Business and Politics (2024), **26**, 200–217
doi:10.1017/bap.2023.35

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

The Paradox of Algorithms and Blame on Public Decision-makers

Adam L. Ozer[1], Philip D. Waggoner[2] and Ryan Kennedy[3]

[1]Verian, LONDON, United Kingdom, [2]Columbia University, New York, NY, USA and [3]University of Houston, Houston, TX, USA

Corresponding author: Ryan Kennedy; Email: rkennedy@uh.edu

Abstract

Public decision-makers incorporate algorithm decision aids, often developed by private businesses, into the policy process, in part, as a method for justifying difficult decisions. Ethicists have worried that over-trust in algorithm advice and concerns about punishment if departing from an algorithm's recommendation will result in over-reliance and harm democratic accountability. We test these concerns in a set of two pre-registered survey experiments in the judicial context conducted on three representative U.S. samples. The results show no support for the hypothesized blame dynamics, regardless of whether the judge agrees or disagrees with the algorithm. Algorithms, moreover, do not have a significant impact relative to other sources of advice. Respondents who are generally more trusting of elites assign greater blame to the decision-maker when they disagree with the algorithm, and they assign more blame when they think the decision-maker is abdicating their responsibility by agreeing with an algorithm.

Keywords: algorithms; artificial intelligence; public policy; public opinion; experiments

The use of algorithms in the public sphere is exploding. Algorithms have been applied in criminal justice,[1] voting,[2] redistricting,[3] policing,[4] allocation of public services,[5] immigration,[6] military and intelligence decision-making,[7] and a range of other sensitive fields. Given the resource constraints of government agencies, who lack the resources to build their own systems and to pay the premium associated with hiring high-level software engineers, most of these algorithms are developed in the private sector.[8] For example, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software, which has been used for assessing the risk of defendants in bail hearings and in criminal sentencing in at least four states and was the source of heated controversy about racial bias and a case before the Wisconsin Supreme Court,[9] was developed by Northpointe (now Equivant). Similarly, the PredPol system for identifying areas for additional police presence was developed by PredPol, Inc. (now Geolitica) and has been used by at least 20 U.S. law enforcement agencies to help inform their policing practices. This system has also been highly controversial due to concerns that it disproportionately identifies communities of color for additional police presence.[10]

[1]Surden (2021).
[2]Berman (2015); Ingraham (2017).
[3]Chen and Rodden (2015).
[4]Brayne (2020).
[5]B. Green and Franklin-Hodge (2020).
[6]Molnar (2021).
[7]Ramakrishnan et al. (2014; Scharre (2018).
[8]Fry (2018).
[9]Angwin et al. (2016).
[10]Fry (2018); Benjamin (2019); Mehrotra (2021).

Sociological studies suggest that decision-makers give a high level of credence to algorithms in making decisions in spite of concerns over bias, accuracy, and myopic optimization.[11] Sentencing algorithms, for example, have received broad bipartisan support for their expanded use in Congress.[12] There is also increasing evidence that the public trusts algorithms, at least in terms of their behavior in response to advice, as much or more than other sources of advice in a range of scenarios, including in public policy,[13] although some scholars still note significant aversion to algorithms in some contexts.[14] The seeming increase in public trust in algorithms runs counter to recent studies on trust in human experts, which suggest a growing negative sentiment and general anti-intellectual attitudes among the public,[15] even as they find them more persuasive than non-expert information sources.[16]

In response, ethicists and legal scholars have raised two, interrelated concerns. First, public decision-makers might rely too much on algorithms because of their perceived "objectivity" and "efficiency."[17] These tendencies have been labeled "technogoggles,"[18] "technowashing,"[19] or "math washing"[20] by critics, who see these tools as either a way for public decision-makers to base their decisions on "more objective" criteria and/or a way for avoiding responsibility for difficult, discretionary decisions. Scholars worry that the use of algorithms adds a sense of legitimacy to otherwise contested decisions[21] and allows for scapegoating the algorithm for mistakes.[22] These claims are quite prominent, appearing in several multi-award-winning and bestselling books[23] and throughout a collection of essays by top scholars in the field of artificial intelligence (AI) ethics.[24]

Second, even if a policymaker has doubts about an algorithm's information, they may feel pressured to comply with the algorithm's recommendations.[25] An elected official or bureaucrat who accepts an algorithm's judgment could pass the blame along to a flawed algorithm if there is an adverse outcome, while one who rejects the algorithm's judgment would have to explain why they rejected the (correct) information given to them.[26] Risk-averse political actors, these scholars fear, will face strong incentives to maintain the political cover algorithms provide. Surden's scenario is worth quoting at length, since, even though we were not aware of it at the time we designed our study, it mirrors our setup in many ways[27]:

> "[J]udges have incentives not to override automated recommendations. Imagine that a Judge was to release a defendant despite a high automated risk score, and that defendant were then to go on to commit a crime on release. The judge could be subject to backlash and criticism, given that there is now a seemingly precise prediction score in the record that the judge chose to override. The safer route for the judge is to simply adopt the automated recommendation, as she can always point to the numerical risk score as a justification for her decision."

He goes on to note that this is ethically problematic for at least three reasons: (1) the numeric scores pose a "problem of false precision," wherein the numeric scores are divorced from practical meaning;

[11]Zuiderwijk, Chen, and Salem (2021); B. Green and Franklin-Hodge (2020); Surden (2021); Goodman (2021); Hannah-Moffat (2015).
[12]See, e.g., the Sentencing Reform and Corrections Act (S. 2123) of 2016.
[13]Kennedy, Waggoner, and Ward (2022); Zwald, Kennedy, and Ozer (2021); Logg (2016); Logg, Minson, and Moore (2019).
[14]Dietvorst, Simmons, and Massey (2015); Gogoll and Uhl (2018).
[15]Merkley (2021); Merkley and Loewen (2021).
[16]Boudreau and McCubbins (2010); Druckman (2001b); Ozer (2020).
[17]Brayne (2020); B. Green and Franklin-Hodge (2020).
[18]B. Green and Franklin-Hodge (2020).
[19]Brayne (2020).
[20]Shane (2019).
[21]Zeide (2021); Perakslis (2020).
[22]Gill (2020).
[23]Benjamin (2019); O'Neil (2016); Brayne (2020).
[24]Surden (2021); Goodman (2021); Molnar (2021).
[25]Surden (2021); Danaher (2016).
[26]Surden (2021); Chesterman (2021).
[27]Surden (2021: 734).

https://doi.org/10.1017/bap.2023.35 Published online by Cambridge University Press

(2) the use of the scores produces a "subtle shifting of accountability for the decision away from the judge and toward the system"; and (3) the use of private, proprietary algorithms produces a "shift of accountability from the public sector to the private sector."

Albright provides some evidence that this process is playing out in actual bail decisions.[28] Looking at bail decisions in Kentucky, she finds that a lenient recommendation by the sentencing algorithm increases the likelihood of a lenient bail decision by about 50% for marginal cases. She posits that this is because judges believe that, if they make a retrospectively incorrect decision, at least part of the blame will fall on the recommendation instead of themselves. There have also been reports of the inverse mechanism. For example, in a story by *The New Yorker* about AI systems for evaluating the mental health of students in schools, a school therapist suggested that he would be unlikely to go against an AI evaluation that a student was potentially suicidal because of potential liability.[29]

These concerns are compounded by the development of most of these algorithms within the private sector. Intellectual property protections for the algorithms means that the public is often unaware of what data is being used or how that data is being modeled to produce the resulting predictions.[30] While advocates for the use of algorithms note that this is not too much different from the inscrutability of human motivations that may underlie particular decisions,[31] a growing chorus of concerns have been raised that algorithms allow biases to scale,[32] create negative feedback effects,[33] and increase discretion of agencies to pursue otherwise controversial or biased practices.[34]

Yet, there are reasons to doubt whether the blame dynamics suggested in this literature will be manifest in popular opinion. While AI experts prefer to emphasize the unique mathematical and technical aspects of AI,[35] the public tends to anthropomorphize AI, emphasizing the characteristics of AI that reflect human characteristics and expertise.[36] Indeed, some studies even suggest that the public attributes intentionality to algorithm actions.[37] This anthropomorphization may undermine the hypothesized distinction between AI and other forms of expert advice amongst the public. It may also undermine the blame dynamics hypothesized in the theoretical literature. Bertsou finds that the public supports the role of experts primarily in *how decisions are implemented* not *what decisions are made*.[38] This is consistent with Agrawal, Gans and Goldfarb's distinction of AI systems improving *prediction* to help humans make better *judgments*.[39] Even in studies that find higher weight given to advice from algorithms than from human experts, a large majority of respondents in all conditions give their own evaluation the highest weight,[40] suggesting hesitancy among the public to delegate even predictive tasks to an outside source, whether expert or algorithm. Thus, the blame dynamics posited by the above literature may not be manifested amongst the public if algorithms are viewed similarly to other expert advice, trust in which has been declining, or the public official is viewed as abrogating their obligation to use their own judgment.

This study focuses on evaluating the concerns of ethicists and legal scholars with regards to both the dislocation of blame (when a decision-maker makes a mistake in concurrence with an algorithm) and the magnification of blame (when a decision-maker makes a mistake in disagreement with an algorithm). We conducted a pre-registered experiment on two representative samples of the U.S. population to directly test these concerns in the judicial decision-making context (Study 1), a context notable for both its salience in the literature on algorithms in public policy and its centrality as an

[28]Albright (2023).
[29]https://www.newyorker.com/magazine/2023/03/06/can-ai-treat-mental-illness
[30]O'Neil (2016); Surden (2021).
[31]Fry (2018).
[32]Benjamin (2019).
[33]O'Neil (2016).
[34]Brayne (2020).
[35]Krafft et al. (2020).
[36]Salles, Evers, and Farisco (2020); Zhang and Dafoe (2019).
[37]Stuart and Kneer (2021).
[38]Bertsou (2021).
[39]Agrawal, Gans, and Goldfarb (2018).
[40]Kennedy et al. (2022); Kennedy et al. (2022).

https://doi.org/10.1017/bap.2023.35 Published online by Cambridge University Press

example in much of the literature laid out above.[41] It is also an area in which the private sector has been very active in developing tools of decision-makers.

Contrary to what legal scholars and ethicists have assumed, there is no significant decrease in blame for mistakes associated with a decision-maker agreeing with an algorithm–if anything, the amount of blame appears to slightly *increase*–though we are careful not to make too much of these substantively small effect sizes and they do not show up in our third sample (Study 2, discussed below). While we do find some increase in blame associated with disagreeing with an algorithm, this increase is not differentiable from agreeing with an algorithm or disagreeing with another source of advice. Again, these differences were small and did not show up in our third sample. Moreover, these experiments provide no evidence that the results are a product of demographics or general algorithm aversion.[42]

To explain this counter-intuitive result, we conducted a second pre-registered experiment (Study 2) on a new representative sample. In this sample we found no significant difference in blame from when the decision-maker was advised by an algorithm or when they decided on their own–a result that is not terribly surprising, given the relatively small effect sizes in Study 1. However, while there are not significant differences in blame between scenarios, on average, there are two specific factors that seem to be particularly important when it comes to agreement or disagreement with an algorithm's advice. We found some evidence that respondents' who are more trusting of experts generally place more blame on the decision-maker when they disagree with an algorithm's advice, while those less trusting of experts place less blame on the judge when they disagree with the advice. This helps explain why we do not find a larger magnifying effect of disagreeing with the algorithm's advice on blame—most of the sample was below the threshold of expert trust above which we see an increase in blame, suggesting that most of the sample receiving this treatment felt that rejection of the advice was justifiable and, perhaps, even appropriate. There was also some evidence that agreeing with the algorithm's advice results in respondents viewing the decision-maker as abdicating their duty to use their own judgment. In other words, perceptions that the judge was using the algorithm's judgment in place of their own increases, rather than deflects, the blame placed on the judge.

In sum, we find no evidence to support the concerns of ethicists and legal scholars, at least in the judicial context. These findings are consistent with more general findings on the role of expertise in the policy process[43] and suggest that the use of algorithms is not (yet?) a special case when it comes to expert advice.

Study 1

Building on the scenarios laid out in the legal, ethics and political science literature, this study looks at what happens when a judge uses an algorithm in making a decision about whether to jail or release a defendant, and compare this with the situation where they use their own judgment, advice from a human source, or the combined advice of an algorithm and a human source.[44]

Treatments

We asked respondents to read a brief scenario, very similar to that in Surden and developed in consultation with a law enforcement professional with 20+ years experience evaluating defendants for judges in three states.[45] The scenario involved a judge making a sentencing decision as to whether to grant probation to a defendant in a repeat drunk driving case. In all scenarios, the judge decides to release the defendant on probation and the defendant is subsequently involved in another drunk driving accident that kills a pedestrian.

[41]Angwin et al. (2016); Surden (2021); Fry (2018).
[42]Dietvorst, Simmons, and Massey (2015, 2018); Gogoll and Uhl (2018).
[43]Bertsou (2021).
[44]Kennedy, Waggoner, and Ward (2022); Surden (2021); Danaher (2016); Angwin et al. (2016); Dressel and Farid (2018).
[45]Surden (2021).

https://doi.org/10.1017/bap.2023.35 Published online by Cambridge University Press

Table 1. Summary of experimental treatments. This table gives a reference for the control condition, #1, and the 6 additional treatment conditions. For analysis, condition #1, where the judge makes the decision on their own is the baseline. Full description of the vignettes can be found in SI.1

	Treatment	Scenario
1	Control – Judge only	Judge grants probation to the defendant.
2	Algorithms agrees	A computer algorithm designed by computer scientists and criminal justice experts recommends probation. The judge agrees and grants probation to the defendant.
3	Algorithms disagree	A computer algorithm designed by computer scientists and criminal justice experts recommends imprisonment. The judge disagrees and grants probation to the defendant.
4	Human agrees	An experienced probation officer recommends probation. The judge agrees and grants probation to the defendant.
5	Human disagrees	An experienced probation officer recommends imprisonment. The judge disagrees and grants probation to the defendant.
6	Human and algorithms	An experienced probation officer, along with a computer algorithm designed by computer scientists and criminal justice experts, recommend probation. The judge agrees and grants probation to the defendant.
7	Human and algorithms disagree	An experienced probation officer, along with a computer algorithm designed by computer scientists and criminal justice experts, recommend imprisonment. The judge disagrees and grants probation to the defendant.

Each respondent was randomly assigned to one of four conditions: 1. judge decides with no additional input (control condition), 2. judge decides with assistance of algorithm, 3. judge decides with assistance of a probation officer, or 4. judge decides with assistance of a probation officer and algorithm. For the advice (non-control) conditions, respondents were also randomized between whether the advice was for probation and the judge agreed, or the advice was for imprisonment and the judge disagreed. The full set of treatments are outlined in Table 1. Respondents were then asked how much blame they placed on the actors involved in the scenario for the adverse outcome.[46] Our main variable of interest is the degree of blame placed on the judge in each advice condition. This measure of blame ranges from 1 ("none at all") to 10 ("a great deal").[47] It is re-scaled to range from 0 to 1, so effects can be interpreted as the proportion increase in the scale.

Although there was little clear empirical guidance from previous literature about what we expect in this particular experiment, we pre-registered the following hypotheses[48]:

> "Hypothesis 1: When an error occurs, a policymaker's (judge) reliance on advice from an algorithm will reduce the level of blame compared to relying on his/her judgment alone. Conversely, disregarding the algorithm's advice will increase the level of blame.
>
> Hypothesis 2: The reduction in blame from relying on an algorithm will be similar to that of reliance on advice from a trained bureaucrat.
>
> Hypothesis 3: When an error occurs, a policymaker's reliance on advice from a hybrid system involving both an algorithm and a trained bureaucrat will reduce the level of blame more than relying on either source alone."

Hypothesis 1 is drawn directly from the concerns of ethicists and legal scholars laid out above. Hypothesis 2 draws from Kennedy, Waggoner and Ward indicating individuals trust advice from automated systems as much or more than advice from other sources.[49] Hypothesis 3 is also drawn from

[46]See SI.1 for full formatting and wording of conditions.

[47]Stuart and Kneer (2021).

[48]Pre-registered on OSF (DOI 10.17605/OSF.IO/ZG37Q, see SI.5) on June 22, 2021. All protocols were reviewed and approved by IRB (STUDY00001247), see SI. 16. All data and code are available on Harvard's Dataverse (link to be included on publication).

[49]Kennedy, Waggoner and Ward (2022).

https://doi.org/10.1017/bap.2023.35 Published online by Cambridge University Press

Kennedy, Waggoner and Ward,[50] where hybrid systems, involving the judgment of both an expert and computer, are more trusted than either source alone.

Sample

The analysis was based on two demographically representative samples of the U.S. population. The first survey sample was collected in June 2021 using Lucid's Theorem platform. 1,500 respondents participated, of which 923 (62%) passed the attention checks and were utilized in the study.[51] Lucid draws from a range of survey panels and automatically assigns participants to match U.S. census demographics, and has been shown to replicate a range of well-established experimental results[52] and is also utilized by many of the most prominent companies in the survey industry to get data.[53] This data was originally collected just prior to the pre-registration, but the data was not inspected or analyzed until after the registration.[54] The second sample collected 1,842 respondents as part of a Time-sharing Experiments in the Social Sciences survey, which contracts with the National Opinion Research Center at the University of Chicago through their AmeriSpeak panel. This program has been used for a range of influential social science survey experiments.[55] Data from this sample was not received until four months after the pre-registration. The samples are pooled for analysis, and no significant differences in results were noted between studies.

Analysis

The responses were analyzed using OLS regression of the form:

$$\textbf{\textit{blame judge}} = \boldsymbol{\alpha} + \sum_{i=1}^{6} \boldsymbol{\beta_i}(\textbf{\textit{treatment}}_i)$$

where α is the overall intercept and β_i is the estimated slope coefficient (effect) of each of the i treatments, with the control condition omitted as a baseline.[56] Confidence intervals were calculated from 1,000 simulated draws from the distribution of the coefficient estimates.[57]

Results

The results suggest that, *irrespective of whether the judge agrees or disagrees with the algorithm*, the degree to which the public says the judge is to blame for the adverse outcome *increases slightly* compared to when the judge makes the decision without assistance. When the algorithm recommends probation and the judge agrees, respondents, on average, place 9% *more* blame upon the judge in light of the tragic outcome, compared to when the judge decides without assistance. When the algorithm recommends jail and the judge disagrees, respondents place about 6% more blame on the judge (Figure 1b). Figure 1a also shows that there is no tradeoff in blame, i.e. respondents did blame the algorithms for the mistake under the agreement condition, but this did not reduce the culpability assigned to the judge.

We note that these effects are small and, therefore, should be interpreted with some caution.[58] Cohen's d for the results ranged from 0.16 to 0.28 (see SI.8), which is in the negligible to medium-small range. While technically statistically significant in this study, such small effects are unlikely to have a strong impact on overall evaluations of the judge, and, as we note in SI.12, there was no impact found for the likelihood of voting for the judge in an election. Moreover, such small effects are less likely to regularly replicate, and, as we detail below, the significance level, though not the direction of the relationship,

[50]Kennedy, Waggoner and Ward (2022).

[51]See SI.2 & SI.3 for full description of samples and attention checks.

[52]Coppock and McClellan (2019).

[53]Enns and Rothschild (2022).

[54]This sample was originally planned as a pilot, but we received acceptance from TESS shortly after fielding and decided it was better to hold off analysis until after we received the TESS data.

[55]Mutz (2011).

[56]Gerber and Green (2012).

[57]King, Tomz, and Wittenberg (2000); Gelman and Hill (2006).

[58]Ioannidis (2005).

https://doi.org/10.1017/bap.2023.35 Published online by Cambridge University Press

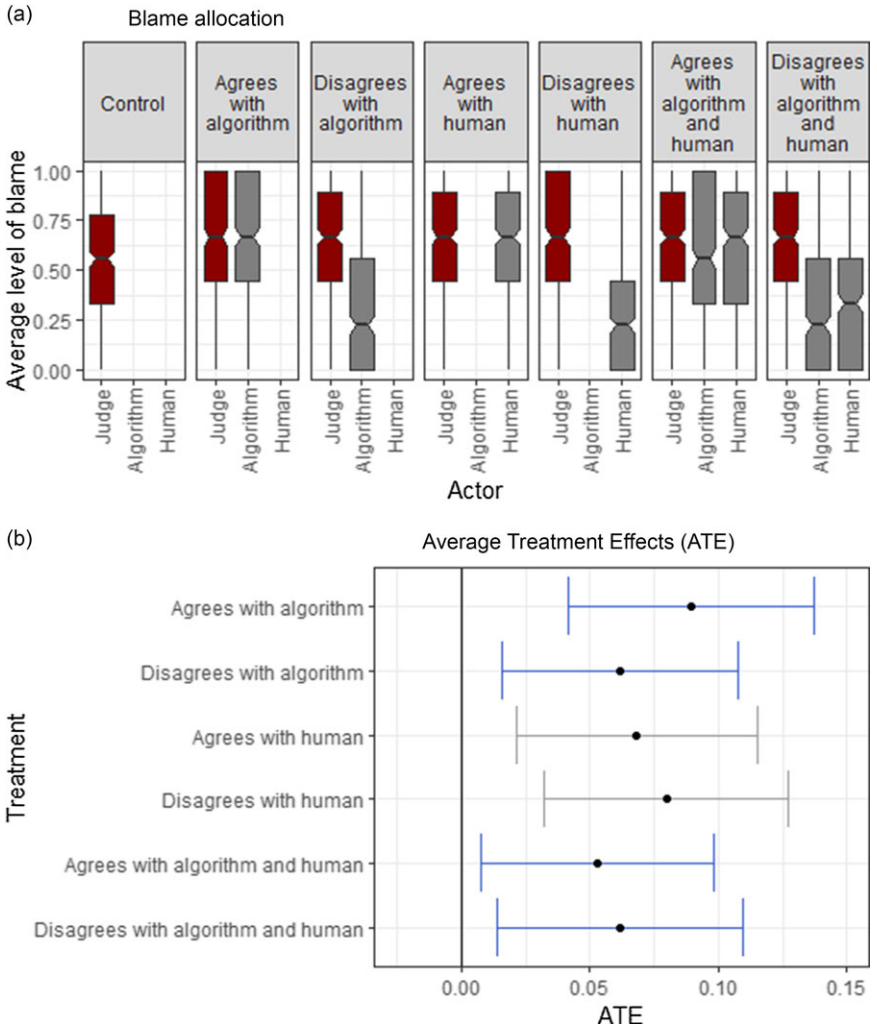

Figure 1. (a): Distribution of blame placed by respondent on each actor involved in decision for each treatment condition. Responses are re-scaled to between 0 and 1, such that differences can be interpreted as the proportion of the scale difference in average response. The main variable of interest, blame placed on the judge, is highlighted in crimson. Boxplots show the median values with a horizontal line with the boxes spanning the 25th and 75th percentiles and whiskers spanning the 1.5*IQR range. The notches show mark the interval of $(1.58 * IQR)/\sqrt{n}$, which is roughly equivalent to a 95% confidence interval. (b) Average Treatment Effects (ATE) for each treatment condition with 95% confidence intervals. ATE was calculated relative to the control condition. ATE ranged from 5.3% when the judge agreed with both the probation officer and algorithm to 9% when the judge agreed with the algorithm. Tabular results, details of the study, and a range of robustness checks are available in the [*SI.4, SI.9, SI.11 & SI.14*].

changes in Study 2. Interpreting these as null effects, however, still runs counter to popular expectation, and suggests agreement with an algorithm's advice will not buffer decision-makers from blame, nor will decision-makers necessarily receive more blame when they disagree with an algorithm.[59]

[59]A-priori power calculations are detailed in SI.6. We checked the retrospective power of the experiment as well (Gelman and Carlin 2014). Since we have little clear guidance on the expected size of effects from previous studies, we tested retrospective power for both the smallest and largest effect sizes from Study 1. For the smallest effect sizes (0.16 for the judge agreeing with the human and algorithm), retrospective power is 0.604 and the Type M error calculation is 1.285, meaning the results are, on average, an overestimation of about 29% of the hypothesized population effect. This is not terribly surprising, given how small this effect is. For the largest effect size (0.28 for the judge agreeing with the algorithm), the retrospective power is 0.962 and the Type M error calculation is 1.022, meaning that there is likely about a 2.2% overestimation. Type S error is 0 for both, meaning that there is no measurable chance of a significant relationship in the opposite direction in either situation.

https://doi.org/10.1017/bap.2023.35 Published online by Cambridge University Press

We also conduct hypothesis tests for the significance of differences between the treatment arms.[60] Table A9 in the SI shows that there are no significant differences in blame placed on the judge based on the source of advice. Contrary to what is posited by the theory literature above, whether the advice comes from an algorithm, human or a combination of the two, the differences are not statistically significant ($p > 0.1$). Thus, we fail to reject the null hypothesis that there is no differential impact based on the source of advice.

Why the concerns of ethicists and legal scholars are not borne out empirically is difficult to discern from this study. We first note that this does not appear to be a simple example of algorithm aversion,[61] as we observe similar increases in blame under the human and combined conditions in Figure 1b.

Figure 2 tests for effect moderation based on demographic characteristics[62] and generalized trust in algorithms.[63] Analysis for moderation is conducted using parallel within-treatment regression analysis to estimate the average treatment moderation effect (ATME),[64] since, unlike traditional treatment-by-covariate interactions, these values have a causal interpretation.[65] The process has the form

$$Y_i = \alpha_0 + \gamma_0 S_i + X_i'\beta_0 + \varepsilon_i \quad \forall i: T_i = 0$$

$$Y_i = \alpha_1 + \gamma_1 S_j + X_i'\beta_1 + \varepsilon_j \quad \forall j: T_j = 1$$

$$\delta_{PR} = \gamma_1 - \gamma_0$$

$$Var(\delta_{PR}) \sim Var(\gamma_0) + Var(\gamma_1)$$

where S_i is the potential mediator variable, X_i' is the set of other variables, T_i is the level of the treatment (in this case treated as present or absent, though it extends intuitively to our multi-treatment context), γ is the OLS coefficient for the potential mediator, and δ_{PR} is the ATME.

We found no evidence of moderation of treatment effects based on standard respondent demographics (gender, age, race, and education). There does seem to be an ideological dimension to respondents' pattern of blame, with respondents who identify more strongly with the Republican Party placing significantly more blame on the judge when they agree with the algorithm (ATME = 0.04, 95% confidence interval = [0.02, 0.06]), disagree with the algorithm (ATME = 0.03, 95% confidence interval = [0.01, 0.05]), or disagree with both the algorithm and human sources (ATME = 0.03, 95% confidence interval = [0.01, 0.05]). Greater trust in algorithms does reduce blame for the judge agreeing with the algorithm somewhat and produces the most promising results (ATME = −0.110, 95% confidence interval = [−0.23, 0.004]), it has no discernable effect when the judge disagrees with the algorithm (ATME = −0.025, 95% confidence interval = [−0.14, 0.09]) and produces opposite and insignificant results when the judge agrees with combined advice from a human and algorithm (ATME = 0.03, 95% confidence interval = [−0.09, 0.05]) or disagrees with this combined advice (ATME = 0.05, 95% confidence interval = [−0.07, 0.17]).[66]

Study 2

Given the surprising results from Study 1, we conducted a further study to both try replicating the results a third time and further explore why we received these results. We pre-registered three additional hypotheses.

[60]Gelman and Stern (2006).

[61]Dietvorst, Simmons, and Massey (2015, 2018); Gogoll and Uhl (2018); Dawes (1979).

[62]Hoff and Bashir (2015).

[63]Kennedy, Waggoner, and Ward (2022); Kim, Ferrin, and Rao (2008).

[64]Bansak (2021).

[65]More traditional analysis for heterogeneous treatment effects estimating conditional average treatment effects (CATEs) through treatment-by-covariate interactions was also tested and is reported in SI.11 (Gerber and Green 2012). The results are similar.

[66]See SI.11 for more detailed results.

https://doi.org/10.1017/bap.2023.35 Published online by Cambridge University Press

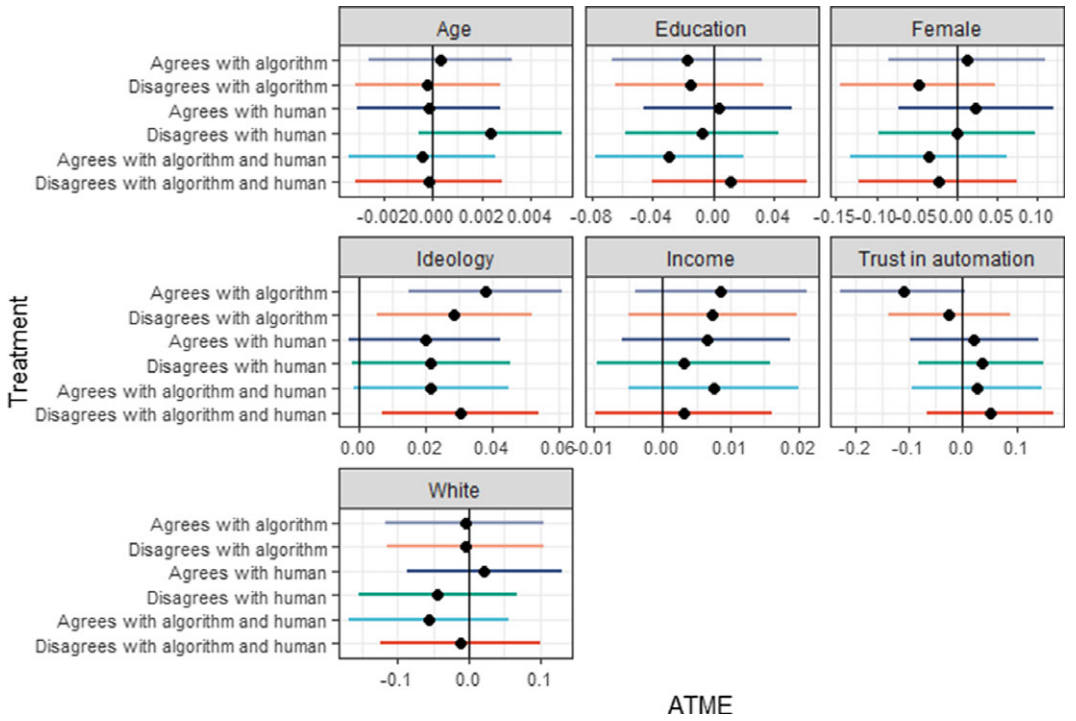

Figure 2. Moderation analysis with respondent characteristics. Values are ATME estimates, calculated using parallel within-treatment regressions for demographic and attitude characteristics. 95% confidence intervals are calculated using the variance formula noted in the methods section. Full tabular details of models are available in SI.11.

(1) Perceptions of blame are moderated by general trust in expert advice,[67] with those more trusting of experts showing decreased blame when the judge agrees with the algorithm and increased blame when the judge disagrees.[68] The basic idea behind this hypothesis is that algorithms, as suggested in Study 1, may be viewed similarly to other sources of expertise. For those with greater trust in expert advice generally, agreeing with an algorithm, even if it ends up being incorrect, will seem a natural and justifiable course of action. Conversely, for those who are skeptical of expert advice, the rejection of such expert advice, and reliance on one's own intuition, will be more likely viewed as a reasonable course of action.

(2) Use of advice changes expectations of accuracy, with an increase in blame under the advice conditions being a result of greater expectations that the judge *should have* gotten the ruling correct. Given that previous studies have found people to have relatively high behavioral trust in algorithms, especially in these situations,[69] it is possible that the use of algorithms increases the expectations of a correct ruling. Failure to meet these expectations may result in greater blame for an incorrect ruling.

(3) Mistakes made when using advice are, retrospectively, seen as an abdication of responsibility,[70] with those who either think the judge did not use their own judgment or should have relied more on their own judgment increasing the blame placed on the judge. The treatment, and the subsequent incorrect ruling, may be increasing perceptions that the judge

[67]Bertsou (2021); Merkley (2021).
[68]Surden (2021).
[69]Kennedy, Waggoner, and Ward (2022); Logg, Minson, and Moore (2019).
[70]Chesterman (2021).

https://doi.org/10.1017/bap.2023.35 Published online by Cambridge University Press

should have relied on their own judgment. Such retrospective biases, that an obvious course of action was not taken by a political decision-maker, are not unusual in other areas of politics, even when the actor has little to no control over the outcome.[71]

The sample for this study was collected in March 2022 from Luc.id with a sample size of 1,400 participants. The study was once again pre-registered with these hypotheses prior to being fielded.[72]

Treatments

The experimental design was nearly identical to the previous study with two notable exceptions. First, we included only three treatments from the prior study: (1) control, (2) judge agrees with the algorithm's recommendation, and (3) judge disagrees with the algorithm's recommendation. We did this both to focus on the most relevant treatments and to ensure appropriate statistical power for mediation analysis—conducting this for all of the treatments from Study 1 would have increased costs well beyond our research budget.[73] Second, we included three additional measures which we used to test for moderation and mediation effects. The first measure was an index of respondents' trust in experts, with respondents rating their degree of distrust for seven different types of experts.[74] The second measure assessed the post-treatment expectation that the judge should have made an accurate decision. We measured this by asking respondents on a scale from "never" to "always," how often they think the judge should have made the correct decision in the scenario. The third measure assesses the degree to which respondents believe that the judge is abdicating responsibility based on the treatment. This was measured using two post-treatment questions. The first assessed the degree to which the respondent thought the decision reflected the judge's evaluation versus that of the advice-giver. The second assessed the degree to which the respondent thought the decision *should have* reflected the evaluation of the judge or the advice-giver.[75]

Sample

Analysis is based on a demographically representative sample of the U.S. population gathered from Lucid. Of the 3,656 respondents that participated, 1,423 (40%) passed the attention check and participated in the study.

Analysis

Analysis for moderation based on trust in experts was done using the same within-treatment parallel-regression method discussed in Study 1 for estimation of the ATME. Analysis for causal mediation followed the protocol developed by Imai et al. and Imai, Keele, and Tingley,[76] and involved the estimation of two equations

$$M_i = \alpha_1 + \beta_1 T_i + X_i'\zeta_1 + e_{i1}$$

$$Y_i = \alpha_2 + \beta_2 T_i \ + \gamma M_i + X_i'\zeta_2 + e_{i2}$$

Y_i is the outcome of interest, T_i is the treatment, M_i is the potential mediator, X_i is a set of control variables, and α_1 and α_2 are intercepts. Since T is randomly assigned, it is independent of the error terms, e_{i1}, $e_{i2} \perp\!\!\!\perp$ X. The first equation estimates the effect of the treatment on the mediator. The second equation simultaneously tests the effect of the mediator and the treatment on the outcome. The average

[71]Achen and Bartels (2017).

[72]More details on the sample and the pre-registration can be found in SI.13.

[73]See SI.6.

[74]Merkley (2021).

[75]Pre-registered on OSF (DOI 10.17605/OSF.IO/ZG37Q, SI.13) and approved by IRB (MOD00004154) see SI.16. Full wording and formatting in SI.14.

[76]Imai et al. (2011); Imai, Keele, and Tingley (2010).

https://doi.org/10.1017/bap.2023.35 Published online by Cambridge University Press

causal mediation effect (ACME) is estimated by calculating $\beta_1 * \gamma$. Confidence intervals for this value were estimated using nonparametric bootstrapping.[77]

Results

In Figure 3 we replicated the analysis conducted above on the direct effect of the treatments. Interestingly, while the direction of the treatment effect remained the same, and still contradicted the expectations from previous literature about increasing blame when the judge disagreed with the algorithm and decreasing blame when the judge agreed with the algorithm, the magnitude of the results is lower and does not reach standard levels of statistical significance ($p > 0.05$). While these results lack statistical significance, we should note the general consistency with previous results and that this lack of statistical significance does not necessarily prevent successful analysis of moderation or mediation.[78] Moreover, as noted above, such issues are not entirely surprising, given the relatively small effect size in the previous experiments. The results still provide relatively strong evidence against the hypotheses of the theoretical literature. Using the test developed by Gelman and Carlin,[79] the probability of Type S error–i.e., the probability that we would see a significant effect in the opposite direction–is 1.1% for when the judge agrees with the algorithm and 0.8% for the judge disagreeing with the algorithm. The results still refute the concerns laid out by previous scholars, although with weaker evidence, and there is no evidence of a statistically significant difference between agreeing or disagreeing with the algorithm ($p > 0.1$).

Figure 4 shows evidence that trust in experts moderates the response of a judge disagreeing with an algorithm. Figure 4a–c show the regression line for trust in experts in each treatment condition. In the control condition, the effect is nearly flat–trust in experts does not affect blame when the judge is making the decision on their own. When the judge agrees with the algorithm, there is a slight, but insignificant, decrease in blame. Finally, when the judge disagrees with the algorithm, there is a significant ($p < 0.001$) and positive relationship between blame and the amount of trust the respondent places in experts. Comparing 4A and 4C, it is notable that the blame placed on the judge only exceeds that of the control condition at the highest levels of trust in experts, encompassing a minority of our sample. For those less trusting of expert advice, the rejection of advice from an expert appears to be seen as justified. Figure 4d summarizes these results. Respondents with the highest level of trust in experts place about 33% more blame on the judge than those with the least trust in experts, when the judge disagrees with the algorithm. Conversely, they place about 17% less blame on the judge when the judge agrees with the algorithm.[80]

We find little evidence that changes in expectations mediate the amount of blame. Figure 5 shows these results. Figure 5b and d show that individuals with higher accuracy expectations do place significantly more blame on the judge for their decision ($p < 0.001$). However, there is no significant impact in 5A between agreeing with the algorithm and expectations that the judge should have arrived at a correct decision, and 5C shows that the relationship between disagreeing with the algorithm and the expectation of accuracy is *negative*.[81] Expectations may be an important explanation of blame generally, but they do not appear to link use of advice and greater blame.

There is some evidence for Hypothesis 3–perceived reliance on advice is viewed as an abdication of the judge's responsibility, and this increases blame (i.e., the judge should have figured it out on their own). Both post-hoc evaluations of the relative role of the judge and the algorithm and assessments of which should have had greater weight have significant average causal mediation effects (ACMEs) when the judge agrees with the algorithm, but are unrelated with the judge disagreeing with the algorithm.[82] Figure 6a shows that there is a significant relationship ($p < 0.001$) between the judge agreeing with the

[77]Imai et al. (2011). See SI.14 for more details.
[78]Bollen (1989); Hayes (2017).
[79]Gelman and Carlin (2014).
[80]SI.14, Table A14.
[81]SI.14, Tables A15 & A16.
[82]SI.14, Tables A17–A20.

https://doi.org/10.1017/bap.2023.35 Published online by Cambridge University Press

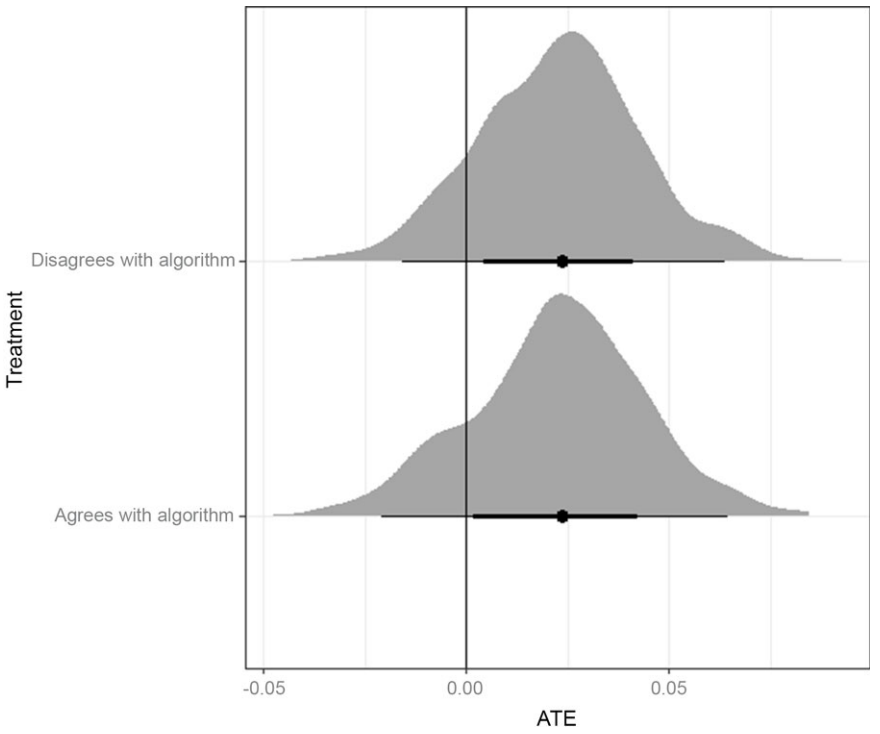

Figure 3. Average Treatment Effects (ATE) for algorithm treatments in Study 2 with 66% and 95% confidence intervals and distribution of estimated coefficients. Dots indicate mean ATE across 1,000 coefficient simulations, with the 66% confidence interval indicated by the bold line and the 95% confidence interval by the narrow line. Distributions show the full distribution of the 1,000 simulations.

algorithm and respondents suggesting that the judge should have been more hands-on in making the judgment. Similarly, Figure 6b shows the significant relationship ($p < 0.001$) between this attitude that the judge should have more control over the decision and the amount of blame placed on the judge for the decision. Figure 6d summarizes this relationship, showing that about 66% of the effect of agreeing with the algorithm is mediated by respondents saying that the judge should have relied more on their own judgment in these situations. This is also borne out in the comments left by respondents, who regularly emphasized the importance of the judge exercising their own judgment (e.g., "they hold the office").[83] There is certainly some level of retrospective bias in this result, but such evaluations are not uncommon in assessing the performance of public officials, even for circumstances beyond their control.[84] There are, however, also some reasons for being cautious about the mediation results. Mediation analysis, in general, is criticized by some scholars for being vulnerable to confounding.[85] There are also some sample-specific issues and indications that the observed ACMEs are not overly robust.[86] Nevertheless, this does suggest an interesting path for future inquiry, and, at a minimum, provides significant evidence that use of algorithms risks perceptions of dependence and re-assessments of the role they should play in decision-making when inevitable mistakes are made. Indeed, some of this could account for the quick turn of public sentiment against private company generated algorithms like COMPAS and PredPol in recent years, as the problems of accuracy and bias have become more apparent. A number of jurisdictions have dropped their contracts with the associated companies in recent years or decided not to sign new contracts under increased public scrutiny.[87]

[83]SI.15.
[84]Achen and Bartels (2017).
[85]Bullock, Green, and Ha (2010); D. P. Green, Ha, and Bullock (2010).
[86]See SI.14, Figures A5–A6.
[87]Mehrotra (2021).

https://doi.org/10.1017/bap.2023.35 Published online by Cambridge University Press

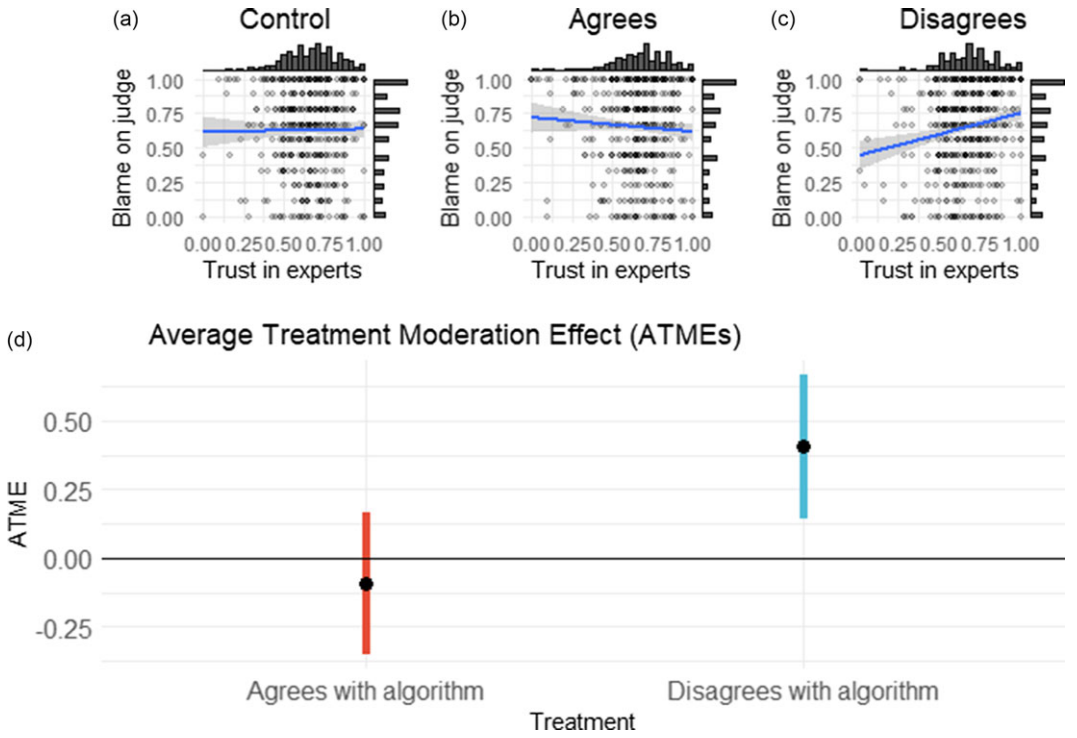

Figure 4. Moderating effect of trust in experts on blame. Figure 5a shows the distribution of respondents' trust in experts and blame on the judge for the control condition, with the OLS regression line and 95% confidence intervals, 5b shows the same information for the treatment condition in which the judge agrees with the algorithm, and 5c shows this information for the condition in which the judge disagrees with the algorithm. Figure 5d shows the estimated ATME, with 95% confidence intervals for both the agreement and disagreement treatments. Full tabular results are available in SI.14.

Discussion

These results have profound implications for policymakers as the use of algorithms in public decision-making grows. First, while more work is needed to tease out the extent to which the results above hold in different circumstances, there is a clear indication that public decision-makers will be held accountable for their decisions and any adverse consequences of those decisions. We find no significant evidence that use of algorithms for advice decreases blame on decision-makers when they agree with the advice or uniquely increases blame when they disagree with the advice. Nor do we find that algorithms hold any special place as a source of advice relative to human sources. This is consistent with other work on experts in the policy implementation process, which finds that the public believes experts should assist in *how* a decision is implemented, not *what* decisions are made.[88] In addition, these results are consistent with studies suggesting the public is increasingly resistant to expertise as a justification for policy action.[89] If a policymaker is contemplating use of an algorithm to assist in decision-making, the basis for that decision will need to be made based on efficacy, not whether it will shield them from criticism or resolve the anxiety of incorrectly assessing risk. From the perspective of ethicists and legal scholars, the results may be something of a double-edged sword. While we found none of the issues with democratic accountability about which some worry, the results also suggest that algorithms, usually implemented under the label of "evidence-based practices," will not provide a shortcut for addressing the mass incarceration problem in the U.S.

Second, to the extent that a policymaker believes an algorithm will result in better decision-making, the logic underlying the algorithm's decision process needs to be explainable. Legitimacy in the

[88]Bertsou (2021).
[89]Merkley (2021).

https://doi.org/10.1017/bap.2023.35 Published online by Cambridge University Press

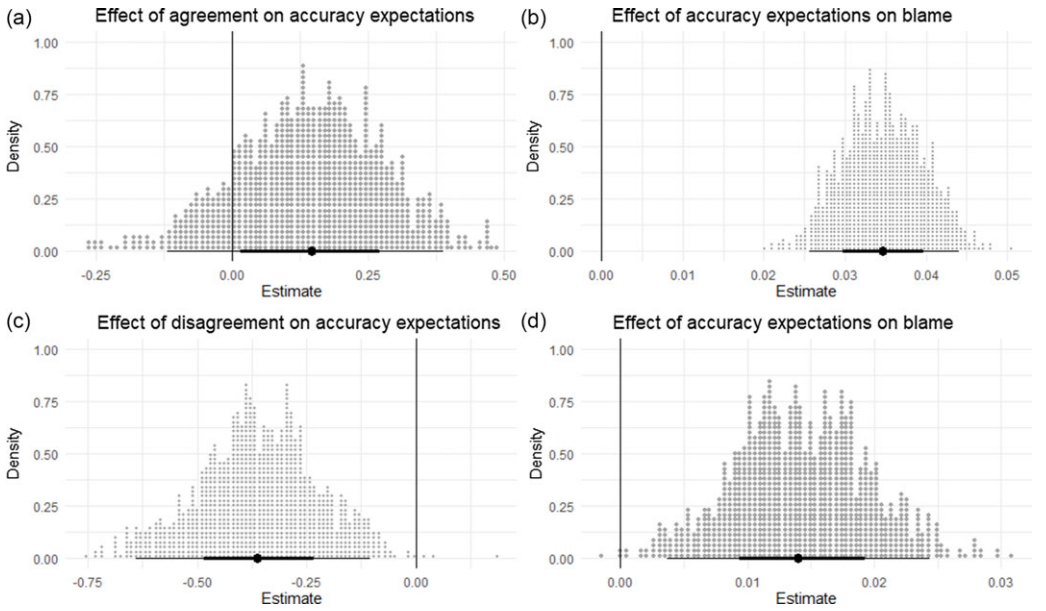

Figure 5. Tests for mediation based on changes in respondents' expectations. Plots show the distribution of coefficient estimates from 1,000 simulations from the coefficient distributions. The dots at the bottom indicate the point estimate from the model, with the bold bar showing the 66% confidence interval and the non-bold bar showing the 95% confidence interval. Figure 3a plots the estimates of the effect of agreement on the expected accuracy of the judge's decision. Figure 3b shows the estimates of the effect of accuracy expectations on the blame placed on the judge in the context of agreement. Figure 3c shows the estimates of the effect of disagreeing with the algorithm on expectations of accuracy. Figure 3d shows the estimates of the effect of expected accuracy on the blame placed on the judge in the context of disagreement. Full tabular results are available in SI.14.

government sphere, and especially in the legal area, requires justification for actions.[90] This addresses the concern raised in Study 2 about respondents viewing concurring with the algorithm as an abdication of the judge's responsibilities to use their own judgment. Judges must be able to explain why the algorithm influenced their decision, beyond simply parroting the final analysis of the algorithm, as was infamously done by the judge in the Loomis v. Wisconsin case. This can be problematic with the modern architecture of computerized decision aids, which can rely on complicated machine learning architectures that are difficult to explain in natural language.[91] Being able to explain the complicated mathematics of a statistical model is, however, not really what is being demanded of these decision aids. Using the tools of behavioral and counterfactual analysis, algorithms have a unique ability to answer questions about how specific factors would change the results from the algorithm. Even if policymakers cannot explain the complex weighting of the machine learning model used to generate the forecasts, this provides direct insight into the impacts of sensitive factors like race and income on recommendations.[92] This allows for decision-makers to justify the basis for the decision and address accusations of unfairness. Auditing tools leverage this fact to identify potential ethical issues with algorithms,[93] and such tools may also prove useful in elucidating the decision process and providing the required explanations and justifications for policy decisions. State-of-the-art documentation and auditing is, however, still relatively rare. Legislation currently being considered in Congress (H.R. 6580; S. 3572), requiring auditing and impact assessments for some entities may assist in this process.

Finally, this research offers fertile ground for further research exploring the perceived role and implementation of algorithms in public policy and politics, as well as how these are framed for the

[90]Chesterman (2021).
[91]Rahwan et al. (2019).
[92]Rahwan et al. (2019); Cowgill and Tucker (2017).
[93]Saleiro et al. (2018).

https://doi.org/10.1017/bap.2023.35 Published online by Cambridge University Press

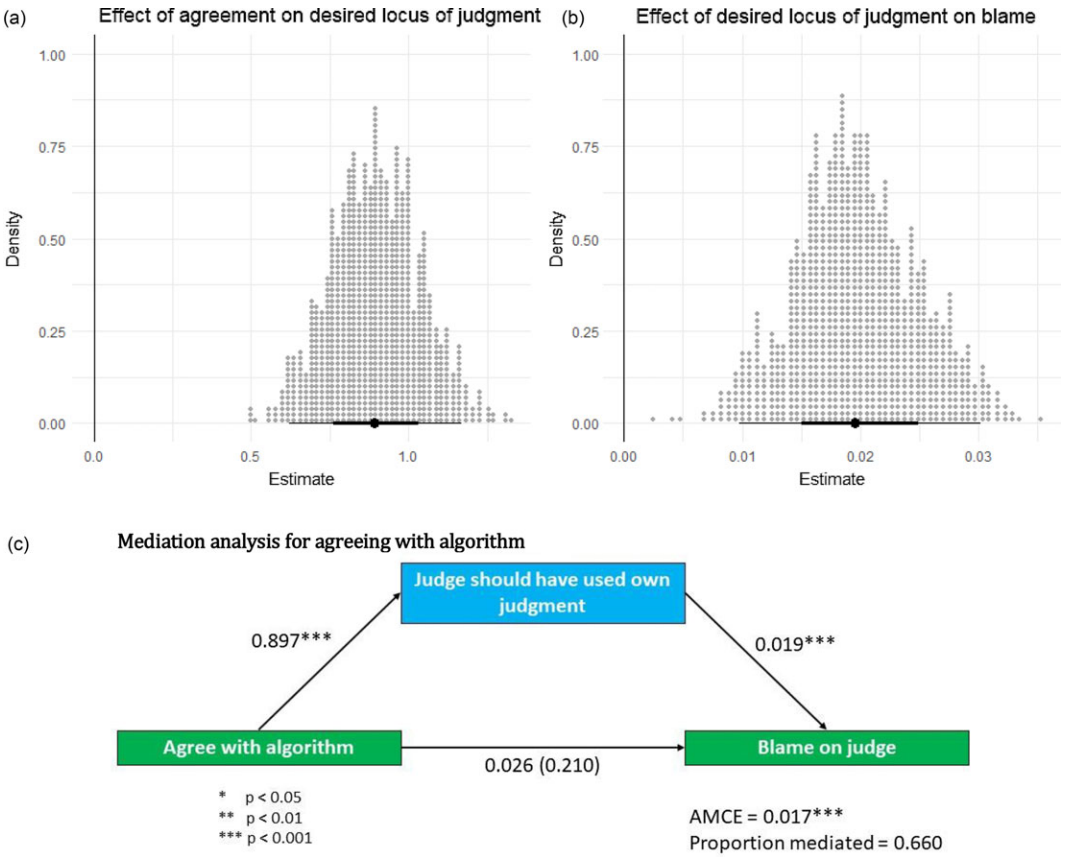

Figure 6. (a): Treatment effect of disagreeing with an algorithm moderated by respondent's trust in experts. Boxplots show the distribution of estimates for the moderated causal effect from 1,000 simulated draws from the coefficient distribution. (b): Treatment effect of agreeing with an algorithm mediated by respondents' evaluation of whose judgment should be used in making such decisions. The total indirect effect (average causal mediation effect (ACME)) is 0.017, with a 95% confidence interval, calculated from 1,000 bootstrapped samples, of [0.007, 0.030] ($p < 0.001$). About 66% of the relationship between agreeing with the algorithm and the increased blame on the judge is explained by this mediated effect [*SI.14*].

public and the effects of this framing. More specifically, we believe it is vital to continue to explore how the implementation of computer algorithms may exacerbate distrust among both the public and elites, affecting potential policy implementation and subsequent success. Moreover, we believe our results highlight the need to further explore the role of algorithms in a similar context to human experts, as the type of role the algorithm is designed to serve may impact the degree of support and trust from the public. Other contexts may also open the possibility of exploring more complete counterfactuals. In this study, the only counterfactual being evaluated is when the judge releases someone who goes on to commit a new crime. We could not evaluate whether the results were similar if someone was jailed unnecessarily, since whether they would have committed a crime if released cannot be realized. Yet, there is some evidence that the public does not evaluate false positives and false negatives in the same way,[94] and the implementation of these algorithms is inherently tied to these calculations of relative risk.[95] Other scenarios may help provide a more complete picture of how different types of errors affect blame on public officials. Finally, we note that, while the authors of this study attempted to produce experimental treatments that were relatively minimal while maintaining realism of the scenarios, consistency with theory, and clarity of the treatment, there were differences between how the treatment conditions and the control condition were worded in order to ensure fidelity to theory and clarity of the

[94]Dressel and Farid (2018); Kennedy, Waggoner, and Ward (2022).
[95]Kennedy (2015).

https://doi.org/10.1017/bap.2023.35 Published online by Cambridge University Press

treatment. This raises the possibility that responses to advice conditions might be highly sensitive to framing effects.[96] While still suggesting that the concerns of theorists that algorithms have unique and significant influence, in themselves, are likely overblown at present, future scholars may find a productive area of research in exploring how the framing of expert and/or algorithm advice affects public perceptions, or even exploring how public perceptions would be shaped by more intensive interventions like public deliberation.[97] Our study provides a baseline on which these studies can proceed.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/bap.2023.35

References

Achen, Christopher H., and Larry M. Bartels. 2017. *Democracy for Realists: Why Elections Do Not Produce Responsive Government (Princeton Studies in Political Behavior, 4)*. Revised edition. Princeton, New Jersey: Princeton University Press.

Albright, Alex. 2023. "The Hidden Effect of Algorithmic Recommendations." Working Paper: Federal Reserve Bank of Minneapolis.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2018. *Prediction Machines: The Simple Economics of Artificial Intelligence*. Brighton, MA: Harvard Business Review Press.

Angwin, Julia, Jeff Larson, Lauren Kirchner, and Surya Mattu. 2016. "Machine Bias — ProPublica." *ProPublica*. May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Bansak, Kirk. 2021. "Estimating Causal Moderation Effects with Randomized Treatments and Non-randomized Moderators." *Journal of the Royal Statistical Society. Series A* 184 (1): 65–86.

Benjamin, Ruha. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*. First edition. Indianapolis, IN: Polity.

Berman, Ari. 2015. "How the 2000 Election in Florida Led to a New Wave of Voter Disenfranchisement." *The Nation*, July 28, 2015. https://www.thenation.com/article/how-the-2000-election-in-florida-led-to-a-new-wave-of-voter-disenfranchisement/.

Bertsou, Eri. 2021. "Bring in the Experts? Citizen Preferences for Independent Experts in Political Decision-making Processes." *European Journal of Political Research*, no. 1475–6765.12448 (April). https://doi.org/10.1111/1475-6765.12448.

Bollen. 1989. *Structural Equations with Latent Variables*. First edition. Indianapolis, IN: Wiley-Interscience.

Boudreau, Cheryl, and Mathew D. McCubbins. 2010. "The Blind Leading the Blind: Who Gets Polling Information and Does It Improve Decisions?" *The Journal of Politics* 72 (2): 513–527.

Brayne, Sarah. 2020. *Predict and Surveil: Data, Discretion, and the Future of Policing*. Oxford, UK: Oxford University Press.

Bullock, John G., Donald P. Green, and Shang E. Ha. 2010. "Yes, but What's the Mechanism? (Don't Expect an Easy Answer)." *Journal of Personality and Social Psychology* 98 (4): 550–558.

Chen, Jowei, and Jonathan Rodden. 2015. "Cutting through the Thicket: Redistricting Simulations and the Detection of Partisan Gerrymanders." *Election Law Journal* 14 (4): 331–345.

Chesterman, Simon. 2021. *We, the Robots?: Regulating Artificial Intelligence and the Limits of the Law*. Cambridge, UK: Cambridge University Press.

Coppock, Alexander, and Oliver A. McClellan. 2019. "Validating the Demographic, Political, Psychological, and Experimental Results Obtained from a New Source of Online Survey Respondents." *Research & Politics* 6 (1): 2053168018822174.

Cowgill, Bo, and Catherine Tucker. 2017. "Algorithmic Bias: A Counterfactual Perspective." In Workshop on Trustworthy Algorithmic Decision-Making. NSF Trustworthy Algorithms, Arlington, VA.

Danaher, John. 2016. "The Threat of Algocracy: Reality, Resistance and Accommodation." *Philosophy & Technology* 29 (3): 245–268.

Dawes, Robyn M. 1979. "The Robust Beauty of Improper Linear Models in Decision Making." *The American Psychologist* 34 (7): 571.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. 2015. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err." *Journal of Experimental Psychology. General* 144 (1): 114–126.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. 2018. "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them." *Management Science* 64 (3): 1155–1170.

Dressel, Julia, and Hany Farid. 2018. "The Accuracy, Fairness, and Limits of Predicting Recidivism." *Science Advances* 4 (1): eaao5580.

Druckman, James N. 2001a. "Using Credible Advice to Overcome Framing Effects." *The Journal of Law, Economics, and Organization* 17 (1): 62–82.

Druckman, James N. 2001b. "On the Limits of Framing Effects: Who Can Frame?" *The Journal of Politics* 63 (4): 1041–1066.

Enns, Peter K., and Jake Rothschild. 2022. "Do You Know Where Your Survey Data Come From?" *Medium*, 2022. https://medium.com/3streams/surveys-3ec95995dde2.

[96]Druckman (2001b, 2001a).

[97]Fishkin (2011).

Fishkin, James S. 2011. *When the People Speak: Deliberative Democracy and Public Consultation*. Oxford, UK: Oxford University Press.
Fry, Hannah. 2018. *Hello World: Being Human in the Age of Algorithms*: Reprint edition. New York, New York: W. W. Norton & Company.
Gelman, Andrew, and John Carlin. 2014. "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 9 (6): 641–651.
Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. First edition. Cambridge, UK: Cambridge University Press.
Gelman, Andrew, and Hal Stern. 2006. "The Difference Between 'Significant' and 'Not Significant' Is Not Itself Statistically Significant." *The American Statistician* 60 (4): 328–331.
Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. Illustrated edition. New York, New York: W. W. Norton & Company.
Gill, Tripat. 2020. "Blame It on the Self-Driving Car: How Autonomous Vehicles Can Alter Consumer Morality." *The Journal of Consumer Research* 47 (2): 272–291.
Gogoll, Jan, and Matthias Uhl. 2018. "Rage against the Machine: Automation in the Moral Domain." *Journal of Behavioral and Experimental Economics* 74: 97–103.
Goodman, Ellen P. 2021. "Smart City Ethics: How 'Smart' Challenges Democratic Governance." In *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das, 823–839. Oxford, UK: Oxford University Press.
Green, Ben, and Jascha Franklin-Hodge. 2020. *The Smart Enough City: Putting Technology in Its Place to Reclaim Our Urban Future (Strong Ideas)*. Cambridge, MA: The MIT Press.
Green, Donald P., Shang E. Ha, and John G. Bullock. 2010. "Enough Already about 'Black Box' Experiments: Studying Mediation Is More Difficult than Most Scholars Suppose." *The Annals of the American Academy of Political and Social Science* 628 (1): 200–208.
Hannah-Moffat, Kelly. 2015. "Partiality, Transparency, and Just Decisions." *Federal Sentencing Reporter* 27 (4): 244–247.
Hayes, Andrew F. 2017. *Introduction to Mediation, Moderation, and Conditional Process Analysis, Second Edition: A Regression-Based Approach (Methodology in the Social Sciences)*: Second edition. New York, NY: The Guilford Press.
Hoff, Kevin Anthony, and Masooda Bashir. 2015. "Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust." *Human Factors* 57 (3): 407–434.
Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. "A General Approach to Causal Mediation Analysis." *Psychological Methods* 15 (4): 309–334.
Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2011. "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies." *The American Political Science Review* 105 (4): 765–789.
Ingraham, Christopher. 2017. "This Anti-Voter-Fraud Program Gets It Wrong over 99 Percent of the Time. The GOP Wants to Take It Nationwide." *The Washington Post*, July 20, 2017. https://www.washingtonpost.com/news/wonk/wp/2017/07/20/this-anti-voter-fraud-program-gets-it-wrong-over-99-of-the-time-the-gop-wants-to-take-it-nationwide/.
Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8): e124.
Kennedy, Ryan. 2015. "Making Useful Conflict Predictions: Methods for Addressing Skewed Classes and Implementing Cost-Sensitive Learning in the Study of State Failure." *Journal of Peace Research* 52 (5): 649–664.
Kennedy, Ryan, Philip D. Waggoner, and Matthew Ward. 2022. "Trust in Public Policy Algorithms." *The Journal of Politics* 84 (2): 1132–1148.
Kim, Dan J., Donald L. Ferrin, and H. Raghav Rao. 2008. "A Trust-Based Consumer Decision-Making Model in Electronic Commerce: The Role of Trust, Perceived Risk, and Their Antecedents." *Decision Support Systems* 44 (2): 544–564.
King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44 (2): 347–361.
Krafft, P. M., Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo. 2020. "Defining AI in Policy versus Practice." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 72–78. AIES '20. New York, NY, USA: Association for Computing Machinery.
Logg, Jennifer M. 2016. *When Do People Rely on Algorithms?* Berkeley: University of California.
Logg, Jennifer M., Julia A. Minson, and Don A. Moore. 2019. "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment." *Organizational Behavior and Human Decision Processes* 151 (March): 90–103.
Mehrotra, Dhruv. 2021. "Crime Prediction Software Promised to Be Free of Biases. New Data Shows It Perpetuates Them – The Markup." https://themarkup.org/prediction-bias/2021/12/02/crime-prediction-software-promised-to-be-free-of-biases-new-data-shows-it-perpetuates-them.
Merkley, Eric. 2021. "Anti-Intellectualism, Populism, and Motivated Resistance to Expert Consensus." *Public Opinion Quarterly* 84 (1): 24–48.
Merkley, Eric, and Peter John Loewen. 2021. "Anti-Intellectualism and the Mass Public's Response to the COVID-19 Pandemic." *Nature Human Behaviour* 5 (6): 706–715.
Molnar, Petra. 2021. "AI and Migration Management." In *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das, 769–787. Oxford, UK: Oxford University Press.
Mutz, Diana C. 2011. *Population-Based Survey Experiments*. Princeton, New Jersey: Princeton University Press.
O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Reprint edition. Crown.

https://doi.org/10.1017/bap.2023.35 Published online by Cambridge University Press

Ozer, Adam. 2020. "Well, You're the Expert: How Signals of Source Expertise Help Mitigate Partisan Bias." *Journal of Elections, Public Opinion and Parties*, April, 1–21.

Perakslis, Christine. 2020. "Exposing Technowashing: To Mitigate Technosocial Inequalities [Last Word]." *IEEE Technology and Society Magazine* 39 (1): 88.

Rahwan, Iyad, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, Matthew O. Jackson, Nicholas R. Jennings, Ece Kamar, Isabel M. Kloumann, Hugo Larochelle, David Lazer, Richard McElreath, Alan Mislove, David C. Parkes, Alex 'Sandy' Pentland, Margaret E. Roberts, Azim Shariff, Joshua B. Tenenbaum, and Michael Wellman. 2019. "Machine Behaviour." *Nature* 568 (7753): 477–486.

Ramakrishnan, Naren, Patrick Butler, Sathappan Muthiah, Nathan Self, Rupinder Khandpur, Parang Saraf, Wei Wang, Jose Cadena, Anil Vullikanti, Gizem Korkmaz, Chris Kuhlman, Achla Marathe, Liang Zhao, Ting Hua, Feng Chen, Chang Tien Lu, Bert Huang, Aravind Srinivasan, Khoa Trinh, Lise Getoor, Graham Katz, Andy Doyle, Chris Ackermann, Ilya Zavorin, Jim Ford, Kristen Summers, Youssef Fayed, Jaime Arredondo, Dipak Gupta, and David Mares. 2014. "'Beating the News' with EMBERS: Forecasting Civil Unrest Using Open Source Indicators." In *Proceedings of KDD*, 1799–1808. ACM.

Saleiro, Pedro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. 2018. "Aequitas: A Bias and Fairness Audit Toolkit." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1811.05577.

Salles, Arleen, Kathinka Evers, and Michele Farisco. 2020. "Anthropomorphism in AI." *AJOB Neuroscience* 11 (2): 88–95.

Scharre, Paul. 2018. *Army of None: Autonomous Weapons and the Future of War*. First edition. Cambridge, UK: W. W. Norton & Company.

Shane, Janelle. 2019. *You Look Like a Thing and I Love You: How Artificial Intelligence Works and Why It's Making the World a Weirder Place*. Illustrated edition. Victoria Embankment, London: Voracious.

Stuart, Michael T., and Markus Kneer. 2021. "Guilty Artificial Minds: Folk Attributions of Mens Rea and Culpability to Artificially Intelligent Agents." *Proceedings of the ACM on Human-Computer Interaction* 363 (CSCW2): 1–27.

Surden, Harry. 2021. "Ethics of AI in Law: Basic Questions." In *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das, 719–736. Oxford, UK: Oxford University Press.

Zeide, Elana. 2021. "Robot Teaching, Pedagogy, and Policy." In *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das, 790–803. Oxford, UK: Oxford University Press.

Zhang, Baobao, and Allan Dafoe. 2019. "Artificial Intelligence: American Attitudes and Trends." https://doi.org/10.2139/ssrn.3312874.

Zuiderwijk, Anneke, Yu-Che Chen, and Fadi Salem. 2021. "Implications of the Use of Artificial Intelligence in Public Governance: A Systematic Literature Review and a Research Agenda." *Government Information Quarterly* 38 (3): 101577.

Zwald, Zachary, Ryan Kennedy, and Adam Ozer. 2021. *Human Trust in Autonomous Machines: Testing the New U.S. Approach to War*: SSRN.

Cite this article: Ozer AL, Waggoner PD, and Kennedy R (2024). The Paradox of Algorithms and Blame on Public Decision-makers. *Business and Politics* **26**, 200–217. https://doi.org/10.1017/bap.2023.35