When Are Algorithmic Decisions Perceived as Legitimate?

The Effect of Process and Outcomes on Perceptions of Legitimacy of Algorithmic Decisions

Kirsten Martin & Ari Waldman

Forthcoming Journal of Business Ethics

Abstract

Firms use algorithms to make important business decisions. To date, the algorithmic accountability literature has elided a fundamentally empirical question important to business ethics and management: Under what circumstances, if any, are algorithmic decision-making systems considered legitimate? The present study begins to answer this question. Using factorial vignette survey methodology, we explore the impact of decision importance, governance, outcomes, and data inputs on perceptions of the legitimacy of algorithmic decisions made by firms. We find that many of the procedural governance mechanisms in practice today, such as notices and impact statements, do not lead to algorithmic decisions being perceived as more legitimate in general, and, consistent with legitimacy theory, that algorithmic decisions with good outcomes are perceived as more legitimate than bad outcomes. Yet, robust governance, such as offering an appeal process, can create a legitimacy dividend for decisions with bad outcomes. However, when arbitrary or morally dubious factors are used to make decisions, most legitimacy dividends are erased. In other words, companies cannot overcome the legitimacy penalty of using arbitrary or morally dubious factors, such as race or the day of the week, with a good outcome or an appeal process for individuals. These findings add new perspectives to both the literature on legitimacy and policy discussions on algorithmic decision-making in firms.
INTRODUCTION

There is a new challenge to firm legitimacy—namely, the increasing use of algorithms and computers to make decisions that directly impact people’s lives (Citron and Pasquale 2014). Businesses use algorithmic decision-making (ADM) systems to make hiring, firing, and promotion decisions (Ajunwa 2020). Banks assess loan eligibility and credit risk algorithmically (Citron and Pasquale 2014). Social media platforms rely on ADM to moderate and curate content (Vincent 2020; Gillespie 2019). Recommendation algorithms nudge us toward products, news, entertainment, and information (Zuboff 2019).

And yet, despite the increasingly frequency with which management decisions are being made algorithmically, the public is skeptical. Roughly six in ten are concerned that ADM will be biased and unfair; two-thirds think algorithmic financial scoring and hiring decisions are unacceptable (Smith 2018). Studies show that people are largely concerned about privacy risks, biases, fairness, and usefulness of ADM (Araujo et al. 2020). They think ADM systems are less “authentic” and, therefore, less ethical than identical human decisions (Jago 2019). And Newman, Fast, and Harman (2020) have shown that individuals perceive ADM used by human resources departments as reductionist and unfair. This suggests that ADM may materially impact corporate legitimacy because a firm’s legitimacy is a product of and positively correlated with its actions (Chung et al. 2014; Santana 2012).¹ Within business ethics, the perceived legitimacy of stakeholders – consumers, users, employees, etc. – are important to firms as key audiences to

¹ We recognize, following Easley and Lenox (2006), that decision legitimacy merits separate evaluation and study. And, the link between decision legitimacy and overall firm legitimacy is of critical importance to business and business ethics scholars.
their actions, and legitimacy is seen as a fundamental concept in business ethics (Chen et al. 2020). In their reassessment of legitimacy as a stakeholder attribute, Ali (2017: 157) has called identifying stakeholder perceptions of legitimacy one of three “pivotal” questions for legitimacy. In other words, understanding what is or is not legitimate is a “process that involves perceptions” (p. 160) because stakeholder and firm legitimacy is “based on the perceptions of the society, stakeholders, and management” (p. 164).

Given the risks that ADM systems pose to corporate legitimacy, this study seeks to determine whether and under what circumstances, if any, the use of ADM to make commercial decisions impacts perceived legitimacy of the decision. To be clear, this study does not seek to normatively defend the practice of using ADM to making social, economic, or policy decisions about people. Nor does it suggest that adoption of any particular form of governance makes the use of ADM legitimate in a democratic society. Rather, we build on the empirical work of Elsbach (1994), Finch et al. (2015), and Jahn et al. (2020) in that we seek to empirically measure perceptions of legitimacy of commercial actors. We do so in a novel context of ADM and also add to the literature by considering the relative importance of factors in determining legitimacy. We use the terms legitimacy dividend and legitimacy penalty to describe the positive and negative effects that the presence or absence of a condition has on perception of the legitimacy of an algorithmic decision.

Using factorial vignette survey methodology to survey individuals’ normative judgments about algorithmic decisions, we ran nine surveys to measure the relative importance of governance, outcomes, and inputs on perceived ADM decision legitimacy. We find that although decision
importance is negatively associated with perceptions of legitimacy, and outcomes are positively associated with perceptions of legitimacy, only particularly robust procedural governance increases perceptions of legitimacy when algorithmic decisions lead to negative outcomes or outcomes with which individuals disagree. That said, neither positive outcomes nor procedural governance can correct for the significant legitimacy penalties associated with ADM systems that use arbitrary and morally dubious factors such as race. Given that predictive ADM systems reify biases in society, our results have significant impact on firm decisions and public policy.

We contribute to legitimacy theory by explicitly testing the perceived legitimacy of ADM. Extending this type of analysis to algorithmic and computer-driven decisions, considering the degree to which the public is aware and skeptical of such decisions (Araujo et al. 2020), we extend the scope of legitimacy scholarship. First, we build on work by Elsbach (1994), Finch et al. (2015), and Jahn et al. (2020), which has explored positive drivers of decision legitimacy by demonstrating factors’ relative importance. Second, we contribute to the business ethics scholarship on legitimacy by considering it through the lens of independent variables—decision importance, procedural governance, and outcomes—that have not been studied in the business ethics literature (Tyler 1990; Badas 2019; Gibsen and Caldiera 2009). Finally, we contribute to public policy as regulators and firms struggle to understand how to improve ADM systems for all stakeholders. This research suggests that testing for arbitrary or morally dubious factors is more important for perceptions of legitimacy of the decision and the company than offering a better notice or impact statements, two particularly common recommendations in the literature (Kaminski 2019; Reisman et al. 2018; Katyal 2019). Current law and new regulatory proposals
prioritize procedural rules for ADM that, our study shows, will not increase perceived legitimacy.

THEORETICAL CONTEXT

Algorithmic Decision-Making in Firms

This study focuses on algorithmic-decision making (ADM) systems, which, following Calo (2017), we define generally as processes involving algorithms, or sequences of logical, mathematical operations, to implement policies by software. Some ADM tools are powered by artificial intelligence of varying maturity and types; our study focuses on algorithms developed from training data generally. We also recognize that there exists a range of automation, with some decisions almost fully computerized while others are merely augmented with technology (Martin 2018; Araujo et al. 2020). Defining all ADM systems is not our goal. As this study focuses on perceptions of legitimacy of algorithmic decisions of private, commercial firms, we assume a definition that recognizes the role of data inputs, computers, and automation of decisions.2

Unfortunately, researchers have shown that such algorithmic systems are risky. Because ADM systems make probabilistic predictions about the future, these programs may make mistakes

---

2 In conceptualizing ADM in this way, we follow a long line of scholars studying popular perceptions of artificial intelligence (AI) generally (Araujo et al. 2020; Dodge and Kitchin 2007). This paper investigates the development of an algorithm from training data and the use of that algorithm to make a decision about individuals. We do not study the underlying technique used to develop the algorithm (neural networks, supervised or unsupervised learning, etc). We leave that for future study.
about the ambiguous situations they are trying to predict; probabilities are necessarily
generalized, with individual cases falling through the cracks (Eubanks 2018; Hu 2016). Studies
have also shown that ADM’s predictive capabilities may be exaggerated (Salganik et al. 2020;
Dressel and Farid 2018; Jung et al. 2017). ADM’s opacity makes algorithmic systems difficult to
interrogate and hold accountable (Cheng 2013; Loi, Ferrario, and Vigano 2020). Algorithmic
systems also provide an incentive for surveillance and data collection because they need large
information sets for model training and analysis (Martin 2019). This creates the circumstances
for invasions of privacy and erosion of privacy norms (Zwitter 2014).

Finally, ADM systems are biased (Johnson 2020, n.d.; Winner 1980). They rely on the corpus of
data on which they are based and, therefore, data that is biased along race, gender, sex, and
socioeconomic lines will lead to discriminatory results (Noble 2018; Caliskan et al. 2017; Katyal
2019; O’Neil 2016). Developers make value-laden assumptions about how to treat missing data
and outliers that also impacts the bias of the developed algorithm (Martin 2022).

Recently, scholars have begun investigating not only the harms and discrimination created by
ADM but also how individuals perceive decisions that rely on algorithms. This work has paralleled
existing theory on the perceptions of fairness, trust, and legitimacy of firms and decisions.
Legitimacy

At the organizational level, entities are perceived to be legitimate when their actions are “are desirable, proper, or appropriate within some socially constructed system of norms, values, beliefs, and definitions” (Suchman 1995: 574). Legitimate firms pursue “socially acceptable goals in a socially acceptable manner” (Ashforth and Gibbs 1990: 177). Legitimacy is an “attitude” (Jahn et al. 2020: 546) or “perception” (Suddaby et al. 2017: 451) that is closely related to institutional credibility.

Firms seek legitimacy for ethical and strategic reasons in a manner that is similar to those seeking to be trustworthy and fair. For most strategy and business ethicists, being fair, trustworthy, and legitimate is both strategically sound as well as a moral imperative and an end in itself (Freeman et al. 2020; Phillips 1997; Pirson et al. 2017). The normative argument for firms caring about legitimacy parallels the argument for why firms should care about being perceived as trustworthy and fair. Firms should seek to understand the concerns of their stakeholders – employees, users, customers, communities – based on obligations of care (Wicks et al. 1994), fairness (Phillips 1997), and an obligation to create value for stakeholders (Freeman et al. 2020). As an actor within a larger industry or marketplace, firms have an obligation to be perceived as legitimate and act in a manner that is “desirable, proper, or appropriate within some socially constructed system of norms, values, beliefs, and definitions” (Suchman 1995: 574) because a firm’s legitimacy impacts the perceived legitimacy of the larger industry or

---

3 The phrase “legitimacy of the firm” references the perceived legitimacy by others as we build on legitimacy as an “attitude” (Jahn et al. 2020: 546) or “perception” (Suddaby et al. 2017: 451). This is parallel to referring to the trustworthiness of a firm when trustworthiness is defined as the belief of others as to the trust factors of the firm (Pirson et al. 2017)
marketplace. In other words, a firm acting in a manner that is not “desirable, proper, or appropriate” within the system of values has harmful spillover effects on similarly situated firms and markets, which then harms others.

In addition, firms should care about whether their actions are perceived as legitimate for the survival of their own firm since stakeholders, including employees, suppliers, and customers, among others, rely on the firm to survive and thrive. A firm that is perceived as more credible and, thus, perceived as a more legitimate company has greater customer loyalty, political leeway, is perceived to make decisions for appropriate reasons, and enjoys voluntary acceptance and compliance (Finch et al. 2015; Deephouse et al. 2017: 32; Chung et al. 2014; Suchman 1995).

When a firm is perceived to be legitimate, it can generate public support and loyalty (Boyd 2000; Coombs 1992). In addition, organizations with high perceptions of legitimacy have more flexibility in decision-making and less public and political interference. In other words, if an organization wants to maintain customer loyalty, stability within a regulatory environment, and achieve market success, it needs to be perceived as legitimate, both with respect to the organization as a whole and its individual decisions (Suchman 1995). Those firms that make illegitimate decisions are considered illegitimate themselves, and those firms cannot survive in the market (Ruef and Scott 1998). Therefore, businesses have powerful ethical, economic,

4 While these two arguments – an overtly normative or ethical argument in one paragraph is separated from the more ‘strategic’ argument in a following paragraph, the authors do not see these as being so separate (Freeman 1994; Freeman et al. 2007). However, we would like to thank a reviewer for reminding us that some readers need to see the argument made separately. We see acting ethically as good for the firm and that actions that create value for stakeholders also create value for the firm.

5 Tom Tyler (2006) and colleagues (Tyler and Huo, 2002: 102) define legitimacy more narrowly, as “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.” Both Suchman and Tyler recognize that there is a moral valence to legitimacy.
political, and social incentives to ensure that both their individual decisions and entire organizations are perceived by the public to be legitimate.

A number of factors are known to influence perceptions of organizational legitimacy. Organizational legitimacy has been positively associated with conformity to norms (Suddaby et al. 2017), negatively correlated with extrinsic or self-interested motives (Jahn et al. 2020), and dependent upon transparency and communication (Elsbach 1994). Corporate social responsibility actions and rhetoric impact perceptions of legitimacy of firms (Bronn and Vidaver-Cohen 2009; Du and Vieira 2012; Castello and Lozano 2011), as do transparent executive responses to crises (Beelitz and Merkl-Davies 2012).

Although there are different types of legitimacy in the organizational literature (Suddaby et al. 2017; Palazzo and Scherer 2006), this study is based on the theory that individual firm decisions impact organizational legitimacy—namely, the perception that a firm is operating according to shared norms, values, and rules (Chung et al. 2014). That perception is based, at least in part, on what Boyd (2000) and Coombs (1992) call issue legitimacy—namely, the perception that a particular corporate activity is appropriate, understandable, and done for the right reasons.

Within this line of scholarship, organizational legitimacy is driven by the perceived legitimacy of an organization’s specific claims, decisions, and behavior (Santana 2012; Easley and Lenox 2006; Dowling and Pfeffer 1975). To the extent that a firm uses ADM to make salient decisions about its stakeholders, the legitimacy of those decisions will have a significant impact on the legitimacy of the organization as a whole. This parallels findings that the perceived fairness of an
algorithmic decisions impacts someone organizational commitment to the firm (Newman et al. 2020).

**Decision Legitimacy**

At the level of individual decisions, motives are also important drivers of perceptions of legitimacy in the legal studies literature on legitimacy, which mostly focuses on popular perceptions of legitimacy of institutions like the Supreme Court or the police. For instance, Badas (2016) showed that when a court hands down a decision with which one disagrees, individuals think the decision was motivated by extrinsic factors like politics and ideology and is, therefore, illegitimate. Badas (2019), Bartels and Johnson (2013), and Christenson and Glick (2015) also suggest that policy disagreements and ideology influence perceptions of the legitimacy of a given court decision and the institution of the judiciary as a whole; when a court makes a decision with which individuals disagree or negatively affects those individuals, they tend to see the court as less legitimate. At the same time, Caldiera and Gibson (2009) and Gibson, Caldiera, and Spence (2005) have shown that repeated exposure to symbols of authority create a positivity bias that insulates institutions like the Supreme Court from legitimacy penalties.

And in canonical studies of legal legitimacy, Tyler (1990) and Sunshine and Tyler (2003) showed that popular perceptions of legitimacy 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 overcome 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).

Similar findings have been made in organizational studies. Elsbach (1994) used a within-subject experimental design to test the cattle’s industry’s legitimizing strategies, finding that the public expects companies to communicate about how they are conforming to accepted norms and values. Finch et al. (2015) examined perceptions of the legitimacy of the environmentally unfriendly oil sands industry in California, finding that participants who favor environmental values judged the industry as less legitimate than those who supported economic development and outcomes. Like similar studies in the legal literature described above, these studies suggest that process—in this case, notice—and outcomes may be associated with popular perceptions with legitimacy.

Both perceptions of fairness and legitimacy have been applied to ADM. Scholars have investigated respondent-level attributes and decision factors that contribute to perceptions of fairness with ADM systems (Araujo et al. 2020; Nagtegaal 2021; Newman et al. 2020). However, according to Kaina (2008), legitimacy is a distinct concept from fairness and trust. Within work on judgements about algorithmic decisions, nascent work on legitimacy of algorithmic decisions has suggested that governance is important to the legitimacy of ADM. For example, Danaher et al. (2017) examine the legitimacy of ADM governance mechanisms (rather than the legitimacy of the decision), and Lünich and Kieslich (2021) studied attributes of respondents (i.e. general trust) in their perception of legitimacy of a vaccine distribution
algorithm. Starke and Lünich (2020) empirically examine input, process, and output legitimacy of EU decisions. What Starke and Lünich (2020) did not do, and what we attempt to do here, is empirically assess the relative importance of inputs, process, and outcomes to the perceived legitimacy of algorithmic decisions.

**HYPOTHESES**

We next develop hypotheses as to the drivers of perceived legitimacy of algorithmic decisions. Based on legitimacy and algorithmic decision scholarship, we expect the type of decision, inputs, process, and outcomes to impact the perceived legitimacy of algorithmic decisions.

*Decision Type.* Social scientists studying algorithms argue that the type of decision being made algorithmically may drive norms around appropriate governance. Where Nagtegaal (2021) finds that the complexity of the decision impacts the perceptions of fair ADM, others have theorized that the degree the decision is important in the life of the individual will drive perceptions of algorithmic decisions. For example, Tufekci (2015) highlights the importance of governing “gatekeeping” algorithms; Calo (2017), Burell (2016), O’Neil (2015), and Martin (2019) argue that “consequential” or “pivotal” decisions should be given the most attention. In other words, scholars have theorized that individuals’ perception of the legitimacy of commercial use of ADM varies with their assessment of the decision’s importance in their lives. This leads to our first hypothesis:

\[ H1: \text{As decision importance increases, individuals’ perception of the legitimacy of using an algorithm to make the decision decreases.} \]
In other words, we predict a larger legitimacy penalty for algorithmic hiring and firing decisions or access to health care decisions than for decisions about music playlists or social media curation.

*Outcomes.* Sociolegal scholarship on legitimacy also suggests that decision outcomes matter for popular perceptions of the legitimacy of the institutions and processes led to those outcomes. For example, Badas (2019) theorizes that disagreement with the outcome of institutional decisions negatively impact perceptions of the legitimacy of those institutions, and Gibson and Caldiera’s (2009) “positivity theory” posits that agreement with the outcome an institutional decision contributes to a larger legitimacy dividend than the legitimacy penalty associated with disagreement with outcomes.

\[ H2: \text{With all other factors held constant, a good, positive outcome is associated with a legitimacy dividend, or an increase in the perceived legitimacy of the decision.} \]

*Arbitrary Rationales.* As the algorithmic accountability literature makes clear (O’Neil 2016, Noble 2018), predictive algorithms based on discriminatory data and discriminatory modeling produce unjust results. For justice scholars, the idea of using arbitrary reasons or basing decisions on factors outside the context of the decision would render the decision unjust (Nozick 1974; Walzer 2008). For algorithmic decisions, using race-based factors in ADM contributes to illegal, and seemingly not legitimate, decisions (Barocas and Selbst 2016). And theoretically, using race as a factor for decisions that have nothing to do with race can delegitimize decisions in law (Ellis and Diamond 2003). The extent to which the use of arbitrary factors delegitimizes ADM has been theorized but has not been empirically tested. This leads to our third hypothesis:
H3: The use of arbitrary or race-based factors has a negative impact on perceived legitimacy.

Governance. One of the most dominant theories in legitimacy studies suggests that people perceive even adverse decisions as legitimate as long as fair procedural governance mechanisms are in place (Tyler 1990). Danaher et al. (2017) also focuses on the legitimacy of the governance mechanism of algorithmic decisions. This is the approach taken by many current and proposed data protection laws, which require organizations to communicate to customers the “logic behind” its algorithms and/or complete impact assessments (Kaminski 2019). Indeed, current policy debates center not on whether ADM requires some form of accountability mechanism, but on what that governance regime should look like (Pasquale 2019). The theoretical argument within legal legitimacy scholarship is that robust governance – impact assessments (Kamiski 2019b), audit trails, detailed explanations, publicly accessible code, systems testing (Citron 2007: 1305-1313), human-in-the-loop (Rahwan 2018; Froomkin, Kerr, and Pineau 2019; Jones 2017), impact assessments (Reisman et al. 2018, Katyal 2019), contestability (Crawford et al. 2021), and codes of conduct, impact assessments, and whistleblower protections (Katyal 2019) – will legitimize decisions regardless of the outcome. Within ethics scholarship, the theorizing around governance and legitimacy is mixed. For example, de Fine Licht and de Fine Licht (2020) theorize that providing justifications for algorithmic decisions could provide grounds for perceived legitimacy, whereas, Leicht-Deobald et al. (2019) suggest that algorithm-based decisions may be perceived as more legitimate because it is difficult for individuals to question a complex, computerized system.
In other words, the literature suggests that perceptions of legitimacy should increase as the robustness of procedural governance mechanisms increases. However, endogeneity theory (Edelman 2016) suggests that many legal rules are developed through an endogenous process involving regulated entities themselves. If individuals perceive those rules unfair, unjust, or the product of corporate self-interest (Jahn et al. 2020), procedural governance may not be able to provide the legitimacy dividends that corporations hope. This leads to two additional hypotheses:

**H4**: Any governance regime brings a legitimacy dividend to algorithmic decision-making relative to no governance, but there is a larger legitimacy dividend as procedural governance becomes more robust.

**H4b**: Robust procedural governance mechanisms moderate the legitimacy penalties associated with bad outcomes.

Scholars have also assumed that good governance through procedural guardrails could legitimize decisions made on arbitrary or morally dubious ground (Rahwan 2018; Froomkin, Kerr, and Pineau 2019; Jones 2017; Kaminski 2019). However, ongoing concerns about algorithms’ racial biases may call that into question. Using racist or arbitrary factors may be so delegitimizing that no amount of procedure could fix it. This leads to our final hypothesis:

**H4c**: There is a legitimacy penalty associated with algorithms that use arbitrary or race-based factors, regardless of the form of procedural governance used.
Figure 1: Figure of Hypotheses
STUDY DESIGN

We used the factorial vignette survey methodology to explore the relative importance of governance, procedure, and outcomes on the perceived legitimacy of decisions based on a firm’s ADM system. A factorial vignette survey presents 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). The general narrative remains consistent across all vignettes with theoretically important factors (based on the hypotheses) randomly generated with replacement. The vignette factors (explained in Tables 1-4 below) are derived from scholarship and systematically varied in the vignette, thereby offering a mechanism to test the theoretical relationships in the hypotheses (Dickel and Graeff, 2018).

This methodology allows researchers to measure the relative importance of the factors of the vignettes affect participants’ attitudes, judgments, or views. The results are theoretically generalizable; the vignettes combine the benefits of laboratory experiments with those associated with experiments in the field (Oll et al. 2018). As noted by Dickel and Graeff (2018), factorial vignette surveys provide a more realistic setting and triggers more honest answers than typical survey instruments (see also Weinberg et al 2014; Auspurk, Hinz, Sauer, and Liebig, 2014).

In this study, decision factors were independently varied with replacement and the respondents judged the degree to which the decision described was legitimate. Each respondent was

---

6 This design avoids two types of weaknesses 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.
presented with 30 short vignettes describing an algorithmic decision made by a firm. In general, the vignettes’ narrative had four elements: a decision-maker, a decision type, the outcome of the decision, and procedural governance associated with the decision-making process, if any. 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.

The use of the single rating task in the factorial vignette survey methodology supports the inductive measurement of the concept – here, the legitimacy of a company’s decision – through the analysis of the factors in the vignette. Through the analysis, we measured the relative importance of the vignette factors in driving the perception of legitimacy of the respondent. This methodology, with a single rating task, has been used to measure the relative importance of vignette factors on just wages (Jasso 2007), just punishments (Hagan et al. 2008), the trustworthiness of an organization (Pirson et al. 2017), and privacy expectations (Martin and Nissenbaum 2020). The vignette factors constitute the theoretically important constructs that may drive the perception of rating task (trust, fairness, privacy, legitimacy). The methodology is useful to allow the researcher to identify how each factor drives the perception of the dependent variable.

**Vignette Template**

The vignette template supported including different factors (below) in the vignette. The template was dynamically created as the respondent took the survey. For each survey run, a different combination of factors was included. The factors – decision type, outcome, governance, and
arbitrary – and the levels of each factor are described below in Tables 1-4. The vignette template is as follows – the labels are added only for clarity as well as the number of levels or options within each factor:

**Decision:** A computer program determines [Decision – 5 levels]. To make the decision, the program uses [Decision 2].

**Outcome:** Based on the program, [Outcome – 2 levels].

**Arbitrary:** It turns out, the decision was partly based on [Arbitrary – 3 levels].

**Governance:** The individual is told [Governance – 6 levels].

**Factors**

**Decision Type.**

The first set of hypotheses center on the type of decision and its importance in people’s lives. We included five decision types in our study: which advertisements people see online, which songs are suggested by music app, which applicants are hired for a job, which insurance claims are filled, and which video content is taken down by an online platform. These decision types were chosen based on how pivotal they are in society based on current literature (O’Neil 2016; Tufekci 2015). The degree the decision was deemed pivotal was verified by a pre-test survey of 1,024 respondents on Amazon Turk. The respondents were assigned one of two conditions – either the importance to an individual or the importance to society – and asked: “Please rate the degree to which the following decision represents a critical decision affecting someone’s life” or “Please rate to which the following decision represents a critical decision in society.” The results are in Table A1 in the Appendix.
Ads and music were deemed the least pivotal and hiring and insurance decisions were the most pivotal for both the individual and society based on the pre-test. In the results below, highly pivotal decisions were operationalized as vignettes with hiring or insurance decisions and low pivotal decisions were operationalized as vignettes with ads and music decisions.

Table 1: Types of Decisions

<table>
<thead>
<tr>
<th>Decisions</th>
<th>As Operationalized in the Vignettes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ads</td>
<td>which ads a person sees online … their predicted preference and interests.</td>
</tr>
<tr>
<td>Music</td>
<td>what songs are suggested in a person’s playlist… their predicted preferences and interests</td>
</tr>
<tr>
<td>Video</td>
<td>whether someone's posted videos are taken down online &amp; their past and predicted behavior</td>
</tr>
<tr>
<td>Hired</td>
<td>who is hired at a company….the applicant’s predicted likelihood to be successful at that job</td>
</tr>
<tr>
<td>Insurance</td>
<td>which insurance claims are paid …. the likelihood the policy holder will fight the claim</td>
</tr>
</tbody>
</table>

Outcomes.

In order to test the role of the outcome in the perceived legitimacy of ADM decisions, we varied the outcome from being either positive for the individual in the vignette (someone was hired) or negative for the individual (someone was not hired). The good/bad outcome was contextualized to the decision type as in Table 2.
Table 2: Outcome Factor for Vignettes

<table>
<thead>
<tr>
<th>Decision</th>
<th>As Operationalized in the Vignettes: Bad v good outcome.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ads</td>
<td>a user sees ads for well-paying jobs</td>
</tr>
<tr>
<td></td>
<td>a user sees primarily ads for poorly paying jobs</td>
</tr>
<tr>
<td>Music</td>
<td>a user hears songs that are totally different from those</td>
</tr>
<tr>
<td></td>
<td>that they like</td>
</tr>
<tr>
<td></td>
<td>a user hears songs that are similar to those that they</td>
</tr>
<tr>
<td></td>
<td>like</td>
</tr>
<tr>
<td>Hired</td>
<td>an applicant is not hired at the grocery store.</td>
</tr>
<tr>
<td></td>
<td>an applicant is hired at the grocery store.</td>
</tr>
<tr>
<td>Insurance</td>
<td>a policy holder is then told that the claim is fully</td>
</tr>
<tr>
<td></td>
<td>paid</td>
</tr>
<tr>
<td></td>
<td>a policy holder is then told that the claim will not be</td>
</tr>
<tr>
<td></td>
<td>paid</td>
</tr>
<tr>
<td>Video</td>
<td>a user who posted a video then sees that the video is</td>
</tr>
<tr>
<td></td>
<td>removed</td>
</tr>
<tr>
<td></td>
<td>a user who posted a video then sees that the video is</td>
</tr>
<tr>
<td></td>
<td>not removed</td>
</tr>
</tbody>
</table>

Arbitrariness.

To test hypothesis 3, we included the types of factors both arbitrary (day of the week) and discriminatory (their race and ethnicity) as well as a null (their online activities).

Table 3: Additional Factors Included in Algorithm

<table>
<thead>
<tr>
<th>Substance</th>
<th>As Operationalized in the Vignettes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity</td>
<td>their online activity, searches,</td>
</tr>
<tr>
<td></td>
<td>and purchases.</td>
</tr>
<tr>
<td>Day</td>
<td>the day of the week</td>
</tr>
<tr>
<td>Race</td>
<td>their race and ethnicity</td>
</tr>
</tbody>
</table>

Governance.

To test whether the type of governance over an algorithmic decision impacted the degree to which the decision is judged legitimate, we included five options for the vignettes: transparency, in which the organization notifies individuals that a computer program made the decision (Diakopoulos 2020); impact assessment, in which the organization completed an assessment of
the impact of the algorithmic process on fairness and privacy (Yam and Skorburg 2021); audit governance, in which an external entity completed an annual independent audit to ensure algorithmic decisions are not biased (Mittelstadt 2016); human governance, or so-called “human in the loop” of the algorithmic process (Elish 2019); and appeals, in which the decision can be appealed by the individual to an internal review board (Mulligan et al. 2020). Perceptions of the legitimacy of these decisions were compared to a null in which a human made the decision without the help of an algorithm. There are, of course, other possible governance options. We chose these options because they are the most commonly proposed governance mechanisms in the legal studies literature and because they sit on a range from more robust to lax, providing a proxy for the effect on legitimacy of different types of procedural governance (Citron 2007; Martin 2018; Rahwan 2018; Froomkin, Kerr, and Pineau 2019; Jones 2017; Reisman et al. 2018; Katyal 2019; Loi et al. 2020).

Table 4: Governance Factor for Vignettes

<table>
<thead>
<tr>
<th>Governance</th>
<th>As Operationalized in the Vignettes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Only</td>
<td>Entire vignette is about human decision</td>
</tr>
<tr>
<td>Appeals</td>
<td>that automated decisions can be appealed to a review board.</td>
</tr>
<tr>
<td>Human Governance</td>
<td>the organization hired a professional to oversee the automated decision.</td>
</tr>
<tr>
<td>Audited</td>
<td>an independent audit by an external firm is conducted annually to ensure automated decisions are not biased.</td>
</tr>
<tr>
<td>Impact Assessment</td>
<td>the organization did an impact assessment on the automated system.</td>
</tr>
<tr>
<td>Notification</td>
<td>that the organization notifies individuals that a computer program made the decision.</td>
</tr>
</tbody>
</table>
Vignette Creation

When the factors in Tables 1-4 are inserted in the vignette template, the respondent then is able to view and rate the dynamically created vignettes as shown in the examples below. The labels (Decision, Outcome, Arbitrary Governance) are only included to explain the methodology; the respondents did not see these labels.

Example 1:

**Decision:** A computer program determines what songs are suggested in a playlist (e.g., Spotify). To make the decision, the program uses their predicted preferences and interests.

**Outcome:** Based on the program, a user hears songs that are totally different from those that they like.

**Arbitrary:** It turns out, the decision was partly based on their race and ethnicity.

**Governance:** The individual is told an independent audit by an external firm is conducted annually to ensure automated decisions are not biased.

Example 2:

**Decision:** A computer program determines which ads a person sees online. To make the decision, the program uses their predicted preferences and interests.

**Outcome:** Based on the program, a user sees ads for well-paying jobs.

**Arbitrary:** It turns out, the decision was partly based on the day of the week.

**Governance:** The individual is told that the organization notifies individuals that a computer program made the decision.

Rating Task

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 scale not visible to the respondents. The slider option allows the respondent more freedom to rate the vignette.

**Conditions and Sample**

The vignette survey was run under nine conditions described in Table 5. This allowed us to isolate the importance of each factor by comparing legitimacy averages as the factors were included. For example, we tested the impact of including the outcome by comparing the averages of Survey 2 (only decision included) versus Survey 3 (decision and outcome included).

Normally, a factorial vignette survey analysis would include a block analysis to isolate which factors, out of many, were dominant in driving the judgement of the respondents. However, we were concerned that some factors, such as the arbitrary factors, could prove to be so important as to overwhelm the respondent and cloud how they perceived the other vignettes. We therefore spent more time running nine surveys over four months in order to better isolate the importance of each type of factor.

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 20-30 vignettes (depending on the condition) taking approximately 10 minutes; U.S. respondents were paid $1.60-$1.80 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 calculating whether respondents ‘clicked through’ without actually judging the vignette (Coppock 2018; Martin 2019; Tucker 2014; Daly and Natarajan 2015). Finally, the design of the survey is to identify theoretically generalizable results as to the relative importance of factors in driving perceptions of legitimacy of AI decisions.

**Table 5: Conditions and Sample**

<table>
<thead>
<tr>
<th>Survey</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vignette Factors Included</td>
<td>Decision</td>
<td>Decision</td>
<td>Decision</td>
<td>Decision</td>
<td>Decision</td>
<td>Decision</td>
<td>Decision</td>
<td>Decision</td>
<td>Decision</td>
</tr>
<tr>
<td>Outcome</td>
<td>Decision</td>
<td>Decision</td>
<td>Decision</td>
<td>Decision</td>
<td>Decision</td>
<td>Decision</td>
<td>Decision</td>
<td>Decision</td>
<td>Decision</td>
</tr>
<tr>
<td>Arbitrary</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
</tr>
<tr>
<td>Human</td>
<td>Human</td>
<td>Human</td>
<td>Human</td>
<td>Human</td>
<td>Human</td>
<td>Human</td>
<td>Human</td>
<td>Human</td>
<td>Human</td>
</tr>
<tr>
<td>N (respondents)</td>
<td>412</td>
<td>384</td>
<td>417</td>
<td>517</td>
<td>520</td>
<td>469</td>
<td>538</td>
<td>470</td>
<td>500</td>
</tr>
<tr>
<td>Vignettes</td>
<td>10,300</td>
<td>9,600</td>
<td>10,425</td>
<td>12,925</td>
<td>13,000</td>
<td>11,725</td>
<td>16,140</td>
<td>14,100</td>
<td>15,000</td>
</tr>
<tr>
<td>Female</td>
<td>42.3%</td>
<td>41.2%</td>
<td>41.5%</td>
<td>45.7%</td>
<td>45.7%</td>
<td>40.3%</td>
<td>43.9%</td>
<td>41.3%</td>
<td>45.4%</td>
</tr>
<tr>
<td>AgeOver35</td>
<td>53.8%</td>
<td>53.6%</td>
<td>49.4%</td>
<td>53.9%</td>
<td>57.1%</td>
<td>55.9%</td>
<td>54.8%</td>
<td>60.6%</td>
<td>59.0%</td>
</tr>
</tbody>
</table>

7 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).

8 Turk has been used for theoretical generalizability quite successfully, as in the examination of the relationship between concepts or ideas (Kang, Brown, Dabbish, & Kiesler, 2014; Martin & Nissenbaum, 2017a; Redmiles, Kross, Pradhan, & Mazurek, 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; as such 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, e.g., whether cause-effect relationships hold or whether concepts are related (Lynch, 1982).
RESULTS

Type of Decision

Hypothesis 1 states that as decision importance increases, individuals’ perception of the legitimacy of using an algorithm to make the decision decreases. To test hypothesis 1, we regressed the rating task, the degree the decision was judged legitimate, on the vignette factors. The results are in Tables 6a and 6b. The coefficients ($\beta$) measure the relative importance of the given vignette factor to the rating task, and p is the significance of the finding. Pivotal decisions are those decisions deemed more important in the pre-survey test. For each survey run, more pivotal decisions lowered the legitimacy of the decision with $\beta < 0$ and $p < 0.001$. For example, in Survey 2, which only included the decision type in the vignette, including high pivotal decisions lowered the legitimacy rating by -47.20 ($p < 0.001$). This finding held across each survey and combination of factors. Less pivotal decisions made by AI are seen as more legitimate than more pivotal decisions.

<p>| Table 6a: Regression Results of Legitimacy Rating Task on Vignette Factors for Each Survey Run. |</p>
<table>
<thead>
<tr>
<th>Survey 1</th>
<th>Survey 2</th>
<th>Survey 3</th>
<th>Survey 4</th>
<th>Survey 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>HighPivotal</td>
<td>$\beta$</td>
<td>$p$</td>
<td>$\beta$</td>
<td>$p$</td>
</tr>
<tr>
<td>Null = Low Pivotal</td>
<td>-25.68</td>
<td>0.00</td>
<td>-47.20</td>
<td>0.00</td>
</tr>
<tr>
<td>BadOutcome</td>
<td>-12.92</td>
<td>0.00</td>
<td>-43.90</td>
<td>0.00</td>
</tr>
<tr>
<td>null = GoodOutcome</td>
<td>-31.28</td>
<td>0.00</td>
<td>-45.71</td>
<td>0.00</td>
</tr>
<tr>
<td>DaySubstance</td>
<td>-40.26</td>
<td>0.00</td>
<td>-46.57</td>
<td>0.00</td>
</tr>
<tr>
<td>RaceSubstance</td>
<td>null = activity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AppealGovern</td>
<td>4.54</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AuditGovern</td>
<td>-3.01</td>
<td>0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HumanGovern</td>
<td>-6.49</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ImpactAssessGovern</td>
<td>-6.58</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>null = Notice Govern</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>_cons</td>
<td>27.30</td>
<td>0.00</td>
<td>51.35</td>
<td>0.00</td>
</tr>
<tr>
<td>Average</td>
<td>-17.11</td>
<td>25.53</td>
<td>14.14</td>
<td>-14.05</td>
</tr>
<tr>
<td>N (respondents)</td>
<td>412</td>
<td>384</td>
<td>417</td>
<td>517</td>
</tr>
</tbody>
</table>
Table 6b: Regression Results of Legitimacy Rating Task on Vignette Factors for Each Survey Run.

<table>
<thead>
<tr>
<th></th>
<th>Survey 6</th>
<th></th>
<th>Survey 7</th>
<th></th>
<th>Survey 8</th>
<th></th>
<th>Survey 9</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \beta )</td>
<td>( p )</td>
<td>( \beta )</td>
<td>( p )</td>
<td>( \beta )</td>
<td>( p )</td>
<td>( \beta )</td>
<td>( p )</td>
</tr>
<tr>
<td>HighPivotal</td>
<td>-39.34</td>
<td>0.00</td>
<td>-23.30</td>
<td>0.00</td>
<td>-19.06</td>
<td>0.00</td>
<td>-6.51</td>
<td>0.00</td>
</tr>
<tr>
<td>Null = Low Pivotal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BadOutcome</td>
<td></td>
<td></td>
<td>-15.19</td>
<td>0.00</td>
<td>-16.68</td>
<td>0.00</td>
<td>-27.43</td>
<td>0.00</td>
</tr>
<tr>
<td>null = GoodOutcome</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DaySubstance</td>
<td>-27.29</td>
<td>0.00</td>
<td>-32.22</td>
<td>0.00</td>
<td>-22.86</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RaceSubstance</td>
<td>-43.69</td>
<td>0.00</td>
<td>-42.39</td>
<td>0.00</td>
<td>-39.62</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>null = activity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AppealGovern</td>
<td>1.10</td>
<td>0.47</td>
<td>1.65</td>
<td>0.22</td>
<td>8.33</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AuditGovern</td>
<td>1.40</td>
<td>0.36</td>
<td>2.15</td>
<td>0.11</td>
<td>2.96</td>
<td>0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HumanGovern</td>
<td>-2.78</td>
<td>0.07</td>
<td>0.28</td>
<td>0.83</td>
<td>-0.44</td>
<td>0.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ImpactAssessGovern</td>
<td>-1.83</td>
<td>0.23</td>
<td>-2.02</td>
<td>0.13</td>
<td>-4.61</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>null = Notice Govern</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>_cons</td>
<td>34.92</td>
<td>0.00</td>
<td>23.62</td>
<td>0.00</td>
<td>29.03</td>
<td>0.00</td>
<td>34.34</td>
<td>0.00</td>
</tr>
<tr>
<td>Average</td>
<td>-11.38</td>
<td>-20.65</td>
<td></td>
<td>-9.84</td>
<td>19.26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N (respondents)</td>
<td>469</td>
<td>538</td>
<td>470</td>
<td>500</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Outcomes

Hypothesis 2 states a good outcome is associated, an increase in the perceived legitimacy of the decision. To test Hypothesis 2, we regressed the dependent variable, the degree to which the decision was judged legitimate, on the vignette factors for Survey 3 (with only the decision and outcome included). We find a bad outcome creates a legitimate penalty of -43.90 (\( p < 0.001 \)) compared to a good outcome. In other words, a good outcome has a positive impact on perceived legitimacy. This result held even when additional factors are added in Surveys 7, 8, and 9 in Table 6b.

We also tested if the legitimacy dividend associated with a good outcome increases as decision importance increases. Figure 2 illustrates the legitimacy ratings for both good and bad outcomes. The benefit of a good outcome over a bad outcome is larger for a low pivotal decision (Good =
50.21; Bad = -9.31; (t = 35.9793; p < 0.001)). compared to a high pivotal decision (good = 23.19; bad = -3.49; (t = 14.53; p < 0.001)). In Figure 2, the steeper slope represents a greater legitimacy penalty for bad outcomes for low pivotal decisions as compared to high pivotal decisions ($X^2 = 196.02$; $p < 0.001$). This could be because respondents expect a good outcome with low pivotal decisions and then penalize the surprise bad outcome more.

**Figure 2: Impact of Good and Bad Outcome on Pivotal Decisions**

![Impact of Outcome Included](image)

**Arbitrary Factors**

Hypothesis 3 states that *the use of arbitrary or race-based factors has a negative impact on perceived legitimacy*. The inclusion of arbitrary factors in Survey 4 lowers the legitimacy rating, all else being equal, from 25.53 (Survey 2) to -14.05 (Survey 4) ($t = -46.06$; $p < 0.001$). In general, the lowest average legitimacy rating is found in Surveys that include arbitrary factors (Survey 6 -11.38; Survey 7 -20.65; Survey 8 -9.84).
Governance and Legitimacy

Hypothesis 4a stated that the inclusion of any governance regime has a positive impact on perceived legitimacy of algorithmic decision-making relative to no governance, but there is a larger legitimacy dividend as procedural governance becomes more robust. To test Hypothesis 4a, we examine the regression results in Tables 6a and 6b for the surveys with governance included (Surveys 5, 6, 8, and 9). Here we see that offering an appeal process provides a legitimacy dividend for decisions with no outcome or arbitrary factors included (Survey 5). The more robust governance mechanism, to allow an appeal, positively impacts perceived legitimacy ($\beta = 4.54, p<0.01$). However, alternative governance mechanisms such as an impact assessment, including a human in the loop, and having the program audited, lower the legitimacy of the decision compared to mere notice. This was a surprise and counter to the hypothesis.

Outcomes and Governance

Hypothesis 4b states that robust procedural governance mechanisms moderate the legitimacy penalties associated with bad outcomes. In other words, offering an appeal to the algorithmic decision legitimizes even a bad outcome. We tested hypothesis 4b by comparing the legitimacy rating for decisions with an outcome across governance mechanisms. As shown in Figure 3, we found that for bad outcomes, the use of an appeal process does provide a legitimacy dividend ($t = -12.53; p<0.0001$); however, the use of an appeal process had no effect on the perceived legitimacy of algorithmic decisions with good outcomes ($t = -1.74; p = 0.04$). Further, the dividend did not ‘make up’ for the penalty of the bad outcome. The average legitimacy rating for a bad outcome even with an appeal process is still lower than that of a good outcome (with or
without an appeal process included) – however the appeal process does close the gap between good and bad outcome legitimacy as in Figure 3.

**Figure 3: Impact of Adding a Robust Governance Mechanism (Appeal) to AI Decisions with Good and Bad Outcomes**

![Impact of Appeal on Good and Bad Outcomes](image)

**Arbitrary Factors and Governance**

Hypothesis 4c states that *there should be a legitimacy penalty associated with algorithms that use arbitrary or race-based factors, regardless of the form of procedural governance used.* To test Hypothesis 4c, we compared the average legitimacy rating for the inclusion of race and arbitrary factors both with and without governance mechanisms. Figure 4 shows the impact of including the appeal of the decision as a governance mechanism, previously shown to have the greatest legitimacy dividend, on decisions with either race or day of the week. The impact of including an appeal on decisions that used the day of the week was modest but significant (t = 3.481; p < 0.001). The impact of including an appeal for decisions utilizing race was not significant (t = .863; p = 0.19). The legitimacy dividend of including an appeal process improves the perceived legitimacy of decisions that include arbitrary factors such as the day of the week but not when including unjust factors such as race.
To better understand the legitimacy penalty for having arbitrary rationales (race or day of the week) with algorithmic decisions, we compared the impact of including the arbitrary factor of race on decisions with both good and bad outcomes. The results, in Figure 5, show that including race as a factor is a legitimacy penalty for bad outcomes ($t = 26.40; p < 0.001$). In addition, any legitimacy dividends of a good outcome are erased when the decision included race as a factor ($t = 50.52; p < 0.001$). Adding an appeal for governance improves the legitimacy only slightly for good outcomes and race (from $-34.18$ to $-24.59$ with an appeal; $t = 3.14; p < 0.001$) and slightly for bad outcomes and race (from $-43.59$ to $-33.17$ with an appeal; $t = 3.32; p < 0.001$). Including race as an arbitrary and unjust factor in an algorithmic decision stubbornly delegitimizes the decision regardless of a good or bad outcome and regardless of including an appeal.
Figure 5: Impact of Adding a Race and Day Arbitrary Factors to Algorithmic Decisions with Good and Bad Outcomes

Table 7: Summary of Findings
DISCUSSION AND CONCLUSION

This paper explores the conceptual antecedents to the perceived legitimacy of algorithmic decisions. Building on cross-disciplinary empirical measurements of decision, organizational, and institutional legitimacy (Elsbach 1994; Finch et al. 2015; Jahn et al. 2020; Tyler 1990; Badas 2019) and theoretical accounts of legitimacy (Suchman 1995; Palazzo and Scherer 2006), we studied how factors of ADM – the type of decision, the outcomes, the rationale, and the governance – impact the perception of decisional legitimacy.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Finding</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>H1</strong> As decision importance increases, individuals’ perception of the</td>
<td>Less pivotal decisions made by AI are seen as more legitimate than more pivotal decisions.</td>
</tr>
<tr>
<td>legitimacy of using an algorithm to make the decision decreases.</td>
<td></td>
</tr>
<tr>
<td><strong>H2</strong> A good outcome is associated with an increase in the perceived</td>
<td>We find a bad outcome is a legitimate penalty compared to a good outcome. In other words, a good outcome has a positive impact on perceived legitimacy.</td>
</tr>
<tr>
<td>legitimacy of the decision</td>
<td></td>
</tr>
<tr>
<td><strong>H3</strong> The use of arbitrary or race-based factors has a negative impact</td>
<td>The inclusion of arbitrary factors in Survey 4 lowers the legitimacy rating, all else being equal, from 25.53 (Survey 2) to -14.05 (Survey 4) (t = -46.06; p &lt; 0.001).</td>
</tr>
<tr>
<td>on perceived legitimacy.</td>
<td></td>
</tr>
<tr>
<td><strong>H4a</strong> Any governance regime brings a legitimacy dividend to algorithmic</td>
<td>The more robust governance mechanism, to allow an appeal, positively impacts perceived legitimacy. However, alternative governance mechanisms such as an impact assessment, including a human in the loop, and having the program audited, lower the legitimacy of the decision compared to mere notice. This was a surprise and counter to the hypothesis.</td>
</tr>
<tr>
<td>decision-making relative to no governance, but there is a larger legitimacy dividend as procedural governance becomes more robust.</td>
<td></td>
</tr>
<tr>
<td><strong>H4b</strong>: robust procedural governance mechanisms moderate the legitimacy</td>
<td>For bad outcomes, the use of an appeal process does provide a legitimacy dividend. However, the use of an appeal process had no effect on the legitimacy of algorithmic decisions with good outcomes. The average legitimacy rating for a bad outcome even with an appeal process is still lower than that of a good outcome.</td>
</tr>
<tr>
<td>penalties associated with bad outcomes.</td>
<td></td>
</tr>
<tr>
<td><strong>H4c</strong>: There is a legitimacy penalty associated with algorithms that use</td>
<td>The legitimacy dividend of including an appeal process improves the legitimacy of decisions including arbitrary factors such as the day of the week but not when including unjust factors such as race.</td>
</tr>
<tr>
<td>arbitrary or race-based factors, regardless of the form of procedural</td>
<td></td>
</tr>
<tr>
<td>governance used.</td>
<td></td>
</tr>
</tbody>
</table>
Empirically, we used factorial vignette methodology and made several hypotheses. We show that perceived legitimacy varies inversely with decision importance. We hypothesized that the robustness of procedural governance over algorithmic decisions made by firms would vary positively and directly with legitimacy. The data suggest something more nuanced. The only form of procedural governance that carries a legitimacy dividend is the most robust that we studied: a formal appeal to a human decision maker. Those more commonly in practice or in legislative proposals today—notice and impact assessments—either have no impact or carry a legitimacy penalty.

We hypothesized that outcomes, defined as whether an individual agrees with a decision or the decision is good for the individual, would matter more for legitimacy than procedural governance. This was mostly true, although the data suggested that there was no legitimacy dividend for positive outcomes for decisions of low importance. We also hypothesized that the use of morally questionable or arbitrary factors in making decisions would carry legitimacy penalties. Our study suggested that this was the case, erasing all legitimacy dividends associated with robust governance or positive outcomes.

*Algorithms and Business Ethics*

Scholars in business ethics have critically examined ADM within specific applications such as accounting (Gunz and Thorne 2020; Munoko, Brown-Liburd, and Vasarhelyi 2020), financial services (Owen and Abeka Arthur 2019) and HR (Leicht-Deobald, et al 2019), and have identified specific moral implications such as social-media addiction (Bhargava and Velasquez
2020), personalized pricing (Seele, Dierksmeier, Hofsteter, and Schultz 2019; Steinberg 2020), gamification (Kim 2019), and accountability (Martin 2019). Previous work has also critically examined the details of algorithms in ride sharing can create gender discrimination in ratings (Greenwood, Adjerid, Angst, and Meikle 2020). We extend this line of AI Ethics scholarship within business ethics to the specific examination of how ADM will impact legitimacy.

Future work could similarly examine how specific design decisions of ADM (including the use of AI and different types of machine learning) impact business ethics outcomes such as justice, fairness, trust, and legitimacy. For example, scholarship could investigate how data, governance, outcomes, and the use of AI in general impact employee perceptions of fairness. Trust in a firm, and trust factors such as ability, benevolence, and integrity, may be impacted by the use of particular data, the degree to which the firm allows decisions to be contested, and the outcome of the algorithmic decision. Scholars could leverage work on the purpose of the firm to help AI ethics scholars understand how ADM outcomes, designed by computer scientists, may undermine legitimate purposes of the firm. Rather than mistakenly seeing the introduction of algorithmic decision making as fundamentally changing how we assess ethical decisions, business ethics should leverage well researched ethical concepts to illustrate how AI does not change the nature of corporate responsibility. Work in AI Ethics within business ethics can (and should) start to connect the known moral implications of AI with our existing frameworks in business ethics in order to have concrete implications for practice as well as contribute to business ethics theory.

**AI Ethics Research**
Although work on fairness, accountability, transparency, and explainability (FATE) has flourished in the past few years, more work on the perceptions of ADM – parallel to the work on perceptions of human-focused decisions – should continue. For example, scholars have studied whether ADM is perceived as authentic (Jago 2019) or fair (Araujo et al. 2020; Nagtegaal 2021; Newman et al. 2020). More work should be done taking existing work on ethical decision making in business ethics and management to investigate whether the factors important to human decisions maintain the same relative importance to judgements about ADM. This could then help guide AI ethics research on the design of algorithmic decision systems.

This paper contributes to AI ethics scholarship by extending the moral implications that should be a concern to firms. While firms have been told to address transparency and accountability as well as discrimination, this study suggests that firms should also be concerned about legitimacy of the decision and the organization as being impacted by *how the ADM is designed*. Future work could continue to examine design decisions and their impact on different dependent variables including legitimacy, fairness, and trust in addition to how the design decisions impact users and subjects of the ADM. Extending the research to include how the firm itself is impacted by the value-laden decisions may resonate with managers and business scholarship.

*Legitimacy*

The results of this study add to our understanding of issue, organizational, and institutional legitimacy. Business, management, and sociolegal scholars studying legitimacy have shown that
factors like communication (Elsbach 1994), motives (Jahn et al. 2020), procedural fairness (Tyler 1990), ideology (Badas 2019; Finch et al. 2015), and agreement or disagreement with outcomes (Gibson and Caldeira 2009) impact legitimacy. Our study not only adds additional factors—decision importance (H1), outcomes (H2a) and the use of arbitrary or morally dubious variables (H3)—but also tests the interaction among the factors (H4a, H4b, and H4c). In so doing, we add nuance to the existing literature. For example, Tyler (1990) suggests that people tend to think that authorities’ decisions are legitimate when the decision-making process is fair and when individuals have the opportunity to be heard through procedural due process. Our findings are partly in line with this research, but we show that not all procedural governance mechanisms are the same. Only an appeal to a human authority is generally capable of legitimizing algorithmic systems, all other factors held constant. It is possible that other procedural mechanisms do not provide the kind of robust guarantees of fairness that individuals may associate with the right to appeal.

In work studying popular perceptions of the legitimacy of the United States Supreme Court, Badas (2019) and Gibson, Caldeira, and Spence (2005) suggest that ideological agreement or disagreement with an authority’s decision is a significant driver of legitimacy, regardless of fair process. Finch et al. (2015) showed that ideological views about environmental protection and economic interests helped determine perceptions of legitimacy of a high-pollution industry. Our study tests this relationship in the context of private firm decisions and finds, with additional nuance, that outcomes do affect perceptions of legitimacy. Positive outcomes confer legitimacy dividends for decisions of high importance, but not for decisions of low importance, and negative outcomes create larger legitimacy penalties for less important decisions than highly pivotal
decisions. This is in line with current literature, which suggests that individuals credit institutions that make the “right” decisions as they see them without seeking explanations (Lodge and Taber 2013). Although our study did not examine the rationales for legitimacy penalties for outcomes mediated by decision importance, the variance may be because individuals expect that such decisions should be easy to get right and, as such, tend to react negatively when computer get them wrong.

These findings contribute to our understanding of positivity theory (Gibson and Caldeira 2009). Positivity theory attempts to explain why individuals punish institutions less for decisions with which they disagree than they reward them for decisions with which they agree. Gibson, Lodge, and Woodson (2014) suggest that public exposure to performative symbols of authority insulates authoritative institutions from the worst legitimacy penalties associated with ideological or policy disagreement. Our study suggests that this may only be true for decisions of high importance, but the extent, if any, of the mediation of decision importance would have to be studied empirically.

Finally, this research highlights the dangers to legitimacy for any firm that makes decisions via ADM systems that rely on race, proxies for race, or seemingly arbitrary factors that individuals do not see as related to the decision itself. Given how difficult it is to remove racial discrimination from AI systems trained on data that is itself the product of systemic and institutional discrimination, the algorithmic accountability literature is right to increasingly focus on issues of structural fairness rather than mere remediation through governance or better data (Pasquale 2019).
Unfortunately, current policy proposals lag behind this the “second wave” of the algorithmic accountability literature (Pasquale 2019). Article 22 of the European Union’s General Data Protection Regulation places restrictions on private firms that employ algorithms without human intervention to make decisions that have significant effects on individuals. Firms can do so, but only if they adopt “suitable measures to safeguard the data subject’s rights and freedoms and legitimate interests,” which might include “at least the right to obtain human intervention … , to express his or her point of view and to contest the decision (General Data Protection Regulation 2018). Other provisions of the GDPR may also impact algorithmic decision-making systems, as well (Edwards and Veale 2017), but they all share one primary common thread: they rely on procedural due process mechanisms to protect individual rights and ensure fairness (Waldman 2021). Our study calls into question the capacity of such procedural governance to legitimize algorithmic decisions in the eyes of the public, lending credibility to arguments from critical scholars that current law needs to be more robust or, perhaps, disused in certain circumstances and for certain purposes all together (Pasquale 2015, 2018, 2021). Although our study did not explore perceptions of the legitimacy of firms that stop using discriminatory algorithms, the strong legitimacy penalties associated with use implies the possibility of legitimacy dividends for changing course.

Of course, even the procedural guardrails of the GDPR do not exist in the United States. Firms are stepping into uncharted waters in the US, which lacks a federal agency for algorithms and
neither requires notice nor impact assessments for algorithmic decisions (Crawford and Schultz 2013). Our study suggests that US policymakers should be careful following the GDPR’s procedural governance model when, if ever, they start regulating private firm use of algorithmic decision-making systems. If, as we show, that only the most robust procedures, such as offering an appeal akin to Mulligan et al. (2020), confer legitimacy benefits, that outcomes are far more powerful drivers of perceptions of legitimacy, and that almost all kinds of algorithmic decisions are viewed as illegitimate when they use race-based or arbitrary factors, then policymakers should consider more substantive limits on algorithmic inputs and uses rather than procedural safeguards.
REFERENCES


https://doi.org/10.1007/s10551-015-2839-2


