorithm, giving appropriate notice, a hearing option, and maintaining the supervision of a human agent are key components for public support when algorithmic systems are being implemented. The study highlights trade-offs in what the public seeks, and demonstrates that acceptability is highly dependent on a policy context, with applications in domains that concern more basic liberties and interests deemed less acceptable. While policy context was most salient with regards to general acceptability and ratings of overall fairness, the accuracy of the algorithm mattered most for efficiency perceptions.

This study adds to a burgeoning interest in public attitudes towards algorithms and artificial intelligence in government and adds granularity to the debate. Our results show that governments may be in a bind: on the one hand, they can institute safeguards that improve the public's reactions, regardless of concrete outcomes in individual cases. On the other hand, safeguards are costly and may undermine the gains in expediency, accuracy, and resources that incentivize the agency to adopt ADTs in the first place. The social anxiety that exists around automation and algorithms may prod agencies to consider public perceptions carefully. Agencies thus have much to learn from the trade-offs in perceptions of ADTs and their change across policy domains, and unpacking the attributes can support their design. This could improve the perceptions of these programs in the public's eyes, imbuing administrative agencies with support that they often lack, and reduce some of the anxiety towards automation.

References

Aoki, N. (2020). An experimental study of public trust in ai chatbots in the public sector. *Government Information Quarterly*, *37*(4), 101490. https://doi.org/10.1016/j.giq.2020.101490

Arnesen, S., Broderstad, T. S., Fishkin, J., Johannesson, M. P., & Siu, A. (2024, February 19). Knowledge and support for AI in the public sector: A deliberative poll experiment. https://doi.org/10.2139/ssrn.4731109

Associated Press. (2023). *New york city chatbot spreads misinformation* [Accessed: 2024-07-29]. https://apnews.com/article/new-york-city-chatbot-misinformation-6ebc71db5b770b9969c906a7ee4fae21

Bambauer, D. E., & Risch, M. (2021). Worse than human? *Arizona State Law Journal*, *53*(4), 1091–1152. Retrieved November 13, 2022, from https://heinonline.org/HOL/P?h=hein.journals/arzjl53&i=1125

Bansak, K., Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2021, April). Conjoint survey experiments [Google-Books-ID: mQEbEAAAQBAJ]. In J. N. Druckman & D. P. Green (Eds.), *Advances in experimental political science*. Cambridge University Press.

Bansak, K., & Paulson, E. (2023). Public attitudes on performance for algorithmic and human decision-makers [Preprint on OSF]. https://doi.org/10.31219/osf.io/pghmx

Beetham, D. (2013). *The legitimation of power* (2nd) [Originally published in 1991]. Macmillan International Higher Education.

Bitektine, A. (2011). Toward a theory of social judgments of organizations: The case of legitimacy, reputation, and status [29 pages]. *The Academy of Management Review*, *36*(1), 151–179. https://www.jstor.org/stable/29765010

Busuioc, M. (2021). Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review*, *81*(5), 825–836. https://doi.org/10.1111/puar.13293

Cave, S., Coughlan, K., & Dihal, K. (2019). Scary robots: Examining public responses to ai. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 331–337. https://doi.org/10.1145/3306618.3314232

Chatterjee, S., Khorana, S., & Kizgin, H. (2022). Harnessing the potential of artificial intelligence to foster citizens' satisfaction: An empirical study on india [https://doi.org/10.1016/j.giq.2021.101621]. *Government Information Quarterly*, *39*, 101621. https://doi.org/10.1016/j.giq.2021.101621

Chen, B. M., Stremitzer, A., & Tobia, K. (2022). Having your day in robot court. *Harvard Journal of Law & Technology*, *36*(1), 127–169. Retrieved May 12, 2023, from https://heinonline.org/HOL/P?h=hein.journals/hjlt36&i=132

Coglianese, C., & Lehr, D. (2019). Transparency and algorithmic governance. *Administrative law review*, *71*(1), 1–56.

Coppock, A. (2023). *Persuasion in parallel: How information changes minds about politics*. University of Chicago Press.

Dacin, M. T., Oliver, C., & Roy, J.-P. (2007). The legitimacy of strategic alliances: An institutional perspective [19 pages]. *Strategic Management Journal*, *28*(2), 169–187. https://www.jstor.org/stable/20142341

Deephouse, D. L., & Suchman, M. (2008). Legitimacy in organizational institutionalism. In R. Greenwood, C. Oliver, K. Sahlin, & R. Suddaby (Eds.), *The sage handbook of organizational institutionalism* (pp. 49–77). SAGE Publications. https://doi.org/10.4135/9781849200387.n2

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Díez-Martín, F., Blanco-González, A., & Díez-de-Castro, E. (2021). Measuring a scientifically multifaceted concept: The jungle of organizational legitimacy.

European Research on Management and Business Economics, *27*(1), 100131. https://doi.org/10.1016/j.iedeen.2020.100131

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona. *Plos one*, *18*(3), e0279720.

Egami, N., & Hartman, E. (2022). Elements of external validity: Framework, design, and analysis. *American Political Science Review*, 1–19.

Engel, C., & Grgić-Hlača, N. (2021). Machine advice with a warning about machine limitations: Experimentally testing the solution mandated by the wisconsin supreme court. *Journal of Legal Analysis*, *13*(1), 284–340. https://doi.org/10.1093/jla/laab001

Engstrom, D. F., & Haim, A. (2023). Regulating government ai and the challenge of sociotechnical design. *Annual Review of Law and Social Science*, *19*(1), 277. https://doi.org/10.1146/annurev-lawsocsci-120522-091626

Engstrom, D. F., & Ho, D. E. (2020). Algorithmic accountability in the administrative state. *Yale Journal on Regulation*, *37*, 800.

Feinstein, B. D. (2024). Legitimizing agencies. *University of Chicago Law Review*, *91*, 919–1019.

Gaozhao, D., Wright II, J. E., & Gainey, M. K. (2024). Bureaucrat or artificial intelligence: People's preferences and perceptions of government service. *Public Management Review*, *26*(6), 1498–1525. https://doi.org/10.1080/14719037.2022.2160488

Gibson, J. L., Caldeira, G. A., & Spence, L. K. (2005). Why do people accept public policies they oppose? testing legitimacy theory with a survey-based experiment [15 pages]. *Political Research Quarterly*, *58*(2), 187–201. https://www.jstor.org/stable/3595616

Gibson, J. L., Lodge, M., & Woodson, B. (2014). Losing, but accepting: Legitimacy, positivity theory, and the symbols of judicial authority. *Law & Society Review*, *48*(4), 837–866. https://doi.org/10.1111/lasr.12104

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *The Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Grimmelikhuijsen, S. (2022). Explaining why the computer says no: Algorithmic transparency affects the perceived trustworthiness of automated decision-making [Online first]. *Public Administration Review*. https://doi.org/10.1111/puar.13483

Grimmelikhuijsen, S. (2023). Explaining why the computer says no: Algorithmic transparency affects the perceived trustworthiness of automated decision-making. *Public Administration Review*, *83*(1), 241–253.

Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments [Publisher: Cambridge University Press]. *Political Analysis*, *22*(1), 1–30. https://doi.org/10.1093/pan/mpt024

Havasy, C. (2023). Relational fairness in the administrative state. *Virginia Law Review*, *109*. https://doi.org/10.2139/ssrn.4164125

Henning, A., & Langenbach, P. (2024, May 1). Bridging the human-automation fairness gap: How providing reasons enhances the perceived fairness of public decision-making. https://doi.org/10.2139/ssrn.4819145

Hermstrüwer, Y., & Langenbach, P. (2022). *Fair governance with humans and machines*. Retrieved May 12, 2023, from https://www.ssrn.com/abstract=4118650

Hibbing, J. R., & Theiss-Morse, E. (2001). Process preferences and american politics: What the people want government to be [9 pages]. *The American Political Science Review*, *95*(1), 145–153. https://www.jstor.org/stable/3117637

Hutchison, M. L., & Johnson, K. (2011). Capacity to trust? institutional capacity, conflict, and political trust in africa, 2000–2005. *Journal of Peace Research*, *48*, 737–752. https://doi.org/10.1177/0022343311417981

Ingrams, A., Kaufmann, W., & Jacobs, D. (2022). In ai we trust? citizen perceptions of ai in government decision making [https://doi.org/10.1002/poi3.276]. *Policy & Internet*, *14*, 390–409. https://doi.org/10.1002/poi3.276

Jackson, J. (2018). Norms, normativity, and the legitimacy of justice institutions: International perspectives. *Annual Review of Law and Social Science*, *14*(1), 145–165. https://doi.org/10.1146/annurev-lawsocsci-110316-113734

Johnson, D., Maguire, E. R., & Kuhns, J. B. (2014). Public perceptions of the legitimacy of the law and legal authorities: Evidence from the caribbean [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/lasr.12102]. *Law & Society Review*, *48*(4), 947–978. https://doi.org/10.1111/lasr.12102

Joshi, D. (2021). *Algorithmic accountability for the public sector*. AI NOW Institute.

Kaminski, M. E. (2019). The right to explanation, explained. *Berkeley Technology Law Journal*, *34*, 189.

Kaminski, M. E., & Malgieri, G. (2020). Algorithmic impact assessments under the gdpr: Producing multi-layered explanations [Online First]. *International Data Privacy Law*, 1–20. https://doi.org/10.2139/ssrn.3456224

Katzenbach, C., & Ulbricht, L. (2019). Algorithmic governance. *Internet Policy Review*, *8*(4), 1–18.

Kennedy, R. et al. (2024). Net versus relative impacts in public policy automation: A conjoint analysis of attitudes of black americans [https://link.springer.com/10.1007/s00146-024-01975-3 (last visited Dec 4, 2024)]. *AI & Society*. https://doi.org/10.1007/s00146-024-01975-3

Kennedy, R. P., Waggoner, P. D., & Ward, M. M. (2022). Trust in public policy algorithms. *The Journal of Politics*, *84*(2), 1132–1148. https://doi.org/10.1086/716283

Keppeler, F. (2024). No thanks, dear ai! understanding the effects of disclosure and deployment of artificial intelligence in public sector recruitment. *Journal of Public Administration Research and Theory*, *34*(1), 39–52. https://doi.org/10.1093/jopart/muad009

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv:1609.05807 [cs, stat]*. Retrieved July 29, 2020, from http://arxiv.org/abs/1609.05807

Knudsen, E., & Johannesson, M. P. (2019). Beyond the limits of survey experiments: How conjoint designs advance causal inference in political communication research. *Political Communication*, *36*(2), 259–271. https://doi.org/10.1080/10584609.2018.1493009

Kolkman, D. (2020). The usefulness of algorithmic models in policy making. *Government Information Quarterly*, *37*(3), 101488. https://doi.org/10.1016/j.giq.2020.101488

König, P. D. et al. (2024). The importance of effectiveness versus transparency and stakeholder involvement in citizens' perception of public sector algorithms. *Public Management Review*, *26*(7), 1061–1083.

König, P. D., Wurster, S., & Siewert, M. B. (2022). Consumers are willing to pay a price for explainable, but not for green ai. evidence from a choice-based conjoint analysis. *Big Data & Society*, *9*(1). https://doi.org/10.1177/20539517211069632

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, *165*, 74.

Lee, M. K., Jain, A., Cha, H. J., Ojha, S., & Kusbit, D. (2019). Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation [Publisher: Association for Computing Machinery]. *Proceedings of the ACM on Human-Computer Interaction*, *3*. https://doi.org/10.1145/3359284

Levi, M., & Sacks, A. (2009). Legitimating beliefs: Sources and indicators. *Regulation & Governance*, *3*, 311–333. https://doi.org/10.1111/j.1748-5991.2009.01058.x

Levi, M., Sacks, A., & Tyler, T. (2009). Conceptualizing legitimacy, measuring legitimating beliefs. *American Behavioral Scientist*, *53*, 354–375. https://doi.org/10.1177/0002764209338797

Levy, K., Chasalow, K. E., & Riley, S. (2021). Algorithms and decision-making in the public sector [_eprint: https://doi.org/10.1146/annurev-lawsocsci-041221-023808]. *Annual Review of Law and Social Science*, *17*(1), 309–334. https://doi.org/10.1146/annurev-lawsocsci-041221-023808

Lima, G., Grgić-Hlača, N., & Cha, M. (2021). Human perceptions on moral responsibility of ai: A case study in ai-assisted bail decision-making. *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3411764.3445260

Margalit, Y., & Raviv, S. (2023, September 15). The politics of using AI in policy implementation: Evidence from a field experiment. https://doi.org/10.2139/ssrn.4573250

Mashaw, J. L. (2018). *Reasoned administration and democratic legitimacy: How administrative law supports democratic government* [Google-Books-ID: ff5tDwAAQBAJ]. Cambridge University Press.

Medaglia, R., Gil-Garcia, J. R., & Pardo, T. A. (2023). Artificial intelligence in government: Taking stock and moving forward. *Social Science Computer Review*, *41*(1), 123–145.

Meier, K. J., & Bohte, J. (2007). *Politics and the bureaucracy: Policymaking in the fourth branch of government*. Thomson/Wadsworth.

Miller, S. M., & Keiser, L. R. (2021). Representative bureaucracy and attitudes toward automated decision making [https://doi.org/10.1093/jopart/muaa019]. *Journal of Public Administration Research and Theory*, *31*(1), 150–165. https://doi.org/10.1093/jopart/muaa019

Miller, S. M., Song, M., & Keiser, L. R. (2022). The effect of human versus automated interaction on willingness to participate in government programs: The role of representation [Online first, https://doi.org/10.1111/padm.12879]. *Public Administration*, 1–18. https://doi.org/10.1111/padm.12879

Miratrix, L. W., Sekhon, J. S., Theodoridis, A. G., & Campos, L. F. (2018). Worth weighting? how to think about and use weights in survey experiments. *Political Analysis*, *26*(3), 275–291.

Mulligan, D. K., & Bamberger, K. A. (2019). Procurement as policy: Administrative process for machine learning. *Berkeley Technology Law Journal*, *34*, 781. https://doi.org/10.2139/ssrn.3464203

Oetzel, M.-C., & Spiekermann, S. (2014). A systematic methodology for privacy impact assessments: A design science approach. *European Journal of Information Systems*, *23*(2), 126–150. https://doi.org/10.1057/ejis.2013.18

O'Shaughnessy, M., Schiff, D. S., Varshney, L. R., Rozell, C., & Davenport, M. (2021, December 14). *What governs attitudes toward artificial intelligence adoption and governance?* (preprint). Open Science Framework. https://doi.org/10.31219/osf.io/pkeb8

Pasquale, F. (2015). *The black box society.*

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 1.

Pew. (2023). Pew American Trends Panel Poll, 2023 [Dataset] [Roper #31120386, Version 1. Ipsos [producer]. Cornell University, Ithaca, NY: Roper Center for Public Opinion Research [distributor]. Access Date: Dec-29-2023].

Rahman, K. S. (2017). Reconstructing the administrative state in an era of economic and democratic crisis book review. *Harvard Law Review*, *131*(6), 1671–1713. Retrieved October 11, 2022, from https://heinonline.org/HOL/P?h=hein.journals/hlr131&i=1699

Ranchordas, S. (2021). Empathy in the digital administrative state. *Duke Law Journal*, (72).

Raviv, S. (2023, January 18). When do citizens resist the use of AI algorithms in public policy? theory and evidence. https://doi.org/10.2139/ssrn.4328400

Risse, T., & Stollenwerk, E. (2018). Legitimacy in areas of limited statehood. *Annual Review of Political Science*, *21*(1), 403–418. https://doi.org/10.1146/annurev-polisci-041916-020736

Sætra, H. S. (2020). A shallow defence of a technocracy of artificial intelligence: Examining the political harms of algorithmic governance in the domain of government. *Technology in Society*, *62*, 101283. https://doi.org/10.1016/j.techsoc.2020.101283

Schiff, D. S., Schiff, K. J., & Pierson, P. (2022). Assessing public value failure in government adoption of artificial intelligence. *Public Administration*, *100*(3), 653–667.

Schiff, K. J. et al. (2024). Institutional factors driving citizen perceptions of ai in government: Evidence from a survey experiment on policing [https://onlinelibrary.wiley.com/doi/abs/10.1111/puar.13754 (last visited Dec 4, 2024)]. *Public Administration Review*.

Schoon, E. W. (2022). Operationalizing legitimacy. *American Sociological Review*, *87*(3), 478–503. https://doi.org/10.1177/00031224221081379

Shin, D. (2020). How do users interact with algorithm recommender systems? the interaction of users, algorithms, and performance. *Computers in Human Behavior*, *109*, 106344. https://doi.org/10.1016/j.chb.2020.106344

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai [Online first]. *International Journal of Human-Computer Studies*, *146*, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Simmons, R. (2018). Big data and procedural justice: Legitimizing algorithms in the criminal justice system. *Ohio State Journal of Criminal Law*, *15*.

Smith, A. (2019). *More than half of u.s. adults trust law enforcement to use facial recognition responsibly* (tech. rep.). Pew Research Center. https://perma.cc/FUV7-5BDJ

Stagnaro, M. N., Druckman, J., Berinsky, A. J., Arechar, A. A., Willer, R., & Rand, D. G. (2024). Representativeness versus response quality: Assessing nine opt-in online survey samples. https://doi.org/10.31234/osf.io/h9j2d

Starke, C., & Lünich, M. (2020). Artificial intelligence for political decision-making in the european union: Effects on citizens' perceptions of input, throughput, and output legitimacy [Publisher: Cambridge University Press]. *Data & Policy*, *2*, e16. https://doi.org/10.1017/dap.2020.19

Stiglitz, E. H. (2022). *The reasoning state*. Cambridge University Press. https://doi.org/10.1017/9781108662673

Suddaby, R., Bitektine, A., & Haack, P. (2017). Legitimacy. *Academy of Management Annals*, *11*(1), 451–478. https://doi.org/10.5465/annals.2015.0101

Tyler, T. R. (2006). Psychological perspectives on legitimacy and legitimation. *Annu. Rev. Psychol*, *57*, 375–400.

Tyler, T. R., & Lind, A. E. (2001). Procedural justice. In J. Sanders & V. L. Hamilton (Eds.), *Handbook of justice research in law* (pp. 65–92). Springer US. https://doi.org/10.1007/0-306-47379-8_3

van der Voort, H. G., Klievink, A. J., Arnaboldi, M., & Meijer, A. J. (2019). Rationality and politics of algorithms. will the promise of big data survive the dynamics of public decision making? *Government Information Quarterly*, *36*(1), 27–38. https://doi.org/10.1016/j.giq.2018.10.011

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable ai for robotics. *Science Robotics*, *2*(6), eaan6080. https://doi.org/10.1126/scirobotics.aan6080

Waggoner, P. D., Kennedy, R., Le, H., & Shiran, M. (2019). Big data and trust in public policy automation. *Statistics, Politics and Policy*, *10*(2), 115–136. https://doi.org/10.1515/spp-2019-0005

Waldman, A., & Martin, K. (2022). Governing algorithmic decisions: The role of decision importance and governance on perceived legitimacy of algorithmic decisions [Publisher: SAGE Publications Ltd]. *Big Data & Society*, *9*(1), 20539517221100449. https://doi.org/10.1177/20539517221100449

Wang, A. J. (2018). Procedural justice and risk-assessment algorithms. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3170136

Wang, G., Guo, Y., Zhang, W., Xie, S., & Chen, Q. (2023). What type of algorithm is perceived as fairer and more acceptable? a comparative analysis of rule-driven versus data-driven algorithmic decision-making in public affairs. *Government Information Quarterly*, *40*(2), 101803. https://doi.org/10.1016/j.giq.2023.101803

Wenzelburger, G. et al. (2024). Algorithms in the public sector: Why context matters. *Public Administration*, *102*(1), 40–55.

Willems, J., Schmid, M. J., Vanderelst, D., Vogel, D., & Ebinger, F. (2022). Ai-driven public services and the privacy paradox: Do citizens really care about their privacy? [Online first]. *Public Management Review*, 1–19. https://doi.org/10.1080/14719037.2022.2063934

Yalcin, G., Themeli, E., Stamhuis, E., Philipsen, S., & Puntoni, S. (2023). Perceptions of justice by algorithms. *Artificial Intelligence and Law*, *31*(2), 269–292. https://doi.org/10.1007/s10506-022-09312-z

Zhang, B., & Dafoe, A. (2020). U.s. public opinion on the governance of artificial intelligence. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 187–193. https://doi.org/10.1145/3375627.3375827

Online Supplemental Material

This document contains supplemental materials for the study. Below is a brief Table of Contents:

Supplement A: Subgroup Analysis by Policy Domain

This supplement contains AMCE estimates for each policy domain. While the general trends and directions of the effects are mostly consistent across domains, several notable differences emerge. Notably, accuracy matters more in the "morally neutral" policy domains ("permit" and "city"); the Public engagement attribute only affects the "parole" policy domain; and the explanation of the decision matters most for the "unemployment" policy.

Figure 5

AMCE Estimates by Policy Domain: Parole

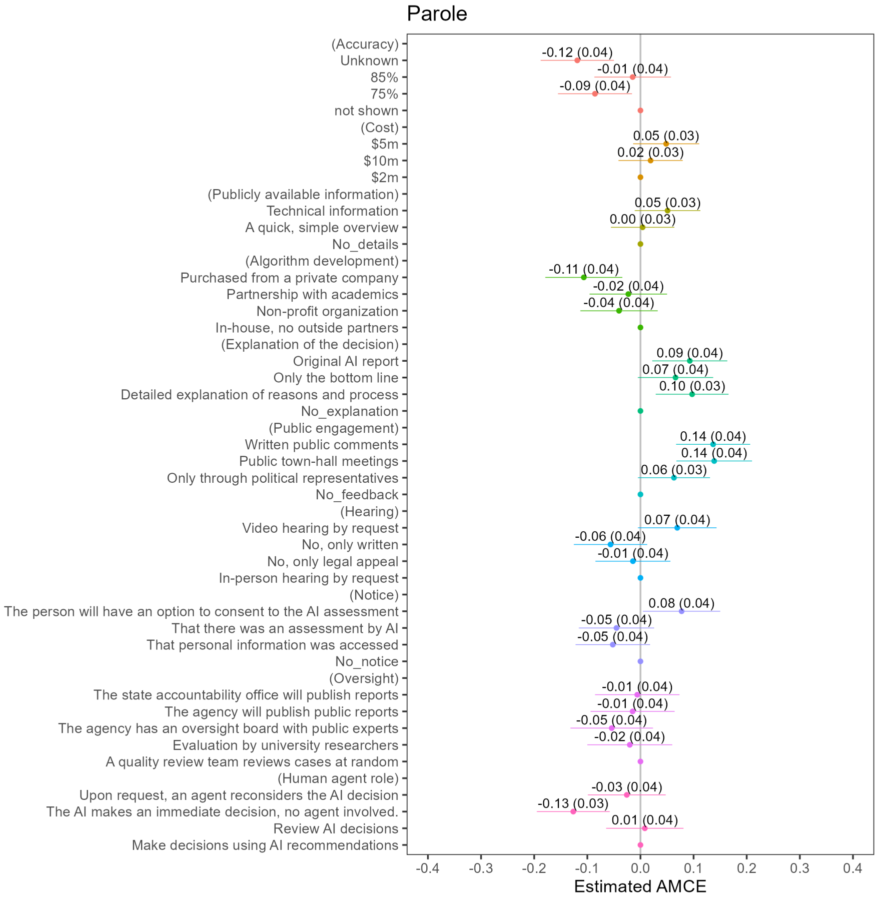


Figure 6

AMCE Estimates by Policy Domain: Unemployment

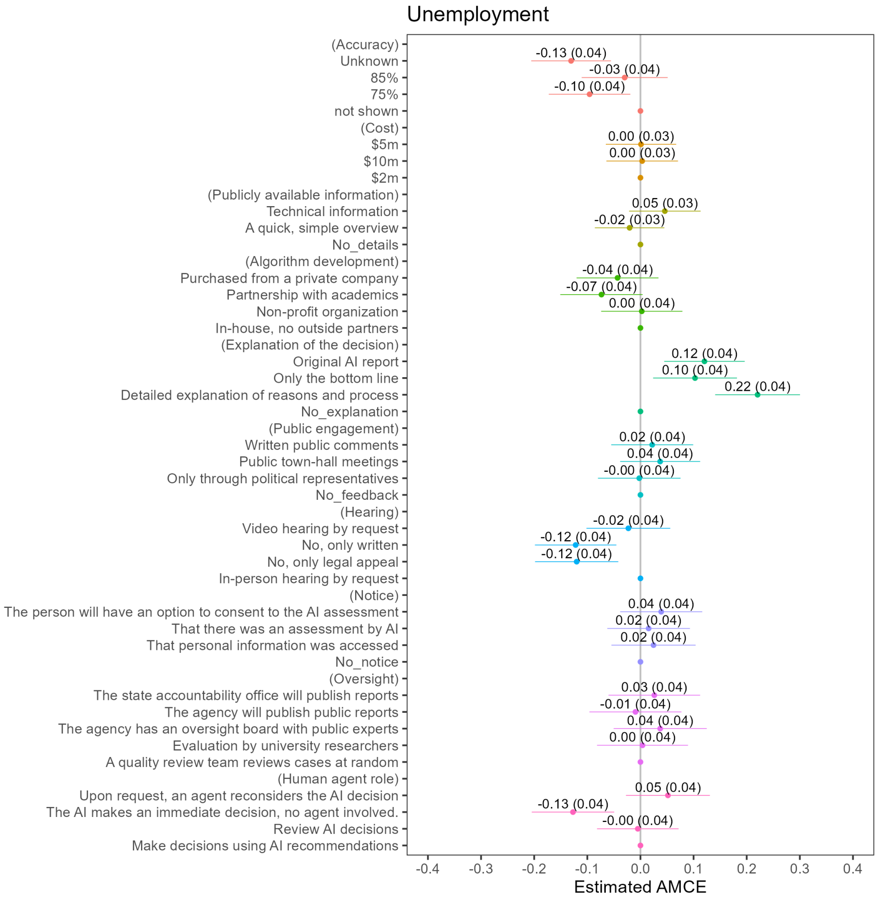


Figure 7

AMCE Estimates by Policy Domain: Refugee Resettlement

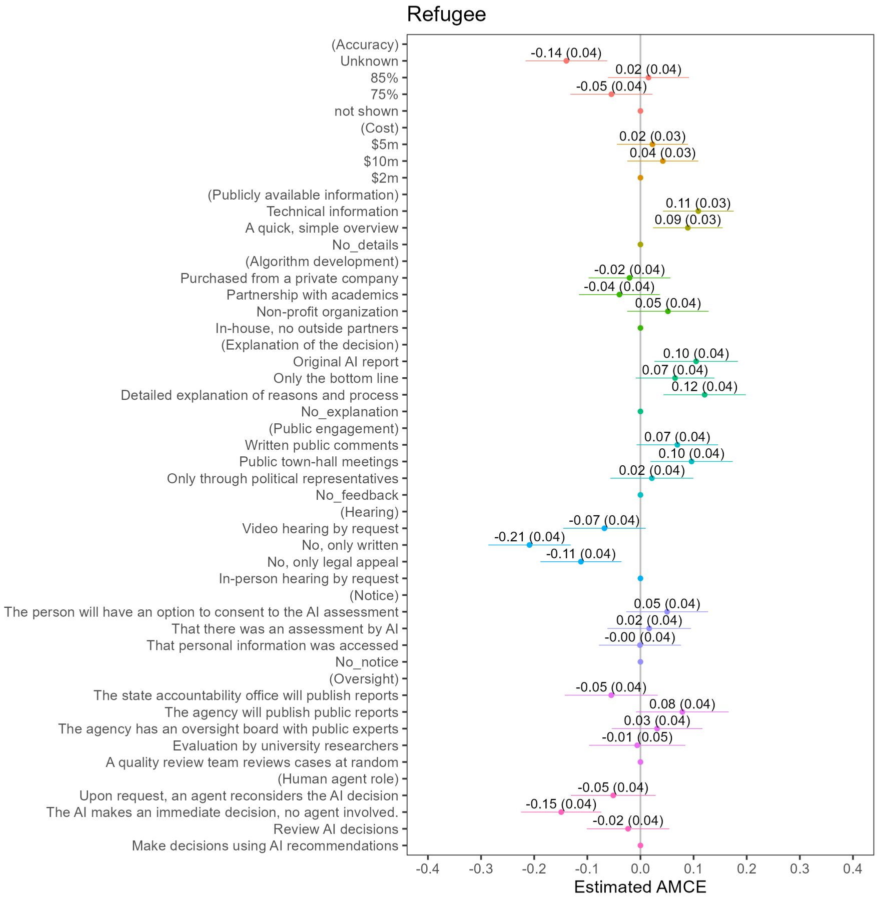


Figure 8

AMCE Estimates by Policy Domain: Permit

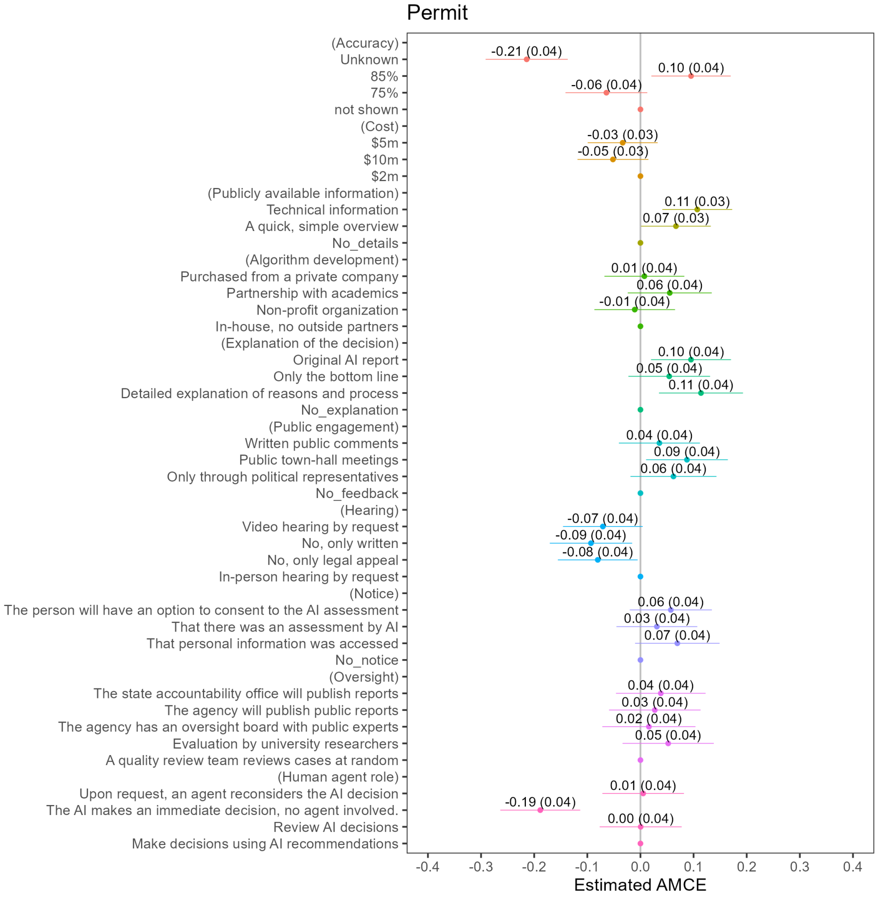
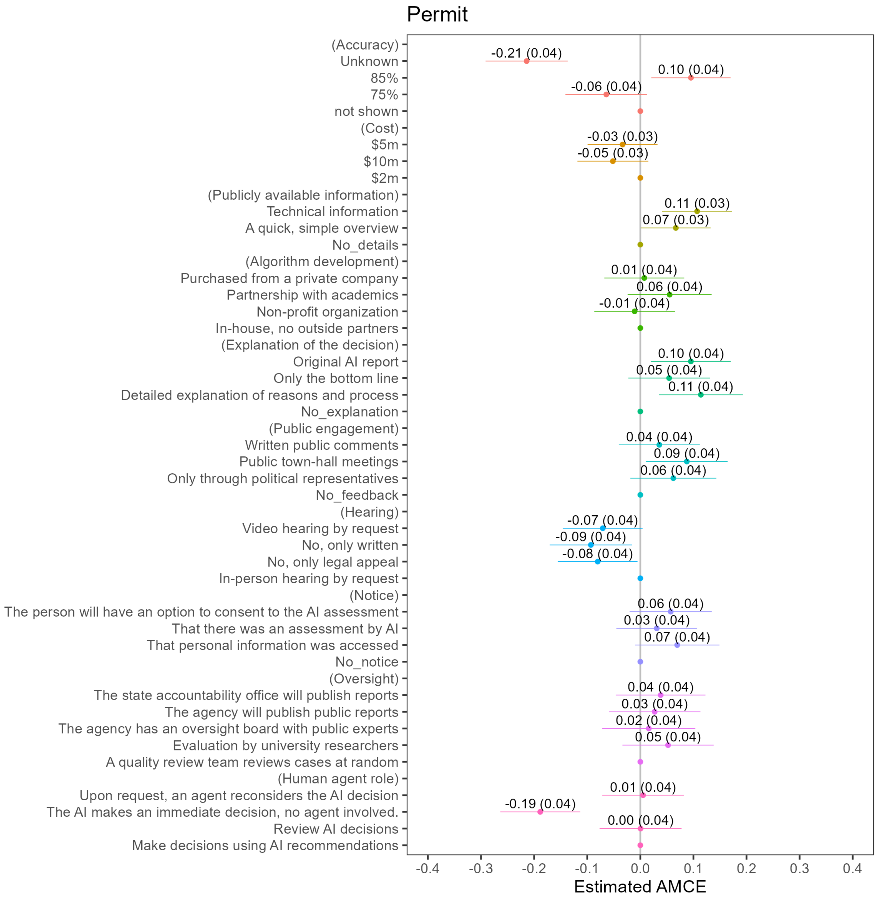


Figure 9

AMCE Estimates by Policy Domain: City

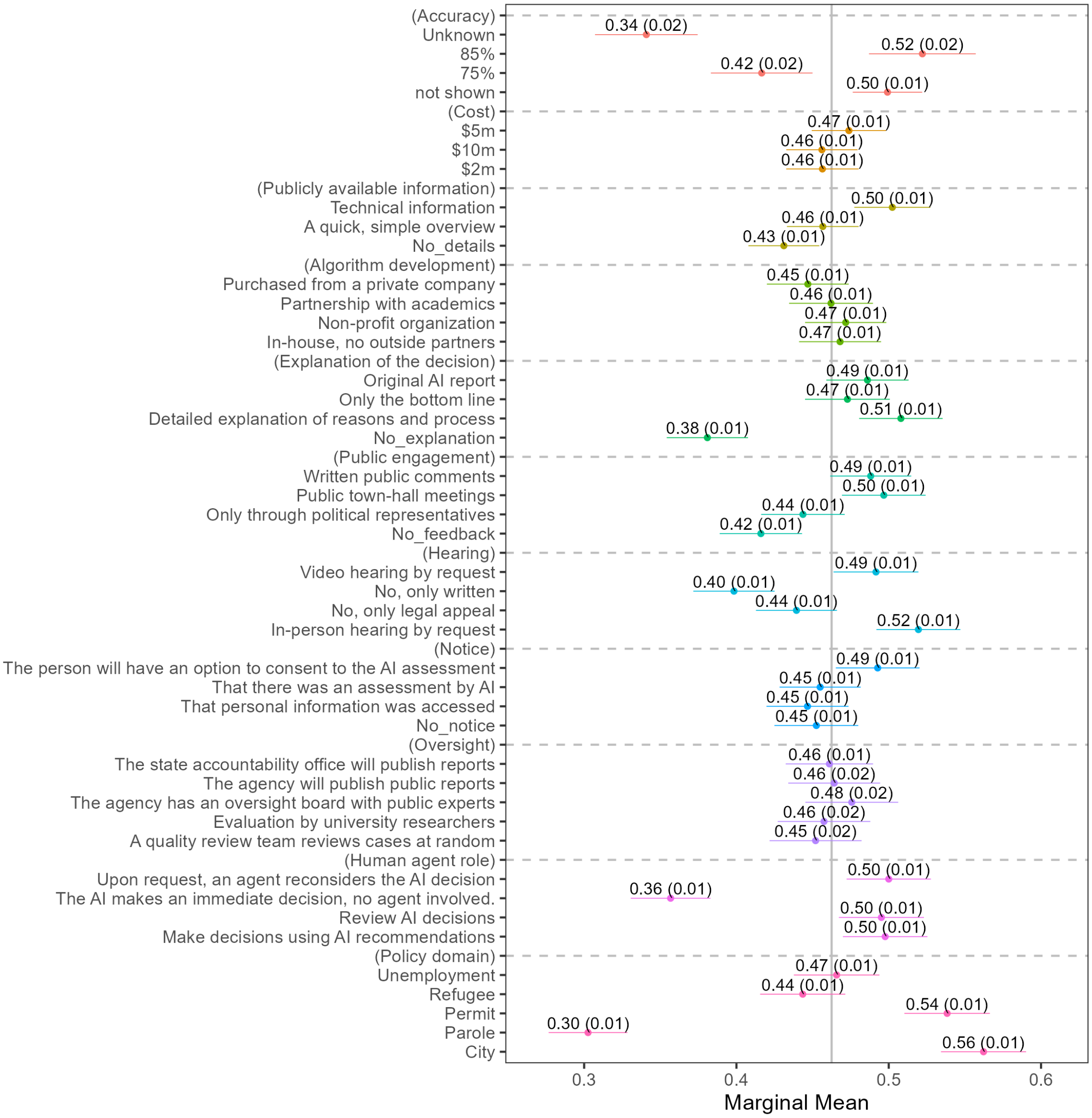


Supplement B: Marginal Means Analysis

Figure 10 presents the Marginal Means framework, allowing for a closer examination of the most notable levels across attributes.

Figure 10

Marginal Means of Main Outcome (Acceptability)



Note: This figure displays the Marginal Means (MM) for acceptability across different attribute levels. Analysis was conducted using a Gaussian Generalized Linear Model (GLM). Points represent estimated marginal means, and bars represent 95% confidence intervals.

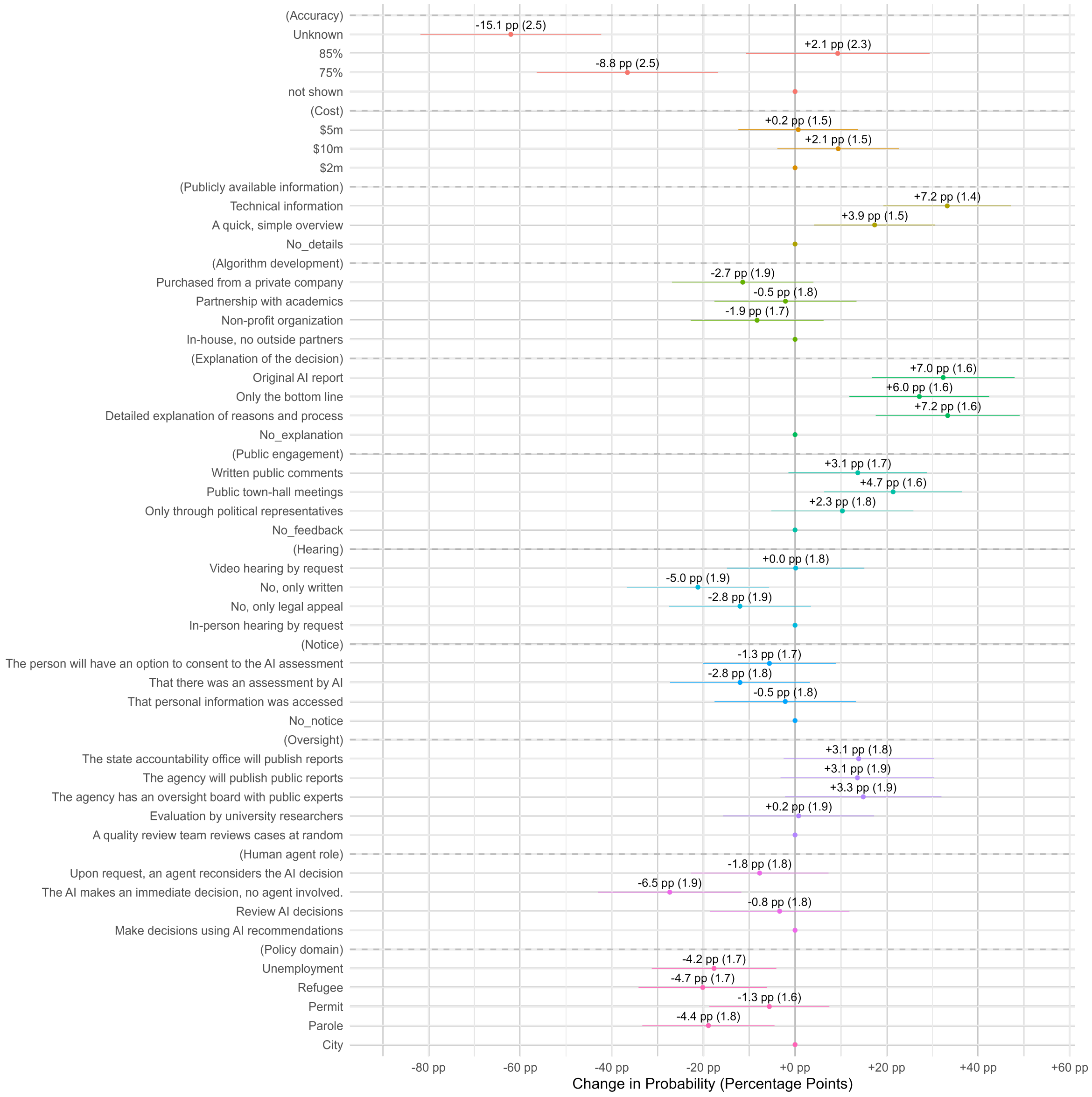
Supplement C: Additional Outcomes – Efficiency and Fairness

Efficiency

64.3% of programs were perceived as “efficient” compared to 46.2% for acceptability and 49.7% for fairness, t = -20.391, df = 12108, p<.001.

Figure 11

AMCE of Secondary Outcome: Efficiency

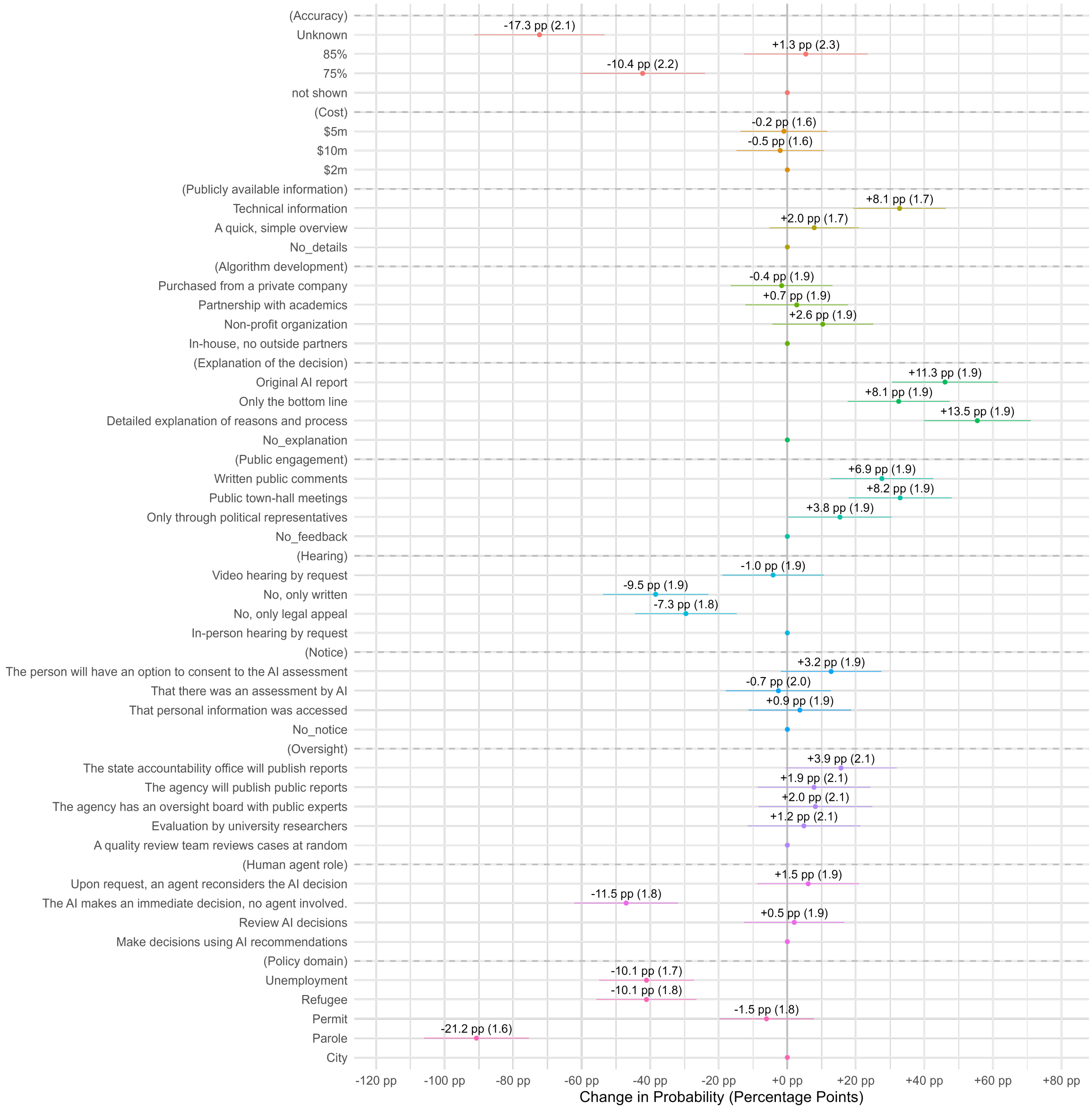


Fairness

49.7% of policies were seen as “fair."

Figure 12

AMCE of Secondary Outcome: Fairness



Supplement D: Statistical Power Analysis

To calculate the statistical power of our conjoint experiment, we followed established methodologies for analyzing forced-choice conjoint designs. With $N = 1,213$ participants each completing 5 choice tasks, we have a total of 6,065 observations.

Key Parameters

- Number of participants (N): 1,213
- Tasks per respondent: 5
- Total observations: 6,065
- Significance level (α): .05
- Desired power threshold: 80%

Calculation Process

For conjoint experiments, the standard error (SE) of the average marginal component effect (AMCE) can be approximated as:

$$SE = \sqrt{\frac{1}{N \times T}} \tag{2}$$

where N is the number of participants and T is the number of tasks per respondent. For our study:

$$SE = \sqrt{\frac{1}{6065}} \approx .0128 \tag{3}$$

The minimum detectable effect size (MDES) at 80% power can be calculated using:

$$MDES = (z_{1-\alpha/2} + z_{1-\beta}) \times SE \tag{4}$$

where:

- $z_{1-\alpha/2}$ is the critical value for a two-sided test at $\alpha = .05$ (1.96)
- $z_{1-\beta}$ is the critical value for power of 0.80 (0.84)
- SE is the standard error calculated above

This yields an MDES of .036 standard deviations at 80% power. For larger effect sizes:

- .05 SD: 97.3% power
- 0.10 SD: >99.9% power
- 0.15 SD: >99.9% power

These calculations account for the clustered nature of the data where observations from the same respondent may be correlated. The high statistical power is achieved by combining a substantial sample size and the efficiency gained from multiple observations per respondent in the conjoint design.

Consideration of Attribute Structure

Our design includes 10 attributes with varying numbers of levels:

- Public information (3 levels)
- Accuracy (4 levels)
- Algorithm development (4 levels)
- Human-agent role (4 levels)
- Explanation (4 levels)
- Notice (4 levels)
- Hearing (4 levels)

- Oversight (5 levels)
- Public engagement (4 levels)
- Cost (3 levels)

While the number of levels and attributes affects the frequency with which each specific level appears in the experiment, the fundamental power calculations for detecting average marginal component effects (AMCEs) depend primarily on the total number of observations (Hainmueller et al., 2014). Our randomized design ensures that each level appears with equal probability within its attribute, maintaining balanced statistical power across all comparisons.

Supplement E: Pilot Study

We conducted a pilot study with a smaller sample and four policy domains in May and June 2022. The results of the pilot study were reproduced in the main article, granting it additional robustness.

Sample

We sample 499 participants from the U.S. adult population, recruited through the Prolific online platform. Participants were paid according to a $10 hourly rate. The survey included a manipulation check, but no participants were disqualified on this or any other basis.

The sample contained about 42% male participants and 57% female participants, with the rest identifying non-binary or other sex categories. The mean age of the sample was 38.6 years (median 36). As per race and ethnicity, participants that identified as White comprised 79.6% of the sample, 10.4% Asian, 6.2% African-American, 5% as Latino (and 4.8% Hispanic; other categories were less than 1% of the sample; participants could choose multiple categories).

As per the extra randomization step, 51% of the sample was shown the attribute of accuracy, while 49% did not have the attribute included in the experiment table.

Results

Overall, 43.7% of the programs were deemed “acceptable,” and 44.7% were seen as “fair” (the difference is not significant). Both outcomes exhibit almost identical AMCE estimates across all attributes and levels.

Correlation between outcomes is presented in Table 8.

We present the results of our conjoint task asking participants to indicate whether they think the government program is acceptable or not in Figure 13. The figure displays the marginal effect each attribute had on the probability a respondent classified the

Table 7
Sample Descriptive Statistics

Statistic	*Mean*	*SD*	Min	Max
Age	38.65	13.76	19	83
Female	0.57	0.50	0	1
Liberal	0.54	0.50	0	1
Democrat	0.49	0.50	0	1
Education High School or less	0.15	0.35	0	1
High AI familiarity	0.41	0.49	0	1
No CS formal training	0.61	0.49	0	1
No programming experience	0.60	0.49	0	1
AI is beneficiary	0.58	0.23	.00	1.00
Pro AI Regulation	0.71	0.18	.00	1.00
Favorable of Technology	0.55	0.18	.00	0.95

program as “acceptable,” relative to the omitted level of each experimental factor. Within each, the omitted variable is displayed first (without a 95% confidence interval). Estimated effects are interpreted relative to the omitted categories. If we omit a different level for each category, the direction and appearance of the plot can change, but the relative distance between levels will not.

The pilot study’s results were reproduced in the main article.

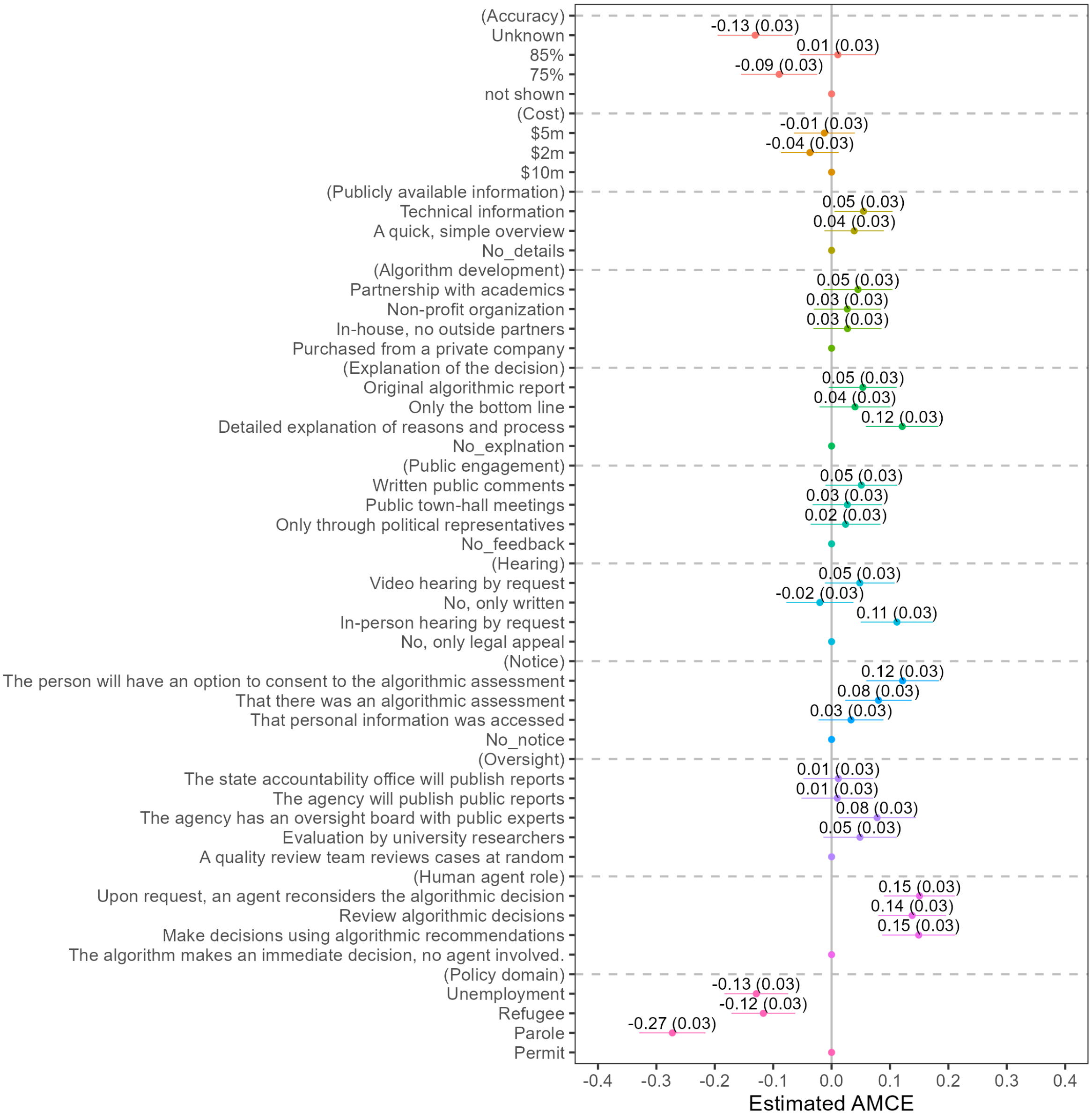
Table 8

Correlation between Outcomes: Acceptability, Fairness and Efficiency

	Acceptability	Fairness	Efficiency
Acceptability	1	0.94	0.61
Fairness	0.94	1	0.60
Efficiency	0.61	0.60	1

Figure 13

AMCE of Main Outcome (Acceptability)


(Accuracy)
Unknown
-0.13 (0.03)
0.01 (0.03)
-0.09 (0.03)
not shown
(Cost)
$5m
-0.01 (0.03)
$2m
-0.04 (0.03)
$10m
(Publicly available information)
Technical information
0.05 (0.03)
A quick, simple overview
0.04 (0.03)
No_details
(Algorithm development)
Partnership with academics
0.05 (0.03)
Non-profit organization
0.03 (0.03)
In-house, no outside partners
0.03 (0.03)
Purchased from a private company
(Explanation of the decision)
Original algorithmic report
0.05 (0.03)
Only the bottom line
0.04 (0.03)
Detailed explanation of reasons and process
0.12 (0.03)
No_explnation
(Public engagement)
Written public comments
0.05 (0.03)
Public town-hall meetings
0.03 (0.03)
Only through political representatives
0.02 (0.03)
No_feedback
(Hearing)
Video hearing by request
0.05 (0.03)
No, only written
-0.02 (0.03)
In-person hearing by request
0.11 (0.03)
No, only legal appeal
(Notice)
The person will have an option to consent to the algorithmic assessment
0.12 (0.03)
That there was an algorithmic assessment
0.08 (0.03)
That personal information was accessed
0.03 (0.03)
No_notice
(Oversight)
The state accountability office will publish reports
0.01 (0.03)
The agency will publish public reports
0.01 (0.03)
The agency has an oversight board with public experts
0.08 (0.03)
Evaluation by university researchers
0.05 (0.03)
A quality review team reviews cases at random
(Human agent role)
Upon request, an agent reconsiders the algorithmic decision
0.15 (0.03)
Review algorithmic decisions
0.14 (0.03)
Make decisions using algorithmic recommendations
0.15 (0.03)
The algorithm makes an immediate decision, no agent involved.
(Policy domain)
Unemployment
-0.13 (0.03)
Refugee
-0.12 (0.03)
Parole
-0.27 (0.03)
Permit
Estimated AMCE