


Governing algorithmic decisions: The role of decision importance and governance on perceived legitimacy of algorithmic decisions

Big Data & Society
 January–June: 1–16
 © The Author(s) 2022
 Article reuse guidelines:
sagepub.com/journals-permissions
 DOI: 10.1177/20539517221100449
journals.sagepub.com/home/bds


Ari Waldman¹  and Kirsten Martin² 

Abstract

The algorithmic accountability literature to date has primarily focused on procedural tools to govern automated decision-making systems. That prescriptive literature elides a fundamentally empirical question: whether and under what circumstances, if any, is the use of algorithmic systems to make public policy decisions perceived as legitimate? The present study begins to answer this question. Using factorial vignette survey methodology, we explore the relative importance of the type of decision, the procedural governance, the input data used, and outcome errors on perceptions of the legitimacy of algorithmic public policy decisions as compared to similar human decisions. Among other findings, we find that the type of decision—low importance versus high importance—impacts the perceived legitimacy of automated decisions. We find that human governance of algorithmic systems (aka human-in-the-loop) increases perceptions of the legitimacy of algorithmic decision-making systems, even when those decisions are likely to result in significant errors. Notably, we also find the penalty to perceived legitimacy is greater when human decision-makers make mistakes than when algorithmic systems make the same errors. The positive impact on perceived legitimacy from governance—such as human-in-the-loop—is greatest for highly pivotal decisions such as parole, policing, and healthcare. After discussing the study’s limitations, we outline avenues for future research.

Keywords

AI governance, legitimacy, algorithmic decision, algorithmic accountability, legitimacy dividend, legitimacy penalty

Introduction

Decision-making is increasingly automated. Governments are now relying on automated systems to make critical decisions affecting entitlements, education, and criminal justice (Citron and Pasquale, 2014; Lavorgna and Ugwudike, 2021; Pasquale, 2014; Wexler, 2018; Whittaker et al., 2018). Houston used an algorithm to determine promotions, bonuses, and jobs for its teachers (Houston Federation of Teachers v. Houston Independent School District, 2017). Idaho relied on an algorithm to cut disability benefits (K.W. v. Armstrong, 2016). U.S. border agents are using algorithms to make detention determinations (Sonnad, 2018). Some police forces leverage artificial intelligence (AI) to pacify populations on the basis of race and forecast crimes before they occur (Joseph and Lipp, 2018; Sheehy, 2019).

The increasing prevalence of automated decision-making belies its risks. Scholars have shown that these systems are data-extractive and biased (Benjamin, 2019;

Katyal, 2019; Kraemer et al., 2011; Loi et al., 2020; Martin, 2019; Noble, 2018). They are “black boxes” with little accountability (Innerarity, 2021; Pasquale, 2014). Citron (2007), Crawford and Schultz (2014), and Berman (2018) argue that algorithmic decision-making threatens due process. And, as de Laat (2019) has argued, algorithmic systems that are used to predict human behavior are polypnoptic in the Foucauldian sense: they subject us to surveillance, encourage a race to the norm, and undermine

¹Northeastern University, School of Law and Khoury College of Computer Sciences, Boston, MA, USA

²University of Notre Dame, Mendoza College of Business, Notre Dame, IN, USA

Corresponding author:

Kirsten Martin, Mendoza College of Business, Information Technology, Analytics, and Operations Department, University of Notre Dame, Notre Dame, IN 46556, USA.

Email: kmarti33@nd.edu

independence. If these critics are correct, algorithmic decision-making should pose a problem for democratic societies, which we define, following Dworkin (1996: 17), as a society in which “collective decisions [are] made by political institutions whose structure, composition, and practices treat all members of the community, as individuals, with equal concern and respect.” When authorities in democratic societies use technological tools that studies show erode privacy, evade accountability, and lead to arbitrary results, those technologies have a legitimacy problem.

Legitimate decisions are those that comport with our values and, importantly, inspire voluntary and willing compliance (Brummette and Zoch, 2016; Tyler, 1990/2006). Many studies show that, in democratic societies, unelected authorities gain legitimacy through a fair process, treating people with respect, and giving individuals opportunities to be heard, among other procedural factors (Sunshine and Tyler, 2003; Tyler and Huo, 2002). Unelected authorities—judges, police, and administrative agencies, for example—depend on perceived legitimacy in order to achieve their policy goals, maintain their positions, and function properly in democratic societies (Grimes 2006).

Given the risks that automated decision-making systems pose to the values of justice, freedom, and equality, this study seeks to determine whether and under what circumstances, if any, the public perceives that the use of algorithmic systems to make public policy decisions is legitimate.¹ More specifically, the prevalence of algorithmic decision-making systems requires us to ask, first, whether the trappings of governance help facilitate voluntary compliance with algorithmic decisions and, second, how, if at all, any of the known systemic problems with algorithmic systems affect those perceptions of legitimacy. Therefore, in this study, we consider the impact of decision type, governance, errors, and other factors on popular perceptions of the legitimacy of algorithmically derived social policy decisions in democratic societies.

Using factorial vignette survey methodology to survey individuals’ normative judgments about algorithmic and human decisions, we find that human governance (human-in-the-loop) increased the perceived legitimacy of algorithmic systems, even when those decisions are likely to result in significant errors. Notably, we also find the perceived legitimacy penalty is greater when human decision-makers make mistakes than when algorithmic systems make the same errors. This may be explained by a lingering effect of human trust in technology (de Vries and Midden, 2008; Figueras et al., 2021; Hoff and Bashir, 2015; Jacovi et al., 2021). Moreover, privacy issues, including data governance and data source, offer neither a perceived legitimacy dividend nor penalty, suggesting individuals perceive public policy decisions made by an algorithm as less a privacy issue than a question of justice or due process.

Theory

The functions and risks of algorithmic decision-making

This study focuses on algorithmic decision-making systems, which, following Calo (2017), we define generally as processes involving algorithms, or sequences of logical, mathematical operations, to implement policies by software. Some algorithmic decision-making tools are powered by AI of varying maturity and types. We also recognize that there exists a range of automation for decision-making, with some decisions almost fully computerized while others are merely augmented with technology (Martin, 2018). Defining artificial intelligence and classifying all algorithmic decision-making systems is not our goal. As this study focuses on perceptions of the legitimacy of decisions made by humans as compared to decisions augmented by machines, a definition that recognizes the role of data inputs, computers, and automation of decisions suffices.

So defined, algorithmic decision-making systems try to identify meaningful relationships and likely patterns in large data sets (Cormen, 2009). Governments use algorithms that analyze internet browsing behavior, purchase histories, residence zip code, criminal history, employment, educational achievement, and family relationships, among myriad other variables, to predict whether someone should receive in-home health services or whether someone is likely to commit another crime after release from prison (Angwin et al., 2016; Crawford and Schultz, 2014; Federal Trade Commission, 2016).

Unfortunately, researchers have shown that because algorithmic systems make probabilistic predictions about the future, they still make mistakes; probabilities are necessarily generalized, with individual cases falling through the cracks (Eubanks, 2018; Hu, 2016). Studies have also shown that AI’s predictive capabilities may be exaggerated (Dressel and Farid, 2018; Jung et al., 2017; Salganik et al., 2020). AI’s opacity makes algorithmic systems difficult to interrogate and hold accountable (Colaner, 2021; Innerarity, 2021; Loi et al., 2020). Algorithmic systems also incent surveillance and data collection because they need large information sets for model training and analysis. This creates the circumstances for invasions of privacy and erosion of privacy norms (Ohm, 2010; Zwitter, 2014). Finally, algorithmic decision-making systems are biased. Algorithmic systems can only be as good as the corpus of data on which they are based and, as such, data that is biased along with race, gender, sex, and socioeconomic lines will lead to biased results (Bridges, 2017; Caliskan et al., 2017; Katyal, 2019; Noble, 2018; O’Neil, 2016; Sühr et al., 2021). Although it is true that some of these concerns may be ameliorated with design changes, algorithms’ deficiencies may threaten the legitimacy of authorities in democratic societies (Berman, 2018).

Background: definitions and related literature

In measuring individuals' perceptions of decisions, scholars have used trustworthiness, fairness, and legitimacy at the individual, organizational, and system levels (Jacovi et al., 2021; Lee, 2018; Veale and Binns, 2017). Each lens is distinct, based on different factors and sources. According to Kaina (2008), legitimacy is a distinct concept from fairness and trust. For example, fairness measures focus on distributive, procedural, and interactional justice components (Skarlicki and Folger, 1997); trustworthiness is driven by the ability, benevolence, and integrity of the subject (Pirson et al., 2019). In this section, we review the literature on legitimacy, describe related studies of algorithmic legitimacy, and identify gaps in the current literature.

Legitimacy. Based on Weber (1947/2012), legitimacy refers to the public's willingness to accept the validity of authorities' actions (Lipset, 1959). Suchman (1995) defines legitimacy as "a generalized perception or assumption that the actions of an entity are desirable, proper, or appropriate within some socially constructed system of norms, values, beliefs, and definitions." Tom Tyler (1990/2006) and his colleagues (Tyler and Huo, 2002: 102) define legitimacy more narrowly, as either "perceived obligation to comply with the directives of an authority, irrespective of the personal gains" or "a quality possessed by an authority, a law, or an institution that leads others to feel obligated to obey its decisions and directives voluntarily." Either way, both Suchman and Tyler recognize that there is a moral valence to legitimacy. When decision-making is perceived to be legitimate, it carries a moral presumption of validity. Therefore, legitimacy is of critical concern to social scientists, lawyers, ethicists, and technologists, especially in democratic societies. As Bohman (1998: 400) has noted, deliberative democracy depends on decisions that "everyone could accept" or "not reasonably reject."

In canonical studies of legal legitimacy, Tyler (1990/2006) and Sunshine and Tyler (2003) showed that popular perceptions of legitimacy and, in turn, a general willingness to accept the decisions of authorities, hinges at least in part on the existence of procedural safeguards and the opportunity to be heard. In Tyler's work, the legitimacy dividend of fair processes overcomes any lingering distrust, opposition, or negative reaction associated with an adverse result (Tyler, 1994). That is, even those individuals who came out worse off due to the actions of authorities, institutions, or law were willing to comply with the law if the process was fair (Easton, 1965). Legal studies scholars have suggested that a fair process can increase perceptions of legitimacy of judicial decisions (Gibson and Caldeira, 2009; Gibson et al., 2003). However, Badas (2019), Bartels and Johnston (2013), and Christenson and Glick (2015) have challenged the conventional wisdom

that process rather than outcome is the most significant factor in judicial legitimacy. Their work has elevated the role of policy disagreements: those who disagree with judicial decisions tend to think the judge or court is less legitimate.

Based on the scholarship discussed in this section, we query whether and how inputs (data), process (types of governance), and outputs (errors) affect perceptions of the legitimacy of algorithmic decisions. This study asks respondents to report their perception of legitimacy based on these factors.

Related studies on legitimizing algorithmic systems. Recent research has begun to identify factors that legitimize algorithmic systems, broadly construed. For instance, de Fine Licht and de Fine Licht (2020) suggest that a limited form of transparency that focuses on providing justifications for decisions could provide grounds for the perceived legitimacy of algorithmic decisions. However, Leicht-Deobald et al. (2019) theorize that algorithm-based decisions may be perceived as *more* legitimate because the individuals cannot question the system (Leicht-Deobald et al., 2019). The literature is theoretical and thus invites empirical research to understand the intersection between transparency and legitimacy.

The second stream of scholarship explicitly links the legitimacy of algorithmic decisions with governance (Danaher et al., 2017). In accord with Tyler (1990/2006), fair procedures and accountability should legitimize algorithmic systems in the eyes of the public. Citron (2007: 1305–1313) was among the first legal scholars to call for replacing old forms of agency adjudication and rule-making with audit trails, educating hearing officers on machine fallibility, detailed explanations, publicly accessible code, and systems testing, among other recommendations. The goal was to bring algorithmic systems under the umbrella of traditional accountability regimes. Other scholars have proposed a right to explanation, which ostensibly entitles individuals to clarity about the process behind a model's development with the goal of improving acceptance of algorithmic decisions (Selbst and Barocas, 2018: 1087) and satisfying an individual's dignitary right to "understand why" results came out the way they did (Zarsky, 2013: 325).

Another strain of scholarship is more specific as to the type of transparency required for legitimate, fair, and ethical algorithmic decisions. Martin (2019) suggests that the design of AI must include a process to identify, judge, and fix inevitable mistakes. In a similar vein, scholars suggest that humans in the loop can fix errors and place guard rails around absurd, unethical, or inappropriate results (Henderson, 2018; Jones, 2017; Rahwan, 2018). Studies have measured the relative importance of transparency in using algorithms (König et al., 2022) and find explainability may not be intrinsically valuable (Colaner,

2021). Other scholars focus on impact statements (Katyal, 2019; Metcalf et al., 2021; Reisman et al., 2018) modeled after environmental or privacy impact assessments, or *ex ante* transparency as to goals and metrics (Loi et al. (2020) to document and assess a system's fairness. These scholars all focus on the capacity of procedures to make "better" algorithmic systems.

Missing from this research, however, is an empirical exploration of the specific and necessary conditions for popular perceptions of the legitimacy of decision-making systems, human and algorithmic alike. That is, it remains unclear what effect, if any, these policies actually have on perceptions of the legitimacy of the ultimate decision. In designing the study, we build on existing studies that attempt to measure the legitimacy of algorithmic systems. Lünich and Kieslich (2021) measured the role of trust and social group preference in the perceived legitimacy of algorithmic versus human decision-making. The authors studied the attributes of the respondents in driving perceived legitimacy and found general trust in automated decisions and a preference for vaccines impacts the perceived legitimacy of the vaccine allocation.²

In previous studies on legitimacy and algorithmic decisions, Danaher et al. (2017) asked participants about algorithmic *governance* (not decisions) with a single question: "What are the barriers to legitimate and effective algorithmic governance" via email. Starke and Lünich (2020) empirically examine input, process, and output legitimacy of EU decisions. Output legitimacy was measured by perceived goal attainment and with different goals listed as well as whether the individual agreed with the decision. Input legitimacy was measured by whether respondents perceived that the correct people participated in the decision—namely, if people like them participated. Process legitimacy was operationalized as comprising of appropriateness, fairness, and perceptions of satisfaction (none were defined). What Starke and Lünich (2020) did not do, and what we attempt to do here, is empirically assess the relative importance of inputs, processes, and outcomes to the perceived legitimacy of algorithmic decisions. As Persson et al. (2013, p. 391) rightly noted, "legitimacy is an inherently abstract concept that is hard to measure directly." Our study contributes to this research by empirically assessing the relative impact of specific governance factors, including types of governance, errors, data use, and biases, on perceptions of legitimacy of automated decision-making.

Study design and hypotheses

We used a factorial vignette survey methodology to survey individuals' normative judgments about both human and algorithmic decisions made by government actors. This allowed us to compare how the legitimacy of algorithmic decisions differed, if at all, from the human decisions, all else being equal.

A factorial vignette survey presents randomized respondents with a series of scenarios where several factors are systematically varied in the vignette; respondents then judge the scenario using a single rating task (Jasso, 2006; Wallander, 2009). This methodology allows researchers to study how different features of the vignettes affect participants' attitudes, judgments, or views. Subsequent statistical analysis allows researchers to determine not just which factors drive the respondents' judgments, but the extent to which changes in the presence and degree of those factors have a significant (or insignificant) effect on participants' views.

In addition, the single rating task in the factorial vignette survey methodology supports the inductive measurement of the concept—here, the legitimacy of the decision—through the analysis of the factors in the vignette. In other words, the researcher is able to measure the relative importance of the vignette factors in driving the perception of legitimacy of the respondent. Previously, the methodology has been used to measure the relative importance of vignette factors on just wages (Jasso, 2007); the relative importance of the vignette factors on just punishments (Hagan et al., 2008); the relative importance of vignette factors on the trustworthiness of an organization (Pirson et al., 2017); the relative importance of vignette factors on privacy expectations (Martin and Nissenbaum, 2020). In each case, the vignette factors constitute the theoretically important constructs that may drive the perception of trust, fairness, or privacy of the respondent; the respondent is provided a single rating task to rate the vignettes. The researcher is able to then identify how each factor drives the perception of the dependent variable.

In this study, decision factors were independently varied with replacement, and the respondents judged the degree to which the decision described was legitimate. This design avoids two types of issues in typical surveys. First, the respondents are forced to use a single rating task while taking into consideration multiple factors at the same time. Second, the design avoids respondent bias where respondents attempt to answer survey questions to appear more ethical or to please the researcher.

Each respondent was presented with thirty short vignettes describing either a human or algorithmic decision. In general, the vignettes' narrative had five elements: a decision-maker, a decision type, source of data for the decision, the decision error rate, and the governance regime, if any, for the decision. The elements are illustrated in Table 1 and described in more detail below. A general outline of the vignettes and samples of how they were presented to survey respondents is provided in Appendix A.

Vignette factors and hypotheses

Decision-Maker: In order to isolate the importance of algorithmic decisions under each condition, we varied whether

Table 1. Vignette factors and levels as operationalized in the vignettes.

Factors	Levels	Operationalized	Mistakes
Decisions	Potholes	which potholes are fixed ... the worst streets in the city ... the city government ... for each neighbourhood	Where wrong potholes got fixed XX% of the time.
	Police	Deciding which neighborhoods are patrolled ... The likelihood people will commit crimes in that neighborhood ... the police ... for each neighborhood	Where the wrong neighborhood is patrolled by police XX% of the time.
	Education	allocation of state funds for education ... the schools with the greatest needs ... the state's board of education... for each school district...	Where the wrong schools received the funds XX% of the time.
	Paroled	who gets released early from prison the estimated risk of recidivism ... the parole board ... for each prisoner.	The wrong prisoner remains in prison XX% of the time.
	Health Services	The services someone receives under Medicare or Medicaid (e.g. in home health aid, wheelchair, physical therapy, etc) the patient's need for assistance ... the state's board for human services ... for each patient.	Patients were incorrectly denied care XX% of the time.
Data Type	General Specific (NULL)	gathered from individuals' online browsing behavior gathered directly for this purpose	
Mistakes	% wrong	10, 20, 30, 40, 50.	
Governance	Human Only (NULL)	Entire vignette is about human decision	
	Null	The organization does not notify citizens or provide oversight over the decision.	
	Human Governance Notification	The organization hires a privacy professional to oversee the decision. The organization notifies citizens that a computer program makes the decision.	

the decision was made by a group of individuals or a computer program by running separate survey conditions. Similarly, Starke and Lünich (2020) found that AI-based decisions were perceived as less legitimate than human-involved decisions, and Leicht-Deobald et al. (2019) make the theoretical case that algorithmic decisions may be perceived as *more* legitimate. Both rightly argue that any measurement of legitimacy of an algorithmic decision should be compared to a similarly situated *human* decision.

Decision: In order to determine whether the perceived legitimacy of human versus algorithmic decision-making varies across types of decisions, we included five types of decisions in the distribution of social goods: which *potholes* will get fixed, how *funds will be allocated to schools*, how police officers will be allocated to *patrol neighborhoods*, which *health services* patients will receive under Medicare or Medicaid, and who will be released on *parole from prison*. These decision types were chosen based on how pivotal they are in society within the

current literature (Burrell, 2016; O'Neil, 2016; Tufekci, 2015). The degree to which decision type was deemed pivotal was verified by a pre-test survey of 400 respondents on Amazon Turk. The respondents were asked: "Please rate the degree to which the following decision represents a critical decision affecting someone's life." Potholes were the least important (-8.00 on average), education and police were in the next tier (42.3/42.4 on average, respectively), and health and parole were the most important (66.9/66.2 on average, respectively). Our first hypothesis is that the perceived legitimacy of algorithmic decision-making varies significantly with decision importance. More specifically,

Hypothesis 1 There is an inverse relationship between decision importance and the perceived legitimacy of an algorithmic decision. As decision importance increases, algorithmic decision-making systems are perceived as less legitimate.

Data Type: Decision-making depends on data inputs or information the decision-maker uses to make a decision. Algorithms also depend on both data inputs and a corpus of training data (Caliskan et al., 2017). Current scholarship and policy debates in privacy law and algorithmic accountability have recognized this, proposing several procedural guardrails around the collection and use of personal information (Citron and Pasquale, 2014; Kaminski, 2019a, 2019b; Katyal, 2019). Research also suggests that individuals generally disapprove of data brokers' data collection and data aggregation practices (Martin and Nissenbaum, 2017). Given the centrality of data in the theoretical legal scholarship on algorithmic systems, we included in the vignettes two data types that could be used for either human or algorithmic decisions: data about an individual that was gathered for the specific purpose of making this decision or general data about an individual that was aggregated across online sources. Our second hypothesis is that the perceived legitimacy of algorithmic decision-making is influenced by the type the data used. More specifically,

Hypothesis 2 There is a positive impact on perceived legitimacy for algorithmic decisions when data inputs are gathered for the specific purpose of making the decision as compared to algorithmic decisions based on general data about an individual collected from the internet.

Mistakes: Decision outcome affects the perceived legitimacy of the algorithmic decision (Starke and Lünich, 2020). Therefore, because both human and algorithmic decision-makers make mistakes, it is worth studying the impact of mistakes on perceived legitimacy. To integrate mistakes into the model, we varied the percent of mistaken outcomes and particularized the mistake to the type of decision. The percent of mistakes was systematically varied (randomly) among 10, 20, 30, 40, and 50%. We then designed the vignettes to ensure that the type of mistake made sense with the type of decision. For example, the element of the vignettes reflecting mistakes looked like this: *the wrong potholes got fixed XX% of the time, the wrong neighborhood is patrolled by police XX% of the time, the wrong schools received the funds XX% of the time, the wrong prisoner remains in prison XX% of the time, and patients were incorrectly denied care XX% of the time.* We hypothesized that legitimacy and error rates are inversely related. More specifically,

Hypothesis 3a Both human and algorithmic decisions experience decreases in perceived legitimacy as mistake rates increase.

Hypothesis 3b Given the expected relationship between decision importance and perceived legitimacy, we also

expect to find that the legitimacy penalty for mistakes to be greater for more pivotal decisions than for less pivotal decisions.

Governance and Mistakes: Current policy debates center not on whether algorithmic decision-making requires some form of accountability mechanism, but on what that governance regime should look like (Pasquale, 2019). For example, de Fine Licht and de Fine Licht theorize that the lack of transparency would decrease the perceived legitimacy of algorithmic decisions (de Fine Licht and de Fine Licht, 2020). To test whether the type of governance over an algorithmic decision impacted the degree to which the decision is perceived to be legitimate, we included three options for the algorithmic vignettes: notification that a computer program is being used, a human overseeing the algorithmic decision, and a control governance option where the organization neither notifies or has a human overseeing the algorithmic decision. All three options are used in practice, with many government agencies using algorithms without full transparency, providing notice of an algorithm's use in a privacy policy, or, including a "human in the loop." There are, of course, other possible governance options. We chose these three options because many of the other, arguably more robust governance regimes in the literature—legal remedies or nondiscrimination, for example—may not be familiar to nonexperts and, therefore, simple and accessible options were more likely to lead to better results. We hypothesize that a governance regime will make algorithmic decision-making systems appear more legitimate than those without any governance at all, and that the decision's importance will be inversely related to the legitimacy dividend that comes with a more robust governance regime. We also expect governance regimes to somewhat insulate algorithmic decision-making from error-related legitimacy penalties.

Hypothesis 4a As compared to no governance, any governance regime makes the algorithmic system perceived to be more legitimate. Algorithmic decisions with a human-in-the-loop governance mechanism are perceived as more legitimate than mere notice.

Hypothesis 4b As decision-importance increases, the legitimacy dividend from human-in-the-loop governance would decrease.

For each vignette, respondents were instructed to indicate on a slider the degree to which they agreed with the statement: "This decision is legitimate." The left side of the slider indicated "Strongly Disagree" and the right side of the slider indicated "Strongly Agree." The slider was on a scale of -100 to $+100$ with the number not visible to the respondents.

Vignette template

The template for the vignettes in this study is as follows:

In order to decide {Decision Context}, a computer program uses information that {Decision Context} {Data Type} to identify {Decisions Social Goods_alt2}.

Upon review, approximately {Mistakes} of the decisions were incorrect—{Decisions Context} {Mistakes} of the time.

The organization {Governance Factor}

Example:

In order to decide the services someone receives under Medicare or Medicaid (e.g. in home health aide, wheelchair, physical therapy, etc), a computer program uses information that the state’s board for human services gathered specifically for this purpose to identify the patient’s need for assistance.

Upon review, approximately 10% of the decisions were incorrect—where patients were incorrectly denied care 10% of the time.

The organization hires a professional to oversee the decision.

The sample

The surveys were run on Amazon Mechanical Turk, a crowdsourcing marketplace where researchers publish a job (“HIT”) for respondents to take a survey. Each respondent rated 30 vignettes taking approximately 10 min; U.S. respondents were paid \$1.70 and were screened for over 95% HIT approval rate. The survey implementation was designed to minimize a number of concerns with samples from Amazon Mechanical Turk. First, the factorial vignette survey methodology was created to avoid respondent bias in normative judgments—namely, where respondents might try to game the system to appear more ethical or socially desirable. Second, the structure of the data—in two levels with individuals at the first level and vignette ratings at the second level—supports the researcher in calculating whether respondents “clicked through” without actually

Table 2. Sample descriptive statistics.

	Human Decision Survey		Algorithmic Decision Survey	
N	305		294	
Vignettes	9180		8820	
Age				
18–24	30	10%	28	10%
25–34	125	41%	123	42%
35–44	76	25%	83	28%
45–54	39	13%	32	11%
55–64	27	9%	20	7%
65+	8	3%	8	3%
Gender				
Male	156	53%	168	58%
Female	139	47%	122	42%

Table 3. Regressions of vignette factors on legitimacy dependent variable.

	Human Decision		AI Decision	
	β	<i>p</i>	β	<i>p</i>
Decision Type				
EducDecision	−7.29	0.00	−9.59	0.00
HealthDecision	−12.02	0.00	−13.29	0.00
ParoleDecision	−9.84	0.00	−13.55	0.00
PoliceDecision	−2.45	0.10	−7.51	0.00
<i>Null = Pothole Decision</i>				
Type of Data				
AggregatedData	−22.76	0.00	−8.33	0.00
<i>Null = Specific Data</i>				
Governance (AI only)				
NullGovernance	n/a	n/a	−23.02	0.00
HumanGovernance	n/a	n/a	11.08	0.00
<i>Null = Notice Only</i>				
Mistakes				
PercentMistakes	−14.84	0.00	−10.80	0.00
<i>Continuous 1–5 (10%–50%)</i>				
N (Users)	305		294	
N (Vignettes)	9180		8820	

judging the vignette (Coppock, 2018; Daly and Nataraajan, 2015; Martin, 2019; Tucker, 2014).³ Finally, the survey was designed to identify theoretically generalizable results as to the relative importance of factors in driving perceptions of legitimacy of algorithmic decisions.⁴ Thus, the sample is not designed to be nationally representative as the goal of the study is not data generalizability.

In the end, 294 respondents rated algorithmic decisions and 305 respondents rated human decisions. Each time, twice as many respondents were assigned the algorithmic decision vignette to support the additional analysis conducted on algorithmic vignettes below (Table 2).

Table 4. Average legitimacy score across decision type (automated decisions only).

Decision Type	Ave Legitimacy Rating
Potholes (least pivotal*)	1.88
Education	-7.78
Police	-5.47
Health Services	-11.27
Parole (most pivotal)	-13.49

*Degree pivotal based on pre-test as explained above.

Results

Decision type

Hypothesis 1 suggested that as decision importance increases, algorithmic decision-making systems should be perceived as less legitimate. Our analysis lends credibility to our hypothesis. To test this hypothesis, we regressed the rating task, the degree the vignette is legitimate, on the vignette factors. The results are in Table 3. Table 3 includes the coefficients for each decision type with the least pivotal decision—pothole decisions—as the null. Highly pivotal decisions, such as healthcare ($\beta = -13.29, p < 0.001$) and parole ($\beta = -13.55, p < 0.001$), were judged less legitimate for algorithmic decisions compared to decisions judged to not be important (fixing potholes). This can also be seen in Table 4 with the average legitimacy rating for each type of decision. Only pothole decisions were on average rated legitimate, all else being equal (Avg. = 1.88).

Data types

Hypothesis 2 predicted a positive impact on perceived legitimacy (a legitimacy dividend) for algorithmic decisions when data inputs are gathered for the specific purpose of making the decision as compared to algorithmic decisions

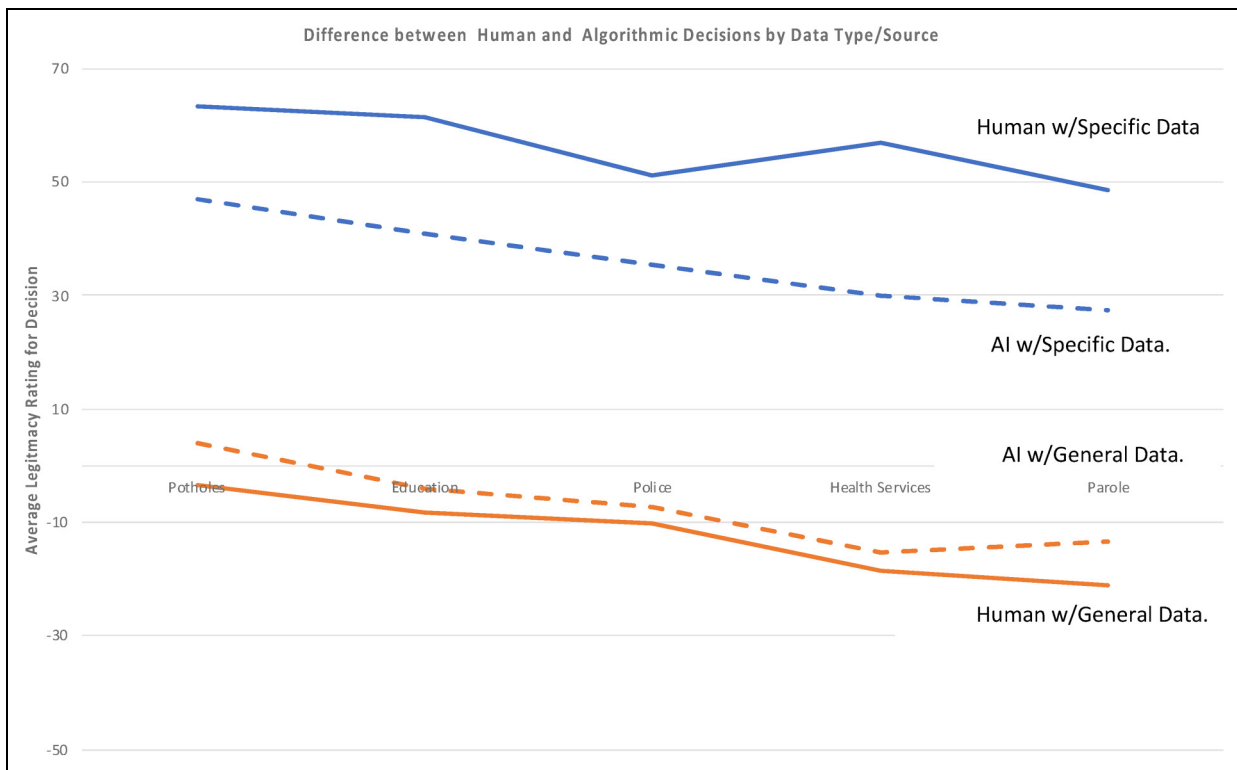


Figure 1. Average legitimacy score across decision type by data type/source for both algorithmic and human decisions.

based on general data about an individual collected from the internet. In general, respondents judged human and algorithmic decisions as less legitimate when based on more general data than when based on specific data for a given decision for both algorithmic and human decision-making as shown in Table 3. The impact on perceived legitimacy for using general data about an individual is negative and significant for both human decisions ($\beta = -22.76, p < 0.01$) and algorithmic decisions ($\beta = -8.33, p < 0.01$).

Figure 1 also demonstrates the positive impact on perceived legitimacy for algorithmic decisions based on internal, specific data compared to commercial, general data by comparing the blue line vs. the orange line. This can be framed in two ways. First, the benefit of using data gathered for this specific purpose can be seen by comparing the solid blue line (human decision) to the solid orange line (algorithmic decision) across decision types. The benefit of using specific data is actually larger for human decisions compared to algorithmic decisions. In addition, the average perceived legitimacy across all decision types is positive for data gathered for a specific purpose but averages negative for commercially gathered data.

In addition, when using general data about an individual, the perceived legitimacy for *human* decision-making (-22.76) is lower than the legitimacy for algorithmic decision-making (-8.33 ; $X^2 = 76.83, p < 0.001$). This would mean that individuals find the use of an algorithm in decision-making to be more legitimate than a human-led decision when using general data about the individual, all else being equal.

Mistakes

Hypothesis 3a suggested that the perceived legitimacy of algorithmically driven decisions will decrease as the

percent of outcomes that are mistakes increases. We found our hypothesis was correct. To test hypothesis 3a, we analyzed the relative importance of mistakes on the legitimacy of decisions in the regression results in Table 3. For human decisions, each increase in the percent mistakes (e.g. from 10–20% or from 20–30%) constitutes a legitimacy penalty of -14.84 ($p < 0.01$) all else being equal. For algorithmic decisions, each increase in the percent mistakes constitutes a legitimacy penalty of -10.80 ($p < 0.01$). Table 5 shows the average legitimacy rating for different percent mistakes estimated. Only decisions estimated at 10% mistakes are rated positively for algorithmic decisions.

In addition, in *hypothesis 3b, we expected the legitimacy penalty for mistakes to be greater for pivotal decisions.* In other words, although we expect mistakes to negatively impact the perceived legitimacy of a decision, in general, we also expect mistakes in highly pivotal decisions to negatively impact perceived legitimacy even more. Table 7 shows the legitimacy penalty for mistakes for each type of decision, where the perceived legitimacy decreases as mistakes increase.

We split the sample by the five types of decision and regressed the rating task—the perceived legitimacy—on the vignette factors for each type of decision. Table 6 shows the coefficient for the mistake factor in a regression run by decision type. For each decision type, from decisions of low importance (potholes) to decisions of high importance (parole), mistakes negatively impact perceived legitimacy and create a legitimacy penalty; however, the legitimacy penalties are not statistically different across decision types.

Finally, we reran the regressions of the legitimacy rating task on the vignette factors with an interaction between each

Table 5. Average legitimacy rating across mistake percentage (automated decisions only).

Percent of Mistakes	Ave Legitimacy Rating (Automated)
10%	17.74
20%	-0.08
30%	-9.23
40%	-19.65
50%	-25.61

Table 7. Average legitimacy rating across governance types.

Governance	Average Legitimacy Rating
AI Decision—Human Governance	8.94
AI Decision—Notice Governance	-3.86
AI Decision—No Governance	-26.39
Human Decision	-8.72

Table 6. Legitimacy penalty for mistakes by decision type.

Legitimacy Penalty	Coefficient for Mistake for Regressions on Decision Sub-Samples (Sample split by decision type)				
	Potholes	Education Decision	Police	Health	Parole
For Mistake (β)	-12.07**	-10.14**	-10.53**	-10.26**	-11.06**

* $p < 0.05$; *** $p < 0.001$.

Table 8. Legitimacy dividend for human-in-the-loop over notice governance by decision type.

Legitimacy Dividend	Regression Results of Sample Split By Decision Type				
	Within split sample by ...				
	Pothole	Education	Police	Parole	Health
Human-in-the-Loop (β) (Null = Notice)	2.34	8.68*	16.77**	18.98**	12.72**
X^2 (p) for cross-sample comparison		<i>n/s</i>	10.40 (0.001)	13.08 (0.001)	5.14 (0.02)

* $p < 0.01$; ** $p < 0.001$.

decision type and the mistake factor. No interaction was significant between decision type and mistake, meaning the coefficient for mistake was consistent across all decision types. Contrary to expectations, the legitimacy penalty for increasing mistakes in the output was consistent across decision types.

Governance

In *hypothesis 4a*, we predicted algorithmic decisions with a human-in-the-loop governance mechanism are perceived as more legitimate than mere notice. We found this to be true. The regression results in Table 3 show that both notice and human oversight result in a significant positive impact on the perceived legitimacy over no governance for algorithmic decisions. In Table 3, having no governance has a negative impact on perceived legitimacy ($\beta = -23.20$, $p < 0.01$) for algorithmic decision-making and including a human-in-the-loop provides a legitimacy dividend ($\beta = 11.08$, $p < 0.01$) compared to mere notification. Measured another way, Table 7 includes the average legitimacy rating for each governance type: human governance of algorithmic decisions has a higher perceived legitimacy rating than even human decisions. Only algorithmic decisions with human governance have a positive legitimacy rating.

Governance and Type of Decision. We then examined whether the increase in legitimacy for including a human in the loop for governance remains consistent across decision types. This was hypothesized in *hypothesis 4b*: as decision importance increased, the legitimacy dividend from human-in-the-loop governance would decrease. In other words, the increase in perceived legitimacy from governance mechanisms would be moderated by the importance of the decision: more important decisions would not see the legitimacy benefit of governance as would decisions of low importance.

To test hypothesis 4b, the sample was split by decision type and the rating task was regressed on the vignette factors including the governance factor. This would allow us to compare the relative importance of governance (notice vs. human-in-the-loop) for types of decisions (pivotal vs. not pivotal). The results of the regression for

the decision type factor are in Table 8. The legitimacy dividend was *lower* for low importance decisions compared to pivotal decisions (health, parole, and even police) thus contradicting hypothesis 4.

All else being equal, the legitimacy benefit from utilizing a human-in-the-loop (rather than mere notice) is greatest for parole, health, and police decisions and lowest for decisions not considered pivotal (i.e. potholes). This suggests that utilizing a human-in-the-loop for highly pivotal decisions has more impact on the perception of the legitimacy of the decision than a mere notice governance regime. In effect, human governance insulated algorithmic decisions of high importance from decreases in perceived legitimacy. We also compared the relative importance of decision type under two conditions: algorithmic decision with mere notice and algorithmic decisions with a human-in-the-loop governance mechanism. Although respondents differentiated between types of decisions for the perceived legitimacy of the decision with mere notice, the type of decision ceased to be statistically significant as a factor in the regression of the perceived legitimacy of algorithmic decisions on the vignette factors when human-in-the-loop was included ($p > 0.03$ for all decisions).

Discussion

Using factorial vignette methodology, this study has shown that the perceived legitimacy of algorithmic decision-making systems varies with decision importance, data source, error rate, and the type of governance regime applied. More specifically, we find evidence for the following statistically significant relationships.

There is an inverse relationship between decision importance and the legitimacy of algorithmic decision-making. Put another way, the perceived legitimacy *decreases* as algorithmic decision-making is applied to decision types deemed more important or pivotal in a person's life. As to outcomes, we found a significant relationship between outcome error rates and legitimacy: as mistake rates increases, perceptions of legitimacy decrease. However, the penalty for more mistakes was constant across the types of decisions and the legitimacy penalty for mistakes is greater for human decisions compared to algorithmic decisions.

We also found that the type of data input impacts the perceived legitimacy. Human decision-makers will often use information gathered from a specific individual to make decisions about government benefits for that individual. A social worker, for example, will make Medicaid allocation decisions based on data they personally gathered from a visit to the home of an individual seeking in-home care. That type of data source—specific data gathered from the decision target—was perceived as most legitimate. The perceived legitimacy decreased significantly when algorithmic systems made decisions based on general data about an individual gathered from the Internet. Interestingly, we found that human-led decisions were perceived as less legitimate than algorithmic decisions when general data was used for the decision. There could be multiple explanations for this. For example, respondents may not believe humans could adequately analyze a larger, more heterogeneous data set as compared to a computer. Given the increasing prevalence of algorithmic decision-making systems, respondents may be experientially primed to accept that some decisions are made by algorithm (Ajunwa, 2020). The responses may also reflect the recognition, well known in the literature, that humans are biased, too (e.g. Ziegert and Hanges, 2005). More research is needed to explain these results.

Finally, we found that governance of algorithmic decision-making matters significantly for legitimacy. Greater perceived legitimacy was associated with human-in-the-loop governance compared to the other form of governance studied—namely, notice. More specifically, human-in-the-loop governance increased the perception of legitimacy of algorithmic decision-making across the board, while notice governance is only sufficient to legitimize nonpivotal decisions. Notably, these results only compare human-in-the-loop governance to notice and cannot speak to the legitimacy penalties or dividends associated with human-in-the-loop governance relative to other forms of procedural or substantive safeguards. That said, when human-in-the-loop governance was included, respondents did not differentiate between types of decisions (pivotal vs. nonpivotal) when judging the legitimacy of the decision.

Implications to theory. These results have important implications for scholars, policymakers, and developers of algorithmic systems. First, following Tyler's (1990/2006) process theory of legitimacy, the results suggest that government actors using algorithmic decision-making tools may be able to make up for some of the legitimacy penalties of mistakes with governance regimes. Indeed, we find that human-in-the-loop governance has the greatest legitimacy dividend for algorithmic decision-making across a range of decision types. That conclusion, however, should not be confused with the suggestion that human-in-the-loop governance is the best governance regime for algorithmic decision-making generally. It was simply the most

substantial compared to the other two regimes studied, which included no governance and mere notice, both of which are actually quite typical. The algorithmic accountability literature includes myriad governance proposals (Citron, 2007; Jones, 2017; Kaminski, 2019a, 2019b; Katyal, 2019; Madden et al., 2017; Reisman et al., 2018; Selbst and Barocas, 2018). There is reason to believe that human-in-the-loop governance is insufficient (Green, 2021). This study may lend credibility to the notion that more robust governance may increase legitimacy, but the effects of other governance regimes must be studied to determine the optimal governance regime. Nor should popular legitimacy be the only factor to consider when developing governance and accountability tools.

Second, we showed that the legitimacy penalty for mistakes is greater for human decision-making than for algorithmic decision-making. In other words, we punish humans for making mistakes more than we do machines. This result is not surprising. Governance mechanisms are supposed to catch mistakes, violations of rights, and other harms. Therefore, having a human in the loop of an algorithmic decision may put individuals' minds at ease that someone is checking or auditing the results, even if human-in-the-loop governance may cover a wide range of practices (Jones, 2017). Humans also tend to put significant faith in machines to function properly and achieve results, and that faith may be a safety net for algorithmic legitimacy (Hoff and Bashir, 2015). It may also shield algorithmic systems from necessary interrogation. This study may suggest that effective governance is necessary for legitimacy of algorithmic decision-making, but it does not suggest that system governance alone is sufficient to manage algorithmic decision-making generally.

Third, we found a significant relationship between data type and legitimacy. We found a significant, negative impact on perceived legitimacy when decisions relied upon general data about an individual versus specific data regardless of who was making the decision. This implicates our understanding of fairness and algorithms because we may want decisions that affect us to be made based on data that is actually about us rather than about others. This has significant implications for algorithmic processes today and our study of fairness (Barocas et al., 2017; Barocas and Selbst, 2016; Lee, 2018), which take aggregated data categorized by latent characteristics and apply it to make predictions about other populations that share similar traits. Our data suggest that individuals may not accept government making decisions the same way digital advertisers make decisions. There is already evidence of this. The United Kingdom's use of a grading algorithm based on general data sparked protests (Hao, 2020).

Implications for practice. The legitimacy penalty associated with algorithmic decision-making systems based on general data about the individual and those with significant

mistakes, combined with the difficulty of using technical fixes to ameliorate these problems, may raise ethical dilemmas for technology designers. Governments rely on skilled engineers to develop algorithms that they can use to make social policy decisions. Designers may want to become involved not only in the technical aspects of algorithmic systems but also in their governance and deployment in order to ensure that their designs are used ethically and in accordance with their values. This study suggests that governance *and* design matter for public perceptions of the legitimacy of algorithmic legitimacy. If designers would like to develop these systems ethically, their input may be valuable on both metrics.

Limitations and areas for future research

Despite these conclusions, this research has certain limitations. Methodologically, factorial vignettes with too many variables can become too complex for study participants (Auspurg et al., 2014). We attempted to account for this problem by both limiting the variables to keep the study simple and limiting the number of vignettes to which each respondent must respond. That said, our vignettes asked respondents to assess legitimacy based on independent variables in the vignettes. Given the flexibility of that term, responses could reflect variance in respondent perceptions of the word legitimacy. The size of the sample set was intended to ameliorate those concerns.

The factors included in this study's analytical model are also limited. We studied decision importance, data collection, mistakes, and governance on the legitimacy of algorithmic decisions. We did not study, for example, the effect of racial, gender, and other forms of bias on legitimacy. Therefore, we do not know if any governance regime could overcome the expected legitimacy penalty that comes with discriminatory outputs. In other words, outcomes focused on mistakes are less morally laden than outcomes that are discriminatory (Martin, 2019); discriminatory outcomes may be *more* important to the perceived legitimacy than mere mistakes. Nor does this study evaluate perceptions of the legitimacy of *corporate* use of algorithmic decision-making. Considering the sources of legitimacy of corporations differ from that of government (Bitektine and Haack, 2015), the relative importance of input, process, and outcome on perceived legitimacy could also differ. Future research could also assess the effects that different types of algorithms, some of which require more data than others, have on perceptions of legitimacy.

Zeng (2020) found that the legitimacy of the Chinese government's use of algorithmic decision-making systems depends in large part on perceptions of the legitimacy of the government and the Chinese Communist Party. We did not test whether popular perceptions of the legitimacy

of algorithmic decision-making depend on the legitimacy of the particular government authority deploying it.

Finally, this study does not speak to whether the use of algorithmic decision-making at all accords with human values such as equality, nondiscrimination, and dignity, among others. It might be the case, as scholars have argued, that technological decision-making systems elevate neoliberal over social values so significantly that their use should be limited (Cohen, 2019). That is a choice society as a whole, whether through the political, regulatory, or judicial process, must make, but it is not a focus of this study.

Conclusion

This study adds to the sociolegal, ethical, and technological literature on algorithmic accountability in at least three ways. First, the study provides empirical evidence for descriptive and normative arguments about the need for robust human governance in the government use of algorithmic decision-making systems. Relatedly, the study finds that transparency is not a sufficient governance model for algorithmic decision-making, countering arguments for greater transparency as a governance solution. Second, this study reinforces the urgent need to develop governance structures before algorithmic decision-making becomes omnipresent. Perceived legitimacy is of central importance to liberal democracy. Opaque algorithmic systems have the capacity to undermine that legitimacy. Developing a better understanding of the conditions of legitimate algorithmic decision-making, not to mention appreciating when even robust governance may not be enough to cancel out legitimacy penalties, is, therefore, of utmost importance to law, society, and technology scholars.


Declaration of conflicting interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Ari Waldman  <https://orcid.org/0000-0002-9264-9231>

Kirsten Martin  <https://orcid.org/0000-0002-0638-0169>

Notes

1. We focus on government decisions in this study because even though private actors use automated systems to make decisions about loans, healthcare, and housing, government use may be especially problematic when it "distribute[s] resources or

- mete[s] out punishment” (Katyal, 2019). We reserve an analysis of commercial use of algorithmic systems for future study.
- In measuring legitimacy, the authors (Lünich and Kieslich, 2021) used four items (I accept the decision; I agree with the decision; I am satisfied with the decision; I recognize the decision).
 - A few tests allow the researcher to identify whether the respondent “clicked through” including whether the range of responses was small (clustered around -100 , 0 , or $+100$) by analyzing either the “range” of responses or the standard deviation. These respondents were not included in the analysis. Previous studies of respondent quality comparing Mechanical Turk to a (more expensive) nationally representative survey through Knowledge Networks showed a significantly greater proportion of the sample was discarded for clicking through for Knowledge Networks (16%) compared to Mechanical Turk (2%) (Martin and Nissenbaum 2020).
 - Turk has been used for theoretical generalizability quite successfully, as in the examination of the relationship between concepts or ideas (Kang, Brown, Dabbish, and Kiesler, 2014; Martin and Nissenbaum, 2017; Redmiles, et al., 2017). In critiques of Turk samples, the Turk results are compared to phone surveys (Kang et al., 2014) as well as online nationally representative samples; the critiques focus on questions of statistical generalizability (Kang et al., 2014; Sharpe Wessling et al., 2017). We offer a theoretical examination, where the findings will support or not support the hypothesized relationships between vignette factors. Such research seeks the generalizability of ideas rather than the generalizability of data patterns within a specific population (Lynch, 1982). Our results focus on theoretical generalizability, for example, whether cause-effect relationships hold or whether concepts are related (Lynch, 1982).

References

- Ajunwa I (2020) The paradox of automation as anti-bias intervention. *Cardozo Law Review* 41: 1671–1744.
- Angwin J, Larson J, Mattu S, et al. (2016) Machine Bias. *ProPublica*. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Auspurg K, Hinz T, Liebig S, et al. (2014) The factorial survey as a method for measuring sensitive issues. In: Engel U, Jann B, Lynn P, Scherpenzeel A and Sturgis P (eds) *Improving Survey Methods: Lessons from Recent Research*. New York: Taylor & Francis, pp. 137–149.
- Badas A (2019) Policy disagreement and judicial legitimacy: Evidence from the 1937 court-packing plan. *Journal of Legal Studies* 48(2): 377–408.
- Barocas S, Hardt M and Narayanan A (2017) Fairness in machine learning. *Nips tutorial* 1: 2.
- Barocas S and Selbst AD (2016) Big data’s disparate impact. *California Law Review* 104.
- Bartels B and Johnston C (2013) On the ideological foundations of supreme court legitimacy in the American public. *American Journal of Political Science* 57: 184–199.
- Benjamin R (2019) *Race After Technology*. Cambridge, UK: Polity.
- Berman E (2018) A government of laws and not machines. *Boston University Law Review* 98: 1277–1355.
- Bitektine A and Haack P (2015) The “macro” and the “micro” of legitimacy: Toward a multilevel theory of the legitimacy process. *Academy of Management Review* 40(1): 49–75.
- Bohman J (1998) Survey article: The coming of age of deliberative democracy. *Journal of Political Philosophy* 6(4): 400–425.
- Bridges K (2017) *The Poverty of Privacy Rights*. Palo Alto, CA: Stanford University Press.
- Brummette J and Zoch L (2016) How Stakeholders’ personal values influence their value expectations for legitimate organizations. *Corporate Communications: An International Journal* 21: 309–321.
- Burrell J (2016) How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1). <https://doi.org/10.1177/2053951715622512>
- Caliskan A, Bryson J and Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. *Science* 356: 183–186.
- Calo R (2017) Artificial intelligence policy: A primer and roadmap. *University of California, Davis Law Review* 51: 399–435.
- Christenson D and Glick D (2015) Chief Justice Roberts’s health care decision disrobed: The microfoundations of the supreme court’s legitimacy. *American Journal of Political Science* 59: 403–418.
- Citron D (2007) Technological due process. *Washington University Law Review* 85(6): 1249–1313.
- Citron D and Pasquale F (2014) The scored society: Due process for automated predictions. *Washington Law Review* 89(1): 1–33.
- Cohen J (2019) *Between Truth and Power*. New York: Oxford University Press.
- Colaner N (2021) Is explainable artificial intelligence intrinsically valuable? *AI & Society* 37(1): 231–238.
- Coppock A (2018) Generalizing from Survey Experiments Conducted on Mechanical Turk: A replication approach. *Political Science Research and Methods* 7(3): 613–628
- Cormen T (2009) *Introduction to Algorithms*. Cambridge, MA: MIT Press.
- Crawford K and Schultz J (2014) Big data and due process: toward a framework to redress predictive privacy Harms. *Boston College Law Review* 55(1): 93–128.
- Daly T and Natarajan R (2015) Swapping bricks for clicks: Crowdsourcing longitudinal data on Amazon Turk. *Journal of Business Research* 68(12): 2603–2609.
- Danaher J, et al. (2017) Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big Data & Society* 4(2): 1–21. DOI: 10.1177/2053951717726554.
- de Fine Licht K and de Fine Licht J (2020) Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy. *AI & Society* 35: 917–926.
- de Laat P (2019) The disciplinary power of predictive algorithms: A Foucauldian perspective. *Ethics and Information Technology* 21: 319–329.
- de Vries P and Midden C (2008) Effect of indirect information on system trust and control allocation. *Behavior and Information Technology* 27(1): 17–29.
- Dressel J and Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4(1): 1–5.
- Dworkin R (1996) *Freedom’s Law*. Cambridge, MA: Harvard University Press.

- Easton D (1965) *A Systems Analysis of Political Life*. New York: Wiley.
- Eubanks V (2018) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: Saint Martin's Press.
- Federal Trade Commission (2016) *Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues*, <https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf>
- Figueras C, Verhagen H and Pargman TC (2021) Trustworthy AI for the People? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*: 269–270.
- Gibson J and Caldeira G (2009) *Citizens, Courts, and Confirmation: Positivity Theory and the Judgments of the American People*. Princeton, NJ: Princeton University Press.
- Gibson J, Caldeira G and Spence LK (2003) Measuring attitudes toward the United States supreme court. *American Journal of Political Science* 47: 354–367.
- Green B (2021) The Contestation of Tech Ethics: A Sociotechnical Approach to Ethics and Technology in Action. arXiv preprint arXiv:2106.01784.
- Grimes M (2006) Organizing consent: The role of procedural fairness in political trust and compliance. *European Journal of Political Research* 45: 285–315.
- Hagan J, Ferrales G and Jasso G (2008) How law rules: Torture, terror, and the normative judgments of Iraqi judges. *Law & Society Review* 42(3): 605–644.
- Hao K (2020) The UK Exam Debacle Reminds Us that Algorithms Can't Fix Broken Systems. *MIT Technology Review*, <https://www.technologyreview.com/2020/08/20/1007502/uk-exam-algorithm-cant-fix-broken-system/>
- Henderson S (2018) A few criminal justice big data rules. *Ohio State Journal of Criminal Law* 15: 527–541.
- Hoff KA and Bashir M (2015) Trust in automation: integrating empirical evidence on factors that influence trust *Human Factors* 57(3): 407–434.
- Houston Federation of Teachers, Local 2415 v. Houston Independent School District, 251 F. Supp. 3d 1168 (S.D. Tex. 2017).
- Hu M (2016) Big data blacklisting. *Florida Law Review* 67(5): 1735–1809.
- Innerarity D (2021) Making the black box society transparent. *AI & Society* 36(3): 975–981.
- Jacovi A, Marasović A, Miller T, et al. (2021) Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 624–635.
- Jasso G (2007) *Studying justice: Measurement, estimation, and analysis of the actual reward and the just reward*. IZA Discussion Papers.
- Jasso G (2006) Factorial survey methods for studying beliefs and judgments. *Sociological Methods & Research* 34(3): 334–423.
- Jones M (2017) The right to a human in the loop: political constructions of computer automation and personhood. *Social Studies of Science* 47: 216–239.
- Joseph G and Lipp K (2018) IBM Used NYPD Surveillance Footage to Develop Technology That Lets Police Search by Skin Color. *The Intercept*, <https://theintercept.com/2018/09/06/nypd-surveillance-camera-skin-tone-search/>
- Jung J, Concannon C, Shroff R, et al. (2017) Simple Rules for Complex Decisions. Working Paper, Stanford University, <https://arxiv.org/pdf/1702.04690.pdf>. Accessed December 4, 2019.
- K.W. v. Armstrong, 180 F. Supp. 3d 703 (D. Idaho 2016).
- Kaina V (2008) Legitimacy, trust and procedural fairness: remarks on Marcia Grimes' study. *European Journal of Political Research* 47(4): 510–521.
- Kaminski M (2019a) Binary governance: Lessons from the GDPR's approach to algorithmic accountability. *Southern California Law Review* 92(6): 1529–1616.
- Kaminski M (2019b) Right to explanation, explained. *Berkeley Technology Law Journal* 34: 189–218.
- Kang R, Brown S and Dabbish L, et al. (2014) *Privacy Attitudes of Mechanical Turk Workers and the US Public*. In: SOUPS, 2014, pp. 37–49.
- Katyal S (2019) Private accountability in the age of artificial intelligence. *University of California, Los Angeles Law Review* 66: 54–141.
- König PD, Wurster S and Siewert MB (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): 1–13.
- Kraemer F, van Overveld K and Peterson M (2011) Is there an ethics of algorithms? *Ethics and Information Technology* 13(3): 251–260.
- Lavorgna A and Ugwudike P (2021) The datafication revolution in criminal justice: An empirical exploration of frames portraying data-driven technologies for crime prevention and control. *Big Data & Society* 8(2): 20539517211049670.
- Lee M (2018) Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5(1): 1–16.
- Leicht-Deobald U, Busch T, Schank C, et al. (2019) The challenges of algorithm-based HR decision-making for personal integrity. *Journal of Business Ethics* 160(2): 377–392.
- Lipset SM (1959) Some social requisites of democracy: Economic development and political legitimacy. *American Political Science Review* 53: 69–105.
- Loi M, Ferrario A and Vigano E (2020) Transparency as design publicity: Explaining and justifying inscrutable algorithms. *Ethics and Information Technology* 23(3): 253–263. <https://doi.org/10.1007/s10676-020-09564-w>
- Lünich M and Kieslich K (2021) Using automated decision-making (ADM) to allocate COVID-19 vaccinations? Exploring the roles of trust and social group preference on the legitimacy of ADM vs. Human decision-making. *arXiv preprint arXiv:2107.08946*: 1–17.
- Lynch Jr JG (1982) On the external validity of experiments in consumer research. *Journal of consumer Research* 9(3): 225–239.
- Madden M, Gilman M, Levy K, et al. (2017) Privacy, poverty, and big data: A matrix of vulnerabilities for poor Americans. *Washington University Law Review* 95: 53–125.
- Martin K (2018) Ethical implications and accountability of algorithms. *Journal of Business Ethics* 160: 1–16.
- Martin K (2019) Designing ethical algorithms. *MIS Quarterly Executive* 18(2): 129–142.
- Martin K and Nissenbaum H (2017) Privacy interests in public records: An empirical investigation. *Harvard Journal of Law and Technology* 31(1): 111–143.

- Martin K and Nissenbaum H (2020) What is it about location? *Berkeley Technology Law Journal* 35(1): 251–309.
- Metcalf J, Moss E, Watkins EA, et al. (2021) Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*: 735–746.
- Noble S (2018) *Algorithms of Oppression*. New York: N.Y.U. Press.
- O’Neil C (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Broadway Books.
- Ohm P (2010) Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review* 57: 1701–1777.
- Pasquale F (2014) *Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.
- Pasquale F (2019) Data-informed duties in AI development. *Columbia Law Review* 119(7): 1917–1940.
- Persson M, Esaiasson P and Gilljam M (2013) The effects of direct voting and deliberation on legitimacy beliefs: An experimental study of small group decision-making. *European Political Science Review* 5(3): 381–399.
- Pirson M, Martin K and Parmar B (2019) Public trust in business and its determinants. *Business & Society* 58(1): 132–166. DOI: 10.1177/0007650316647950
- Pirson M, Martin K and Parmar B (2017) Formation of stakeholder trust in business and the role of personal values. *Journal of Business Ethics* 145(1): 1–20.
- Rahwan I (2018) Society-in-the-Loop: programming the algorithmic social contract. *Ethics and Information Technology* 21: 5–14.
- Redmiles EM, Kross S, Pradhan A, et al. (2017) How well do my results generalize? Comparing security and privacy survey results from MTurk and web panels to the US.
- Reisman D, Schultz J, Crawford K, et al. (2018) “Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability.” AI Now Institute. <https://ainowinstitute.org/aiareport2018.pdf>
- Salganik MJ, et al. (2020) Measuring the predictability of life outcomes with a scientific mass collaboration. *Proc Natl Acad Sci U S A* 117(15): 8398–8403.
- Selbst A and Barocas S (2018) The intuitive appeal of explainable machines. *Fordham Law Review* 87: 1085–1139.
- Sharpe Wessling K, Huber J and Netzer O (2017) MTurk character misrepresentation: Assessment and solutions. *Journal of Consumer Research* 44(1): 211–230.
- Sheehy B (2019) Algorithmic paranoia: The temporal governmentality of predictive policing. *Ethics and Information Technology* 21: 49–58.
- Skarlicki DP and Folger R (1997) Retaliation in the workplace: The roles of distributive, procedural, and interactional justice. *Journal of Applied Psychology* 82: 434–443.
- Sonnad N (2018) US Border Agents Hacked Their ‘Risk Assessment’ System to Recommend Detention 100% of the Time. *Quartz*, <https://qz.com/1314749/us-border-agents-hacked-their-risk-assessment-system-to-recommend-immigrant-detention-every-time/>
- Starke C and Lünich M (2020) Artificial intelligence for political decision-making in the European union: effects on citizens’ perceptions of input, throughput, and output legitimacy. *Data & Policy* 2: 1–17.
- Suchman M (1995) Managing legitimacy: Strategic and institutional approaches. *Academy of Management Review* 20(3): 571–610.
- Sühr T, Hilgard S and Lakkaraju H (2021) Does fair ranking improve minority outcomes? understanding the interplay of human and algorithmic biases in online hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*: 989–999.
- Sunshine J and Tyler T (2003) The role of procedural justice and legitimacy in shaping public support for policing. *Law & Society Review* 37: 513–547.
- Tucker C (2014) The reach and persuasiveness of viral video ads. *Marketing Science* 34(2): 281–296.
- Tufekci Z (2015) Algorithmic Harms beyond Facebook and google: emergent challenges of computational agency. *Colorado Journal of Telecommunications and High Technology* 13: 203–218.
- Tyler T (1994) Governing amid diversity: the effect of fair decision-making procedures on the legitimacy of government. *Law & Society Review* 28(3): 809–831.
- Tyler T (2006/1990) *Why People Obey the Law*. Princeton, NJ: Princeton University Press.
- Tyler T and Huo YJ (2002) *Trust in the Law: Encouraging Public Cooperation with the Police and Courts*. New York: Russell Sage Foundation.
- Veale M and Binns R (2017) Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society* 4(2), SAGE Publications Sage UK: London, England, pp. 1–17.
- Wallander L (2009) 25 Years of factorial surveys in sociology: A review. *Social Science Research* 38(3): 505–520.
- Weber M (1947/2012) *The Theory of Social and Economic Organizations*. Eastford, CT: Martino Fine Books.
- Wexler R (2018) Life, liberty, and trade secrets: Intellectual property in the criminal justice system. *Stanford Law Review* 70: 1343–1429.
- Whittaker M, et al. (2018) *AI Now Report 2018*. AI Now Institute. https://ainowinstitute.org/AI_Now_2018_Report.pdf
- Zarsky T (2013) Transparency in data mining: from theory to practice. In: Custers B (eds) *Discrimination and Privacy in the Information Society*. Vering Berlin Heidelberg: Springer, pp. 301–324.
- Zeng J (2020) Artificial intelligence and China’s authoritarian governance. *International Affairs* 96(6): 1441–1459.
- Ziegert J and Hanges P (2005) Employment discrimination: the role of implicit attitudes, motivation, and a climate for racial bias. *Journal of Applied Psychology* 90(3): 553–562.
- Zwitter A (2014) Big data ethics. *Big Data & Society* 1(2): 1–6.

Appendix A

In general, the vignettes had the following format. Each factor is underlined and the randomly generated level for each factor was created independently and with replacement each time the vignette was created for the respondent. For the human decisions, which were used for comparison, the phrase “a computer program” was replaced with “a group of individuals.” Respondents were assigned either a series of human decisions or automated decisions.

Template:

In order to decide {Decisions Social Goods_alt}, a computer program uses information that {Decisions Social Goods_alt3} {Data Type_alt} to identify {Decisions Social Goods_alt2}.

Upon review, approximately {Mistakes_alt} of the decisions were incorrect—{Decisions Social Goods_alt4} {Mistakes_alt} of the time.

The organization {Governance Factor_alt}

Examples:

In order to decide the services someone receives under medicare or medicaid (e.g., in home health aid, wheelchair, physical therapy, etc), a computer program uses information that the state's board for human services gathered specifically for this purpose to identify the patient's need for assistance.

Upon review, approximately 10% of the decisions were incorrect—where patients were incorrectly denied care 10% of the time.

The organization hires a privacy professional to oversee the decision.

++++

In order to decide which potholes are fixed, a computer program uses information that the city government gathered from individuals' online browsing behavior to identify the worst streets in the city.

Upon review, approximately 50% of the decisions were incorrect—where the wrong potholes got fixed 50% of the time.

The organization hires a privacy professional to oversee the decision.

+++++

In order to decide who gets released early from prison, a computer program uses information that the parole board gathered specifically for this purpose to identify the estimated risk of recidivism.

Upon review, approximately 40% of the decisions were incorrect—where the wrong prisoner remains in prison 40% of the time.

The organization notifies citizens that a computer program makes the decision.

++++