**Supporting Information:**

**The Paradox of Algorithms and Blame on Public Decision-makers**

**Table of Contents**

SI.1. Experimental Instrument Design 1-9

SI.2. Attention Checks 10-11

SI.3. Sample Characteristics 12-13

SI.4. Balance Tests 14-16

SI.5. Preregistration 17

SI.6. Power Analysis 18

SI.7. Separate Sample Results 19-20

SI.8. Tabular Results and Standardized Effect Size 21-22

SI.9. Results with Control Variables 23-25

SI.10. Initial Tests for Moderating Effects 26-28

SI.11. Results for Election Question 29-30

SI.12. Results for Blame on Algorithm and/or Human Advisor 31-32

SI.13. Moderation and Mediation Analysis 33-51

SI.14. Comments from Respondents 52-59

SI.15. IRB Review and Ethics Compliance 60-64

**SI.1: Experimental Instrument Design**

To test the public’s allocation of blame for an adverse outcome in light of the use of an algorithmic decision aid, we designed a vignette experiment [(Mutz 2011)](https://paperpile.com/c/yikrbz/vDQeR). Participants are shown one of seven vignettes, one control condition and six manipulations of that scenario, which are randomly chosen using a pseudo-random number generator.[[1]](#footnote-0)

The first challenge in designing the vignettes is to determine a setting. We decided on a criminal justice setting for a few reasons: (1) it is a well-established area for algorithmic decision aid studies [(Angwin et al. 2016; Dressel and Farid 2018; Surden 2021)](https://paperpile.com/c/yikrbz/Enw8e%2BuU13Q%2BZQnKK); (2) it sidesteps some of the issues of using other branches of government, particularly the expectation of partisan cues [(Arceneaux and Vander Wielen 2017)](https://paperpile.com/c/yikrbz/yt8HB), since it is not unusual for states to have non-partisan methods for judicial selection; (3) it also avoids uncertainty about what constitutes a mistaken decision – just about every respondent can agree that releasing a person who ends up committing a violent offense is a mistake, a consensus that is relatively rare in public policy; (4) it is an area where respondents feel comfortable making independent judgements and are, therefore, more likely to see a mistaken decision as an error (even if they would have made the same decision) [(Dressel and Farid 2018; Kennedy, Waggoner, and Ward 2022)](https://paperpile.com/c/yikrbz/zsXoH%2BEnw8e); and (5) it builds off of previous research suggesting that algorithms in this particular context have a high degree of behavioral trust, even more so than judges [(Kennedy, Waggoner, and Ward 2022)](https://paperpile.com/c/yikrbz/zsXoH).[[2]](#footnote-1)

Moreover, while we were not aware of it at the time the study was designed, the experimental setup for this study is very similar to a scenario cited as problematic by [(Surden 2021)](https://paperpile.com/c/yikrbz/ZQnKK). We mentioned this in the main paper, and it is worth quoting his setup at length to see why he, and other scholars, worry that the incentive structure for policy-makers vis-a-vis abiding by an algorithm’s recommendation is such that it may override their independent judgment:

 “[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.”

[(Surden 2021)](https://paperpile.com/c/yikrbz/ZQnKK) 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; (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.”

In designing the scenario, we consulted with a probation and parole officer with more than twenty years of experience working in three states. Her position has included reviewing the files of offenders and writing reports with sentencing recommendations for judges. Our purpose in this consultation was to ensure that we were describing a scenario that was realistic and impactful. For example, the list of criteria we note as reasons for not giving the defendant time in jail is the same as the criteria that would be used by a probation officer in making a sentencing recommendation to a judge. She also advised us on how to calibrate the description such that both the recommendation to release or jail is prima facie plausible.

We based the description of the algorithm on the wording in [(Kennedy, Waggoner, and Ward 2022)](https://paperpile.com/c/yikrbz/zsXoH), which was sufficient for engendering a relatively high level of behavioral trust.

Respondents read a brief scenario involving a judge making a pre-trial decision as to whether to grant probation to a defendant in a repeat drunk driving case. We randomly assigned respondents to see one of seven scenarios (Table A1): (1) a judge makes the decision (control), (2) the judge makes the decision on the recommendation of an algorithm; (3) the judge disregards the recommendation of the algorithm; (4) the judge makes the decision on the recommendation of a trained probation officer; (5) the judge disregards the recommendation of a probation officer; (6) the judge makes the decision based on advice from *both* an algorithm and probation officer; or (7) the judge disregard the advice of *both* the algorithm and probation officer. In all scenarios, the judge decides *not* to detain the defendant. After the hearing, the defendant commits another drunk driving offense, this time resulting in fatalities.

|   | 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. The judge disagrees and grants probation to the defendant. |

**Table A1. Summary of experimental treatments**. This table gives a quick reference for the control condition, #1, and the 6 additional treatment conditions. For analysis, version 1, where the judge makes the decision on their own is the baseline. Full description of the vignettes can be found below.

After viewing the vignette, respondents indicate how much blame they place on the judge (10-point scale, 1 = None at all, 10 = A great deal). We also include similar measures for the degree of blame placed on the algorithm and/or probation officer in the treatments where this is relevant. We then rescale the dependent variable to run from 0-1 for ease of interpretation.

The exact wording of the vignettes are as follows, along with the blame question that differs between vignettes:

*Vignette 1 (Control Condition):*

A defendant with a prior drunk driving charge was arrested for drunk driving a second time.

Based on the defendant’s otherwise clean criminal history, lack of any other substance abuse charges, high level of education (master’s degree), generally stable family and friend support system, and lack of criminal attitudes, a judge granted the defendant a period of probation.

While on probation for the second offense, the defendant was again driving drunk, but this time struck someone with the vehicle, killing that person instantly.

*The following few pages will ask you some questions based on this information.*

On the following 10 point scale, where 1 indicates "none at all" and 10 indicates "a great deal", indicate the extent to which you think each of the following actors are to blame.

|   | (1) None at all | (2) | (3) | (4) | (5)  | (6) | (7) | (8) | (9) | (10) A great deal |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| The judge |  |  |  |  |  |  |  |  |  |  |

*Vignette 2 (Agrees with Algorithm):*

A defendant with a prior drunk driving charge was arrested for drunk driving a second time.

Based on the defendant’s otherwise clean criminal history, lack of any other substance abuse charges, high level of education (master’s degree), generally stable family and friend support system, and lack of criminal attitudes, a computer algorithm designed by computer scientists and criminal justice experts recommended probation for the defendant for this second offense.

A judge accepted the recommendation of the computer algorithm designed by computer scientists and criminal justice experts,[[3]](#footnote-2) and granted the defendant the recommended period of probation.

While on probation for the second offense, the defendant was again driving drunk, but this time struck someone with the vehicle, killing that person instantly.

*The following few pages will ask you some questions based on this information.*

On the following 10 point scale, where 1 indicates "none at all" and 10 indicates "a great deal", indicate the extent to which you think each of the following actors are to blame.

|   | (1) None at all | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) A great deal |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| The judge |  |  |  |  |  |  |  |  |  |  |
| The computer algorithm |  |  |  |  |  |  |  |  |  |  |

*Vignette 3 (Disagrees with Algorithm):*

A defendant with a prior drunk driving charge was arrested for drunk driving a second time.

Despite the defendant’s otherwise clean criminal history, lack of any other substance abuse charges, high level of education (master’s degree), generally stable family and friend support system, and lack of criminal attitudes, a computer algorithm designed by computer scientists and criminal justice experts recommended imprisonment for the defendant for this second offense.

A judge disagreed with the recommendation of the computer algorithm designed by computer scientists and criminal justice experts, and granted the defendant a period of probation.

While on probation for the second offense, the defendant was again driving drunk, but this time struck someone with the vehicle, killing that person instantly.

*The following few pages will ask you some questions based on this information.*

On the following 10 point scale, where 1 indicates "none at all" and 10 indicates "a great deal", indicate the extent to which you think each of the following actors are to blame.

|   | (1) None at all | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) A great deal |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| The judge |  |  |  |  |  |  |  |  |  |  |
| The computer algorithm |  |  |  |  |  |  |  |  |  |  |

*Vignette 4 (Agrees with Human):*

A defendant with a prior drunk driving charge was arrested for drunk driving a second time.

Based on the defendant’s otherwise clean criminal history, lack of any other substance abuse charges, high level of education (master’s degree), generally stable family and friend support system, and lack of criminal attitudes, a probation officer recommended probation for the defendant for this second offense.

A judge accepted the recommendation of the probation officer and granted the defendant the recommended period of probation.

While on probation for the second offense, the defendant was again driving drunk, but this time struck someone with the vehicle, killing that person instantly.

*The following few pages will ask you some questions based on this information.*

On the following 10 point scale, where 1 indicates "none at all" and 10 indicates "a great deal", indicate the extent to which you think each of the following actors are to blame.

|   | (1) None at all | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) A great deal |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| The judge |  |  |  |  |  |  |  |  |  |  |
| The probation officer |  |  |  |  |  |  |  |  |  |  |

*Vignette 5 (Disagrees with Human):*

A defendant with a prior drunk driving charge was arrested for drunk driving a second time.

Despite the defendant’s otherwise clean criminal history, lack of any other substance abuse charges, high level of education (master’s degree), generally stable family and friend support system, and lack of criminal attitudes, a probation officer recommended imprisonment for the defendant for this second offense.

A judge disagreed with the recommendation of the probation officer and granted the defendant a period of probation.

While on probation for the second offense, the defendant was again driving drunk, but this time struck someone with the vehicle, killing that person instantly.

*The following few pages will ask you some questions based on this information.*

On the following 10 point scale, where 1 indicates "none at all" and 10 indicates "a great deal", indicate the extent to which you think each of the following actors are to blame.

|   | (1) None at all | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) A great deal |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| The judge |  |  |  |  |  |  |  |  |  |  |
| The probation officer |  |  |  |  |  |  |  |  |  |  |

*Vignette 6 (Agrees with Human and Algorithm):*

A defendant with a prior drunk driving charge was arrested for drunk driving a second time.

Based on the defendant’s otherwise clean criminal history, lack of any other substance abuse charges, high level of education (master’s degree), generally stable family and friend support system, and lack of criminal attitudes, a probation officer along with a computer algorithm designed by computer scientists and criminal justice experts, recommended probation for the defendant for this second offense.

A judge accepted the recommendation of the probation officer and the computer algorithm designed by computer scientists and criminal justice experts, and granted the defendant the recommended period of probation.

While on probation for the second offense, the defendant was again driving drunk, but this time struck someone with the vehicle, killing that person instantly.

*The following few pages will ask you some questions based on this information.*

On the following 10 point scale, where 1 indicates "none at all" and 10 indicates "a great deal", indicate the extent to which you think each of the following actors are to blame.

|   | (1) None at all | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) A great deal |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| The judge |  |  |  |  |  |  |  |  |  |  |
| The probation officer |  |  |  |  |  |  |  |  |  |  |
| The computer algorithm |  |  |  |  |  |  |  |  |  |  |

*Vignette 7 (Disagrees with Algorithm and Human):*

A defendant with a prior drunk driving charge was arrested for drunk driving a second time.

Despite the defendant’s otherwise clean criminal history, lack of any other substance abuse charges, high level of education (master’s degree), generally stable family and friend support system, and lack of criminal attitudes, a probation officer along with a computer algorithm designed by computer scientists and criminal justice experts, recommended imprisonment for the defendant for this second offense.

A judge disagreed with the recommendation of the probation officer and the computer algorithm designed by computer scientists and criminal justice experts, and granted the defendant a period of probation.

While on probation for the second offense, the defendant was again driving drunk, but this time struck someone with the vehicle, killing that person instantly.

*The following few pages will ask you some questions based on this information.*

On the following 10 point scale, where 1 indicates "none at all" and 10 indicates "a great deal", indicate the extent to which you think each of the following actors are to blame.

|   | (1) None at all | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) A great deal |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| The judge |   |   |  |  |  |  |  |  |  |  |
| The probation officer |   |   |  |  |  |  |  |  |  |  |
| The computer algorithm |   |   |  |  |  |  |  |  |  |  |

*Electoral Question:*

In addition to the blame question, which differed for each vignette, respondents were also asked how the outcome would affect their likelihood of voting for the judge. Voting for judges is relatively common in the U.S., especially for lower courts. 39 states have elections for at least some judicial positions, a number of them even have elections for higher court positions (including the state supreme courts). Thus, we also asked whether the outcome would affect their likelihood of voting for the judge. The analysis of this question is below, and, as noted in the main text, none of the treatment conditions demonstrated significant impacts.

Here is the wording of the question:

This judge is up for reelection next year.

On a scale from 1 to 10, where 1 indicates "extremely unlikely" and 10 indicates "extremely likely", how likely is it that this incident would negatively influence your chances of voting for the judge?

o (1) Extremely unlikely

o (2)

o (3)

o (4)

o (5) Neither likely nor unlikely

o (6)

o (7)

o (8)

o (9)

o (10) Extremely likely

**SI.2. Attention Checks**

While Lucid has shown itself to produce results comparable with more expensive population surveys [(Coppock and McClellan 2019)](https://paperpile.com/c/yikrbz/Wv7aE), as with most online survey softwares, it is important to check whether the participant was paying attention [(Berinsky, Huber, and Lenz 2012)](https://paperpile.com/c/yikrbz/7MDEw). In our first Lucid sample, we checked whether the respondent was paying attention by asking them to recall which, if any, sources the judge received advice from in the scenario. This question also serves as a manipulation check, confirming that the respondent correctly noticed the source of advice (or lack thereof). For analysis, only those who passed the attention check in the Lucid sample were used in the analysis. This question was omitted from the TESS survey because of space constraints.

Here is the wording of the question:

If you remember back, who did the judge receive advice from in the scenario you read?

o A probation officer

o A computer algorithm

o Both of these sources

o Neither of these sources

This does raise some potential issues. First, it means we are only able to use about 62% of the first Lucid sample (923/1500). Second, some scholars have warned that using post-treatment variables for eliminating observations can cause misestimation of the causal effect when it is correlated with the treatment [(Montgomery, Nyhan, and Torres 2018)](https://paperpile.com/c/yikrbz/QAm3j). We do note such a correlation in the Lucid data, namely with regards to the control condition. Table A2 shows that the proportion of respondents who pass the check is substantially lower in this condition, apparently assuming that a judge will have received some kind of advice even if not specified. We are less concerned about this for our analysis because (1) the results replicate in another sample without any checks and (2) there are substantive reasons for exclusion of those who presume a source of advice that is not there (i.e., the causal effect cannot be estimated if a respondent assumes this).

| **Condition** | **Proportion correctly answering attention check** |
| --- | --- |
| Judge only | 0.46 |
| Agree with algorithm | 0.62 |
| Disagree with algorithm | 0.62 |
| Agree with human | 0.63 |
| Disagree with human | 0.59 |
| Agree with algorithm and human | 0.54 |
| Disagree with algorithm and human | 0.54 |

**Table A2. Proportion of respondents correctly answering attention/manipulation check in the first lucid sample by treatment condition.**

In the subsequent Lucid survey for the analysis of mediation and moderation (see section A13), we utilized an attention check prior to the manipulation using a standard format demonstrated by [(Berinsky, Margolis, and Sances 2014)](https://paperpile.com/c/yikrbz/0uxa0). Those who failed the attention check were excluded from going further in the survey. The attention check was the second question in the survey to ensure that inattentive respondents were removed quickly, without too much disruption. The attention check read as follows:



In total, 1,423 of the 3,656 respondents (38.9%) passed the attention check and were allowed to take the rest of the survey.

**SI.3. Sample Characteristics**

As noted in the main text, two samples were utilized for this study. The first, fielded in June of 2021 utilized Lucid’s Theorem platform (https://luc.id/theorem/), which draws from a range of survey panels and matches respondents such that the resulting sample will reflect census characteristics of the U.S. population. As noted above, the results from this platform have been shown to replicate well-established experimental results in a similar manner to more expensive survey options [(Coppock and McClellan 2019)](https://paperpile.com/c/yikrbz/Wv7aE). However, as noted in the last section, one must also be aware that these low-cost samples can have high levels of inattentiveness [(Aronow et al. 2020)](https://paperpile.com/c/yikrbz/TUW0i). We requested 1,500 respondents in total. Of these, 923 (62%) passed the attention check and were utilized in the study.

The demographic characteristics of this survey sample are listed in Table A2.

|  |
| --- |
| Statistic | N | Mean | St. Dev. | Min | Pctl(25) | Pctl(75) | Max |
|  |
| Female | 923 | 0.519 | 0.500 | 0 | 0 | 1 | 1 |
| Democrat | 923 | 0.445 | 0.497 | 0 | 0 | 1 | 1 |
| Republican | 923 | 0.391 | 0.488 | 0 | 0 | 1 | 1 |
| Age | 923 | 49.183 | 17.336 | 18 | 35 | 64 | 96 |
| Income | 864 | 8.712 | 6.917 | 1.000 | 3.000 | 14.000 | 24.000 |
| White | 923 | 0.779 | 0.415 | 0 | 1 | 1 | 1 |
| Education | 914 | 4.393 | 1.999 | 1.000 | 2.000 | 6.000 | 10.000 |
|  |

**Table A2. Summary statistics for first Lucid sample.**

After the survey was designed, but prior to its posting on Lucid, one of the co-authors was awarded space in the Time-sharing Experiments in the Social Sciences (TESS) survey, which provided a national-representative sample of 1,846 Americans, surveyed in Autumn of 2021. TESS is an NSF-funded program to allow social scientists to field their experiments, alongside other scholars, reducing the overall costs. Projects proposed for TESS undergo a process of blind peer review prior to their being submitted, resulting in a number of changes to the survey in response. These were incorporated into both the Lucid and TESS surveys. TESS has produced a range of influential studies [(Mutz 2011)](https://paperpile.com/c/yikrbz/vDQeR)(see also <https://www.tessexperiments.org/paststudies>). TESS contracts with the National Opinion Research Center (NORC) at the University of Chicago through their AmeriSpeak panel. A summary of the AmeriSpeak Panel sample design used for TESS can be found here: <https://amerispeak.norc.org/about-amerispeak/Pages/Panel-Design.aspx>. NORC is one of the largest and most respected survey research operations in the U.S. and has been in operation since 1941. For this experiment, we used their standard nationally-representative panel without any over-sampling.

The demographic characteristics of this survey sample are listed in Table A3.

|  |
| --- |
| Statistic | N | Mean | St. Dev. | Min | Pctl(25) | Pctl(75) | Max |
|  |
| Female | 1,846 | 0.547 | 0.498 | 0 | 0 | 1 | 1 |
| Democrat | 1,844 | 0.462 | 0.499 | 0.000 | 0.000 | 1.000 | 1.000 |
| Republican | 1,844 | 0.357 | 0.479 | 0.000 | 0.000 | 1.000 | 1.000 |
| Age | 1,846 | 48.738 | 17.168 | 18 | 33 | 63 | 93 |
| Income | 1,846 | 9.904 | 4.155 | 1 | 7 | 13 | 18 |
| White | 1,846 | 0.654 | 0.476 | 0 | 0 | 1 | 1 |
| Education | 1,846 | 3.168 | 1.028 | 1 | 3 | 4 | 5 |
|  |

**Table A3. Summary statistics for TESS sample.**

**SI.4. Balance Tests**

While the role of balance tests is somewhat controversial, with some viewing it as an essential part of experimental reporting, and others viewing it as nonsensical when using randomization [(A. Gerber et al. 2014; Mutz and Pemantle 2015)](https://paperpile.com/c/yikrbz/SIJOz%2BVO7b9), we go ahead and present them below. As the reader will note, in the Lucid sample (Table A4) and the TESS sample (Table A5), there are some instances where there are statistically significant (p < 0.05) differences between groups in pairwise t-tests. This is not unexpected, since, when there are this many comparisons, the probabilities of statistically significant differences by pure chance becomes reasonably high. When standard corrections for multiple comparisons are applied, the number of significant differences drops to near zero. Nevertheless, section A9 below replicates the analysis in the main paper while including the demographic characteristics to eliminate any potential influence on the results. The results remain robust to this change in model specification.

| **Treatment** | **N** | **Female** | **Democrat** | **Republican** | **Age** | **Income** | **White** | **Education** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Judge only (control) | 103 | 0.437 | 0.427 | 0.427 | 53.194 | 8.245 | 0.854 | 4.738 |
| Agrees with algorithm | 140 | 0.564(0.050) | 0.514(0.177) | 0.400(0.669) | 47.279(0.009) | 10.773(0.006) | 0.8(0.312) | 4.532(0.428) |
| Disagrees with algorithm | 148 | 0.581(0.025)(0.775) | 0.446(0.768)(0.243) | 0.372(0.376)(0.623) | 48.48(0.034)(0.556) | 8.732(0.592)(0.016) | 0.736(0.027)(0.194) | 4.055(0.008)(0.044) |
| Agrees with human | 147 | 0.463(0.689)(0.085)(0.042) | 0.497(0.277)(0.763)(0.381) | 0.361(0.289)(0.495)(0.846) | 50.313(0.196)(0.138)(0.363) | 8.645(0.658)(0.012)(0.916) | 0.721(0.012)(0.107)(0.750) | 4.644(0.714)(0.637)(0.012) |
| Disagrees with human | 133 | 0.549(0.088)(0.799)(0.589)(0.149) | 0.414(0.834)(0.094)(0.585)(0.163) | 0.414(0.832)(0.819)(0.474)(0.366) | 48.556(0.042)(0.542)(0.970)(0.397) | 8.431(0.842)(0.007)(0.724)(0.801) | 0.789(0.233)(0.834)(0.285)(0.168) | 4.248(0.062)(0.240)(0.420)(0.098) |
| Agrees with algorithm and human | 127 | 0.488(0.438)(0.214)(0.124)(0.679)(0.327) | 0.362(0.324)(0.013)(0.164)(0.026)(0.405) | 0.409(0.785)(0.875)(0.523)(0.410)(0.946) | 48.992(0.068)(0.419)(0.807)(0.529)(0.839) | 7.857(0.680)(0.001)(0.310)(0.358)(0.517) | 0.827(0.615)(0.598)(0.072)(0.035)(0.468) | 4.216(0.049)(0.198)(0.509)(0.079)(0.897) |
| Disagrees with algorithm and human | 125 | 0.528(0.170)(0.555)(0.382)(0.282)(0.737)(0.527) | 0.44(0.846)(0.224)(0.922)(0.349)(0.669)(0.214) | 0.368(0.364)(0.595)(0.951)(0.900)(0.455)(0.502) | 48.376(0.037)(0.607)(0.961)(0.358)(0.933)(0.778) | 8.068(0.851)(0.002)(0.443)(0.503)(0.683)(0.814) | 0.752(0.064)(0.347)(0.758)(0.540)(0.468)(0.152) | 4.382(0.182)(0.543)(0.181)(0.284)(0.591)(0.512) |

**Table A4. Balance table for Lucid sample.** For each demographic, the mean value is presented in the corresponding cell, with the p-values from an uncorrected paired t-test for all of the conditions above the condition in sequential parentheses. For example, for the last condition (Disagrees with algorithm and human) the first parenthetical value is the p-value the paired t-test with the control condition, the second parenthetical values is the p-value from the paired t-test with the second condition (Agrees with algorithm), etc. None of the p-values are corrected for multiple comparisons. The measure of party ID includes those who report “leaning” towards one party or the other. Income is measured on a Likert scale (1 = < $14,999, 24 = > $250,000).

| **Treatment** | **N** | **Female** | **Democrat** | **Republican** | **Age** | **Income** | **White** | **Education** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Judge only (control) | 252 | 0.548 | 0.532 | 0.317 | 47.778 | 9.964 | 0.631 | 3.19 |
| Agrees with algorithm | 258 | 0.562(0.744) | 0.442(0.042) | 0.391(0.081) | 49.624(0.225) | 9.841(0.738) | 0.643(0.768) | 3.163(0.761) |
| Disagrees with algorithm | 265 | 0.532(0.723)(0.492) | 0.434(0.026)(0.856) | 0.343(0.539)(0.251) | 49.336(0.303)(0.848) | 9.683(0.442)(0.664) | 0.657(0.541)(0.751) | 3.075(0.204)(0.332) |
| Agrees with human | 251 | 0.578(0.498)(0.723)(0.298) | 0.45(0.067)(0.850)(0.711) | 0.39(0.088)(0.981)(0.265) | 48.574(0.604)(0.491)(0.615) | 10.028(0.864)(0.612)(0.346) | 0.697(0.119)(0.203)(0.333) | 3.195(0.959)(0.722)(0.186) |
| Disagrees with human | 276 | 0.482(0.130)(0.063)(0.241)(0.028) | 0.475(0.188)(0.447)(0.342)(0.574) | 0.373(0.182)(0.660)(0.470)(0.680) | 49.025(0.405)(0.688)(0.834)(0.763) | 10.127(0.654)(0.428)(0.215)(0.785) | 0.645(0.736)(0.971)(0.776)(0.208) | 3.203(0.890)(0.652)(0.150)(0.932) |
| Agrees with algorithm and human | 286 | 0.566(0.662)(0.918)(0.418)(0.794)(0.044) | 0.411(0.005)(0.464)(0.581)(0.358)(0.128) | 0.358(0.329)(0.415)(0.723)(0.433)(0.706) | 48.175(0.789)(0.326)(0.428)(0.788)(0.558) | 9.591(0.299)(0.483)(0.795)(0.224)(0.127) | 0.661(0.468)(0.670)(0.917)(0.377)(0.692) | 3.122(0.443)(0.647)(0.593)(0.413)(0.353) |
| Disagrees with algorithm and human | 258 | 0.566(0.679)(0.930)(0.437)(0.789)(0.052)(0.990) | 0.498(0.446)(0.201)(0.142)(0.279)(0.588)(0.041) | 0.323(0.897)(0.105)(0.626)(0.113)(0.227)(0.397) | 48.655(0.564)(0.522)(0.651)(0.957)(0.804)(0.745) | 10.12(0.672)(0.446)(0.229)(0.802)(0.985)(0.138) | 0.643(0.768)(1.000)(0.751)(0.203)(0.971)(0.670) | 3.236(0.614)(0.416)(0.074)(0.651)(0.706)(0.197) |

**Table A5. Balance table for TESS sample.** For each demographic, the mean value is presented in the corresponding cell, with the p-values from an uncorrected paired t-test for all of the conditions above the condition in sequential parentheses. For example, for the last condition (Disagrees with algorithm and human) the first parenthetical value is the p-value the paired t-test with the control condition, the second parenthetical values is the p-value from the paired t-test with the second condition (Agrees with algorithm), etc. None of the p-values are corrected for multiple comparisons. The measure of party ID includes those who report “leaning” towards one party or the other. Income is measured on a Likert scale (1 = < $5,000, 18 = > $200,000).

**SI.5. Preregistration**

We preregistered our hypotheses via the Open Science Framework (osf.io) [(Open Science Collaboration 2015; Soderberg et al. 2021)](https://paperpile.com/c/yikrbz/jVVn2%2B5rhUH) in June of 2021, prior to fielding of both the Lucid and TESS surveys.[[4]](#footnote-3) The preregistration set up three hypotheses:

*H1:* 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.

*H2:* The reduction in blame from relying on an algorithm will be similar to that of reliance on advice from a trained bureaucrat.

*H3:* When an error occurs, a policy maker’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.

Our analysis approach, using OLS regression with blame and election as the dependent variable and treatment status as the independent variable was also included in this pre-registration.

Obviously, from the results in the main text, none of these three hypotheses held up once they were tested. Indeed, the results ended up being almost directly opposite our expectations. In addition to serving the role of reassuring the reviewer that we were not conducting a fishing expedition in designing this survey experiment, the pre-registration also emphasizes just how counter-intuitive these findings are in light of previous exposition on the subject.

**SI.6. Power Analysis**

Using G\*Power, and assuming a medium-small treatment effect for between group comparisons (two-tailed, t-test), we estimate a sample of about 250 in each group will provide us with approximately 90% power [(Faul et al. 2009)](https://paperpile.com/c/yikrbz/SUDX4). This means that the TESS sample is well-powered on its own, while the Lucid sample is underpowered on its own. Pooled, the analysis in the main paper has more than enough power for detection of a medium-small effect.

For the moderation and mediation analysis, we relied on the simulation results from [(Fritz and Mackinnon 2007)](https://paperpile.com/c/yikrbz/Zqftj), which suggests a sample size of 667 is sufficient for 0.80 power for all but the very weakest of conditions, which translates to a sample of 1001 distributed across three treatment arms. Our total n for those experiments was well above this threshold (n = 1423).

**SI.7. Separate Sample Result**

In this section, we present the results of the two survey samples separately. Table A6 presents an OLS regression analysis which assesses the average treatment effect for each treatment for both our nationally-representative sample and census-matched sample. Results refute our preregistered hypotheses, yielding consistent pattern in which receiving advice from the computer algorithm increases the level of blame placed upon the judge in both samples.

The results remain relatively consistent between samples. In both samples, the direction of effects remains consistent. For most of the results, with the exception of when the judge receives advice from both the algorithm and human, the magnitude of the results is also surprisingly consistent. In the one exception to this, respondents in the TESS sample appeared to give the judge quite a bit less blame in the combined advice treatment than did those in the Lucid sample. The statistical significance of the findings is also generally consistent both between samples and with the combined results in the main paper.

These results should assure the reader that the findings in the main paper are not driven by a particular sub-sample, or that they might mask disagreement between the samples. Both samples are quite consistent in suggesting that blame placed on the judge increases when they make decisions based on advice rather than using their own judgment.

|  |
| --- |
|  | *Dependent variable:* |
|  |  |
|  | Blame on judge(Lucid) | Blame on judge(TESS) |
|  | (1) | (2) |
|  |
| Agrees with algorithm | 0.086\*\* | 0.096\*\*\* |
|  | (0.042) | (0.029) |
|  |  |  |
| Disagrees with algorithm | 0.060 | 0.067\*\* |
|  | (0.041) | (0.029) |
|  |  |  |
| Agrees with human | 0.064 | 0.077\*\*\* |
|  | (0.041) | (0.029) |
|  |  |  |
| Disagrees with human | 0.088\*\* | 0.079\*\*\* |
|  | (0.042) | (0.028) |
|  |  |  |
| Agrees with algorithm and human | 0.099\*\* | 0.034 |
|  | (0.043) | (0.028) |
|  |  |  |
| Disagrees with algorithm and human | 0.093\*\* | 0.047 |
|  | (0.043) | (0.029) |
|  |  |  |
| Constant | 0.520\*\*\* | 0.558\*\*\* |
|  | (0.032) | (0.021) |
|  |  |  |
|  |
| Observations | 922 | 1,842 |
| R2 | 0.008 | 0.008 |
| Adjusted R2 | 0.002 | 0.005 |
| Residual Std. Error | 0.322 (df = 915) | 0.326 (df = 1835) |
| F Statistic | 1.245 (df = 6; 915) | 2.558\*\* (df = 6; 1835) |
|  |
| *Note:* | \*p<0.1;\*\*p<0.05;\*\*\*p<0.01 |

**Table A6. Tabular results for blame on judge, separated by sample.** Values are OLS regression coefficients with standard errors in parentheses.Column 1 shows the results for blame placed on the judge in the Lucid sample and column 2 shows this in the TESS sample.

**SI.8. Tabular Results and Standardized Effect Size**

In this section, we present tabular versions of the results from Figure 1 in the main paper so the interested reader can see the exact values represented in the charts. These values in Table A7 are discussed in the main paper, so we forego further discussion here.

|  |
| --- |
|  | *Dependent variable:* |
|  |  |
|  | Blame on judge |
|  |
| Agrees with algorithm | 0.090\*\*\* |
|  | (0.024) |
|  |  |
| Disagrees with algorithm | 0.062\*\*\* |
|  | (0.024) |
|  |  |
| Agrees with human | 0.069\*\*\* |
|  | (0.024) |
|  |  |
| Disagrees with human | 0.081\*\*\* |
|  | (0.024) |
|  |  |
| Agrees with algorithm and human | 0.054\*\* |
|  | (0.024) |
|  |  |
| Disagrees with algorithm and human | 0.061\*\* |
|  | (0.024) |
|  |  |
| Constant | 0.547\*\*\* |
|  | (0.017) |
|  |  |
|  |
| Observations | 2,764 |
| R2 | 0.006 |
| Adjusted R2 | 0.004 |
| Residual Std. Error | 0.325 (df = 2757) |
| F Statistic | 2.924\*\*\* (df = 6; 2757) |
|  |
| *Note:* | \*p<0.1; \*\*p<0.05;  \*\*\*p<0.01 |

**Table A7. Tabular presentation of results in Figure 1 of the main paper.** Values are OLS regression coefficients with standard errors in parentheses.

In addition to the raw results, it is sometimes helpful to characterize the overall size of the effects observed. Larger effect sizes tend to be more robust across replications and can provide a more intuitive explanation of effects for some readers than differences in means or regression coefficients. Figure A1 shows the Cohen’s d estimates for all of the treatments (compared with the control condition). The standardized effects are generally in the area considered “small” for agreeing with the algorithm, agreeing with the human and disagreeing with the human, and negligible for disagreeing with the algorithm, agreeing with both the algorithm and human and disagreeing with both the algorithm and human. While the effect sizes are not large, they are consistent and they still provide significant refutation of previous hypotheses, which suggest the judge should receive less blame for agreeing with advice and more blame for disagreeing with advice when there is an adverse outcome of the decision.



**Figure A.1: Standard effect size (Cohen’s d) for all treatment conditions and 95% confidence intervals.**

**SI.9. Results with Control Variables**

While we technically do not need to add any control variables to recover an estimate of the ATE under conditions of randomization (see section *SI.12* below), but the inclusion of covariates in estimating ATE has a couple of potential advantage – increasing the precision by which the causal effect is estimated and also dealing with any imbalances in respondent characteristics between treatment arms that occur randomly in the random assignment process.

As the reader can see in Table A8 and Figure A2, the results remain essentially the same as those presented in the main paper. The substantive results remain accurate to within thousandths of a point and the levels of statistical significance remain the same.

|  |
| --- |
|  | *Dependent variable:* |
|  |  |
|  | Blame on judge |
|  |
| Agrees with algorithm | 0.089\*\*\* |
|  | (0.024) |
|  |  |
| Disagrees with algorithm | 0.065\*\*\* |
|  | (0.024) |
|  |  |
| Agrees with human | 0.071\*\*\* |
|  | (0.024) |
|  |  |
| Disagrees with human | 0.082\*\*\* |
|  | (0.024) |
|  |  |
| Agrees with algorithm and human | 0.055\*\* |
|  | (0.024) |
|  |  |
| Disagrees with algorithm and human | 0.061\*\* |
|  | (0.024) |
|  |  |
| Female | 0.027\*\* |
|  | (0.013) |
|  |  |
| Age | 0.002\*\*\* |
|  | (0.0004) |
|  |  |
| Party ID | -0.001 |
|  | (0.003) |
|  |  |
| Income | 0.002 |
|  | (0.002) |
|  |  |
| White | -0.031\*\* |
|  | (0.014) |
|  |  |
| Education | -0.017\*\*\* |
|  | (0.006) |
|  |  |
| Constant | 0.480\*\*\* |
|  | (0.033) |
|  |  |
|  |
| Observations | 2,697 |
| R2 | 0.026 |
| Adjusted R2 | 0.022 |
| Residual Std. Error | 0.323 (df = 2684) |
| F Statistic | 6.001\*\*\* (df = 12; 2684) |
|  |
| *Note:* | \*p<0.1;\*\*p<0.05;\*\*\*p<0.01 |

**Table A8. Results of regression models in combined sample including demographic control variables.** Values are OLS regression coefficients with standard errors in parentheses.



**Figure A2: Results of regression models in combined sample including demographic control variables.** Dots are OLS regression coefficients (the estimated Average Treatment Effect (ATE)) with the spike caps showing the 95% confidence intervals estimated from 1,000 simulated draws from the posterior distribution.

**SI.10. Initial Tests for Moderating Effects**

When this research was proposed for TESS, one reviewer suggested that there might be some interesting heterogenous treatment effects based on demographics or the ease respondents feel with algorithms and automation. This is a similar idea to what has been found elsewhere in the algorithm trust literature. In particular younger, male, and more educated respondents tend to show a greater acceptance of advice from machines [(Hoff and Bashir 2015)](https://paperpile.com/c/yikrbz/GXuuG). The reviewer also suggested that there may be some people who are more predisposed to trust technology more generally, again an idea that finds some support in the literature on generalized trust in automation [(Kim, Ferrin, and Rao 2008)](https://paperpile.com/c/yikrbz/cYLW2). To test this, we look at the interaction between whether an algorithm was involved in the decision and the particular individual characteristic [(A. S. Gerber and Green 2012)](https://paperpile.com/c/yikrbz/s89ia).

This involved a few changes. First, we combined all the treatments involving algorithms into a single treatment variable that is compared against the control condition. We do this because heterogeneous treatment effects will, by design, decrease the amount of power within each group being considered. We therefore wish to maximize the power of these tests, which we can do by combining the treatments. Second, while it was a TESS reviewer who suggested its inclusion, there was not room in the TESS survey for a separate trust in automation battery, so this model (6) only includes the Lucid sample.

We should also note that none of these interactive hypotheses were pre-registered, nor did we conduct block randomization by subgroup in the test for these heterogeneous treatment effects. As such, these results should be taken as very preliminary. We report them for two reasons: (1) to test whether our results are limited to a particular subgroup; and (2) as a potential jumping off point for future research among interested readers.

The results suggest that the results in the main paper are not limited to a particular subgroup of the sample population. None of the interactions reach traditional levels of statistical certainty, nor do they have impacts on a scale that would call into question the main effects presented in the main paper.

|  |
| --- |
|  | *Dependent variable:* |
|  |  |
|  | Blame on judge |
|  | (1) | (2) | (3) | (4) | (5) | (6) |
|  |
| Algorithm | 0.067\*\*\* | 0.087\*\*\* | 0.051 | 0.062\* | 0.111\* | 0.129 |
|  | (0.019) | (0.028) | (0.058) | (0.035) | (0.062) | (0.180) |
|  |  |  |  |  |  |  |
| Algorithm \* Female |  | -0.038 |  |  |  |  |
|  |  | (0.039) |  |  |  |  |
|  |  |  |  |  |  |  |
| Algorithm \* Age |  |  | 0.0003 |  |  |  |
|  |  |  | (0.001) |  |  |  |
|  |  |  |  |  |  |  |
| Algorithm \* White |  |  |  | 0.007 |  |  |
|  |  |  |  | (0.042) |  |  |
|  |  |  |  |  |  |  |
| Algorithm \* Education |  |  |  |  | -0.014 |  |
|  |  |  |  |  | (0.018) |  |
|  |  |  |  |  |  |  |
| Algorithm \* Trust in automation |  |  |  |  |  | -0.003 |
|  |  |  |  |  |  | (0.045) |
|  |  |  |  |  |  |  |
| Female | 0.019 | 0.050 | 0.019 | 0.019 | 0.019 | 0.003 |
|  | (0.015) | (0.035) | (0.015) | (0.015) | (0.015) | (0.028) |
|  |  |  |  |  |  |  |
| Age | 0.002\*\*\* | 0.002\*\*\* | 0.002\* | 0.002\*\*\* | 0.002\*\*\* | 0.001 |
|  | (0.0005) | (0.0005) | (0.001) | (0.0005) | (0.0005) | (0.001) |
|  |  |  |  |  |  |  |
| Party ID | -0.0002 | -0.0002 | -0.0001 | -0.0002 | -0.0001 | 0.013\*\* |
|  | (0.004) | (0.004) | (0.004) | (0.004) | (0.004) | (0.006) |
|  |  |  |  |  |  |  |
| Income | 0.002 | 0.002 | 0.002 | 0.002 | 0.002 | 0.002 |
|  | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.004) |
|  |  |  |  |  |  |  |
| White | -0.032\* | -0.032\* | -0.032\* | -0.038 | -0.032\* | -0.029 |
|  | (0.017) | (0.017) | (0.017) | (0.038) | (0.017) | (0.038) |
|  |  |  |  |  |  |  |
| Education | -0.019\*\* | -0.019\*\* | -0.019\*\* | -0.019\*\* | -0.008 | 0.008 |
|  | (0.008) | (0.008) | (0.008) | (0.008) | (0.017) | (0.014) |
|  |  |  |  |  |  |  |
| Trust in automation |  |  |  |  |  | -0.002 |
|  |  |  |  |  |  | (0.041) |
|  |  |  |  |  |  |  |
| Constant | 0.502\*\*\* | 0.485\*\*\* | 0.515\*\*\* | 0.506\*\*\* | 0.466\*\*\* | 0.384\*\* |
|  | (0.038) | (0.042) | (0.058) | (0.045) | (0.061) | (0.178) |
|  |  |  |  |  |  |  |
|  |
| Observations | 1,907 | 1,907 | 1,907 | 1,907 | 1,907 | 550 |
| R2 | 0.021 | 0.022 | 0.021 | 0.021 | 0.022 | 0.029 |
| Adjusted R2 | 0.018 | 0.018 | 0.017 | 0.017 | 0.017 | 0.012 |
| Residual Std. Error | 0.324 (df = 1899) | 0.324 (df = 1898) | 0.324 (df = 1898) | 0.324 (df = 1898) | 0.324 (df = 1898) | 0.319 (df = 540) |
| F Statistic | 5.905\*\*\* (df = 7; 1899) | 5.290\*\*\* (df = 8; 1898) | 5.175\*\*\* (df = 8; 1898) | 5.168\*\*\* (df = 8; 1898) | 5.236\*\*\* (df = 8; 1898) | 1.762\* (df = 9; 540) |
|  |
| *Note:* | \*p<0.1; >\*\*p<0.05; >\*\*\*p<0.01 |

**Table A9. Tests of heterogeneous treatment effects based on demographics and trust in automation.** Values are OLS regression coefficients with standard errors in parentheses.

**SI.11. Results for Election Question**

As noted in the research design section above, we also asked respondents the extent to which the judge’s decision would impact their decision to vote for the judge in an upcoming election. Table A10 shows the results for this question, split by sample. The results suggest that this single incident and the corresponding treatment conditions did not have much impact on likelihood of voting for the judge. While respondents generally leaned towards saying it would affect their vote somewhat (about a 7 out of 10 on average), this changed very little based on treatment conditions and not in a manner consistent across samples.

Similar to how our results for blame suggest that policy-makers will not be shielded from blame by using algorithms in their decision-making process, such advice appears to have no positive effect on the likelihood of election.

|  |
| --- |
|  | *Dependent variable:* |
|  |  |
|  | Election |
|  | (TESS) | (Lucid) |
|  |
| Agrees with algorithm | 0.038 | -0.038 |
|  | (0.029) | (0.040) |
|  |  |  |
| Disagrees with algorithm | 0.010 | -0.061 |
|  | (0.029) | (0.039) |
|  |  |  |
| Agrees with human | -0.002 | -0.021 |
|  | (0.030) | (0.039) |
|  |  |  |
| Disagrees with human | 0.043 | 0.023 |
|  | (0.029) | (0.040) |
|  |  |  |
| Agrees with algorithm and human | -0.032 | -0.008 |
|  | (0.029) | (0.041) |
|  |  |  |
| Disagrees with algorithm and human | -0.005 | -0.021 |
|  | (0.029) | (0.041) |
|  |  |  |
| Constant | 0.604\*\*\* | 0.614\*\*\* |
|  | (0.021) | (0.030) |
|  |  |  |
|  |
| Observations | 1,841 | 923 |
| R2 | 0.005 | 0.007 |
| Adjusted R2 | 0.002 | 0.0003 |
| Residual Std. Error | 0.331 (df = 1834) | 0.306 (df = 916) |
| F Statistic | 1.663 (df = 6; 1834) | 1.050 (df = 6; 916) |
|  |
| *Note:* | \*p<0.1;\*\*p<0.05;\*\*\*p<0.01 |

**Table A10. Results of regression models in separate samples for the election outcome question.** Values are OLS regression coefficients with standard errors in parentheses.

**SI.12. Results for Blame on Algorithm and/or Human Advisor**

As noted in our pre-registration, we also analyzed the level of blame placed on the algorithm and human adviser. The main conclusion from this is that blame placed on these advisors is much higher (roughly equal to that placed on the judge) when the judge accepts their judgment than when the judge does not. Table A11 shows these results. The clear implication is that when the algorithm or human (or both) disagree with the decision of the judge, they receive a much lower amount of the blame – an intuitive result, but one that also verifies that respondents are responding to the scenarios as presented. The blame placed on the algorithm or human decreases by about 30% when the judge disregards their advice, and by about 3-5% when they are both making a recommendation, instead of each separate.

|  |
| --- |
|  | *Dependent variable:* |
|  |  |
|  | Blame on algorithm | Blame on human |
|  | (1) | (2) |
|  |
| Disagrees with algorithm | -0.305\*\*\* |  |
|  | (0.024) |  |
|  |  |  |
| Disagrees with human |  | -0.310\*\*\* |
|  |  | (0.023) |
|  |  |  |
| Agrees with algorithm and human | -0.055\*\* | -0.038\* |
|  | (0.024) | (0.023) |
|  |  |  |
| Disagrees with algorithm and human | -0.307\*\*\* | -0.269\*\*\* |
|  | (0.024) | (0.023) |
|  |  |  |
| Constant | 0.627\*\*\* | 0.604\*\*\* |
|  | (0.017) | (0.016) |
|  |  |  |
|  |
| Observations | 1,600 | 1,602 |
| R2 | 0.147 | 0.152 |
| Adjusted R2 | 0.146 | 0.151 |
| Residual Std. Error | 0.338 (df = 1596) | 0.324 (df = 1598) |
| F Statistic | 91.970\*\*\* (df = 3; 1596) | 95.790\*\*\* (df = 3; 1598) |
|  |
| *Note:* | \*p<0.1;\*\*p<0.05\*\*\*p<0.01 |

**Table A11. Results of regression models for blame on human and algorithm sources of advice.** Values are OLS regression coefficients with standard errors in parentheses. The results suggest that the human or algorithm adviser receives a similar amount of blame as the judge when the judge agrees with them and much lower blame when the judge disagrees with them.

**SI.13. Moderation and Mediation Analysis**

Questions about why particular treatment effects arise can be difficult to answer in experiments [(Bullock, Green, and Ha 2010; Green, Ha, and Bullock 2010)](https://paperpile.com/c/yikrbz/J1gd7%2BldES6), but it is important enough that it is difficult to abandon [(Imai et al. 2011)](https://paperpile.com/c/yikrbz/8Pl30).

To develop our hypotheses for what may explain the increase in blame associated with receiving advice, we reviewed recent literature, as well as looking through comments left by participants. From this, we developed, and pre-registered,[[5]](#footnote-4) three hypotheses based on the following intuition.

First, individuals may place more blame on judges when they have less trust in expert advice, be that from a human or an algorithm [(Bertsou 2021; Merkley 2021)](https://paperpile.com/c/yikrbz/4vHhH%2B9YvWT). Conversely, they may place more blame on the judge when the judge disregards the advice and they trust advice from experts [(Surden 2021)](https://paperpile.com/c/yikrbz/ZQnKK). This posits a moderation of the effect of the treatment based on the level of general trust in expert advice. Hypothesis 1 was pre-registered as:

*H1: There will be heterogenous treatment effects, with participants who are less trusting in expert advice blaming the judge more for using advice than those who are more trusting in expert advice.*

Second, given previous studies finding that people have relatively high behavioral trust in algorithms [(Kennedy, Waggoner, and Ward 2022; Logg, Minson, and Moore 2019)](https://paperpile.com/c/yikrbz/zsXoH%2BtMf6Z), it is possible that the use of an algorithms increases the expectations that the judge would make the correct decision. When those expectations are not met, they will place more blame on the judge when the decision is made incorrectly. This posits a mediation of the treatment effect by changes in expectations. Hypothesis 2 was pre-registered as:

*H2: The increase in blame on judges who use advice from an algorithm will be mediated by an increase in expectations of accuracy of the judgment. Those whose expectations increase due to the presence of outside advice will place more blame on the judge than those who do not.*

Finally, drawing from some of the legal literature on the subject, we considered the possibility that the treatment may induce some degree of feeling that the judge is abdicating their responsibility to make a judgment by taking into account the advice of an outside expert (algorithm) [(Chesterman 2021)](https://paperpile.com/c/yikrbz/cuLHi). This suggests another potential mediating effect, especially where the judge agrees with the algorithm. The treatment may result in respondents having stronger feelings that the judge *should have* relied on their own judgment. While there is a certain degree of retrospective bias involved in this perception, i.e. that this attitude changes in light of seeing the mistake, this is not unusual behavior in politics, even when the actor has little to no control over the outcome [(Achen and Bartels 2017)](https://paperpile.com/c/yikrbz/oXTH6). Hypothesis 3 was pre-registered as:

*H3: The increase in blame on judges who use advice from an algorithm will be mediated by a view that the judge is abdicating their duty to use their own judgment. Those who think the judge is relying more on the algorithm in their decision will place more blame on the judge than those who do not.*

We fielded a new version of the study on Lucid with a sample size of approximately 1,400 respondents. This is a little higher than the target we pre-registered (n = 1,200), and is based on simulation results from [(Fritz and Mackinnon 2007)](https://paperpile.com/c/yikrbz/Zqftj), which suggests a sample size of 667 is sufficient for 0.80 power for all but the very weakest of conditions, which translates to a sample of 1001 for three treatment arms. Table A.12 shows the sample characteristics for this new sample.

|  |
| --- |
| Statistic | N | Mean | St. Dev. | Min | Max |
|  |
| Female | 1,423 | 0.548 | 0.498 | 0 | 1 |
| Democrat | 1,423 | 0.471 | 0.499 | 0 | 1 |
| Republican | 1,423 | 0.354 | 0.478 | 0 | 1 |
| Income | 1,362 | 10.041 | 7.197 | 1 | 24 |
| White | 1,423 | 0.781 | 0.414 | 0 | 1 |
| Education | 1,421 | 4.698 | 1.864 | 1 | 8 |
|  |

 **Table A12: Summary statistics for moderation/mediation study**

Since the focus of the study is on algorithms, we only presented two of the treatment conditions above (agree with algorithm and disagree with algorithm) along with the control condition. As noted above, to address the concerns raised by [(Montgomery, Nyhan, and Torres 2018)](https://paperpile.com/c/yikrbz/QAm3j), we relied on an attention check to filter respondents instead of relying on the post-treatment manipulation checks. We noted that about 77% of the resulting participants did correctly identify the manipulation.

Beginning the analysis, 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). These results can be seen in Table A12. While this lack of statistical significance is surprising, we should note that the general consistency with previous results is more important for evaluating the overall results and does not necessarily prevent successful analysis of moderation or mediation [(Bollen 1989; Hayes 2017)](https://paperpile.com/c/yikrbz/OAK3t%2B4k6rW). The results still refute the concerns laid out by previous scholars, although with weaker evidence.

|  |
| --- |
|  | *Dependent variable:* |
|  |  |
|  | Blame on judge |
|  |
| Agrees with algorithm | 0.026 |
|  | (0.021) |
|  |  |
| Disagrees with algorithm | 0.020 |
|  | (0.021) |
|  |  |
| Trust in experts | 0.025 |
|  | (0.052) |
|  |  |
| Female | 0.011 |
|  | (0.017) |
|  |  |
| Democrat | 0.050\*\* |
|  | (0.025) |
|  |  |
| Republican | 0.069\*\*\* |
|  | (0.025) |
|  |  |
| Income | 0.004\*\*\* |
|  | (0.001) |
|  |  |
| White | -0.035 |
|  | (0.022) |
|  |  |
| Education | -0.001 |
|  | (0.005) |
|  |  |
| Constant | 0.548\*\*\* |
|  | (0.046) |
|  |  |
|  |
| Observations | 1,359 |
| R2 | 0.019 |
| Adjusted R2 | 0.012 |
| Residual Std. Error | 0.313 (df = 1349) |
| F Statistic | 2.874\*\*\* (df = 9; 1349) |
|  |
| *Note:* | \*p<0.1\*\*p<0.05\*\*\*p<0.01 |

**Table A13: Overall treatment effects in moderation/mediation study**

To measure trust in experts, we utilize the scale developed by [(Merkley 2021)](https://paperpile.com/c/yikrbz/9YvWT). The question read:



The responses were added together and scaled from 0 to 1.

Following [(A. S. Gerber and Green 2012)](https://paperpile.com/c/yikrbz/s89ia), the analysis for moderation was conducted by introducing an interaction effect between expert trust and the two treatment conditions, resulting in the model form:

$blame\_{}judge=α+ \sum\_{i = 1}^{2}β\_{i}(treatment\_{i})+expert trust+\sum\_{i = 1}^{2}𝜸\_{i}(treatment\_{i} \* expert trust) $

where $γ\_{i}$ is the change in the effect of the treatment based on the level of trust in experts.

Table A13 shows the results of this analysis. Trust in experts appears to have a significant moderation effect when respondents are given the treatment where the judge disagrees with the algorithm. For those with low trust in experts, disagreement with the algorithm does not increase blame of the judge. Indeed, it is a little less than the control condition for these respondents. This changes quite rapidly, however, and the amount of blame placed on the judge becomes higher than in the control condition for respondents with higher trust in experts. Given that the median level of trust in experts is 0.667 (on a scale from 0 to 1), this may provide some explanation for why we found in the other studies that disagreeing with the algorithm increases blame. This is consistent with the hypothesis presented by [(Surden 2021)](https://paperpile.com/c/yikrbz/ZQnKK), but with an important caveat about the difference between types of individuals who show increased blame in this condition. We find no evidence, however, that trust in experts affects the amount of blame placed on the judge when they agree with the algorithm.

|  |
| --- |
|  | *Dependent variable:* |
|  |  |
|  | Blame on judge |
|  |
| Agrees with algorithm | 0.083 |
|  | (0.081) |
|  |  |
| Disagrees with algorithm | -0.231\*\*\* |
|  | (0.083) |
|  |  |
| Trust in experts | -0.066 |
|  | (0.090) |
|  |  |
| Agrees with algorithm \* trust in experts | -0.085 |
|  | (0.119) |
|  |  |
| Disagrees with algorith \* trust in experts | 0.375\*\*\* |
|  | (0.121) |
|  |  |
| Female | 0.014 |
|  | (0.017) |
|  |  |
| Democrat | 0.051\*\* |
|  | (0.025) |
|  |  |
| Republican | 0.068\*\*\* |
|  | (0.025) |
|  |  |
| Income | 0.004\*\*\* |
|  | (0.001) |
|  |  |
| White | -0.037\* |
|  | (0.022) |
|  |  |
| Education | -0.003 |
|  | (0.005) |
|  |  |
| Constant | 0.615\*\*\* |
|  | (0.066) |
|  |  |
|  |
| Observations | 1,359 |
| R2 | 0.031 |
| Adjusted R2 | 0.023 |
| Residual Std. Error | 0.311 (df = 1347) |
| F Statistic | 3.956\*\*\* (df = 11; 1347) |
|  |
| *Note:* | \*p<0.1\*\*p<0.05\*\*\*p<0.01 |

**Table A14: Test for moderation of treatments by trust in experts**

# New Trust in Experts table based on Bansak’s method for causal moderation testing.

|  |
| --- |
|  | *Dependent variable:* |
|  |  |
|  | Blame on judge |
|  | Control | Agrees | Disagrees |
|  | (1) | (2) | (3) |
|  |
| Trust in experts | -0.074 | -0.167\* | 0.333\*\*\* |
|  | (0.093) | (0.089) | (0.091) |
|  |  |  |  |
| Female | -0.026 | 0.051 | 0.015 |
|  | (0.030) | (0.032) | (0.029) |
|  |  |  |  |
| Democrat | 0.067 | 0.014 | 0.067 |
|  | (0.042) | (0.045) | (0.042) |
|  |  |  |  |
| Republican | 0.070 | 0.024 | 0.108\*\* |
|  | (0.043) | (0.045) | (0.043) |
|  |  |  |  |
| Income | 0.004\* | 0.003 | 0.006\*\* |
|  | (0.002) | (0.002) | (0.002) |
|  |  |  |  |
| White | -0.032 | -0.037 | -0.036 |
|  | (0.037) | (0.040) | (0.037) |
|  |  |  |  |
| Education | -0.008 | 0.011 | -0.012 |
|  | (0.009) | (0.010) | (0.009) |
|  |  |  |  |
| Constant | 0.656\*\*\* | 0.664\*\*\* | 0.371\*\*\* |
|  | (0.076) | (0.085) | (0.074) |
|  |  |  |  |
|  |
| Observations | 439 | 443 | 477 |
| R2 | 0.018 | 0.025 | 0.062 |
| Adjusted R2 | 0.002 | 0.010 | 0.048 |
| Residual Std. Error | 0.305 (df = 431) | 0.320 (df = 435) | 0.310 (df = 469) |
| F Statistic | 1.155 (df = 7; 431) | 1.609 (df = 7; 435) | 4.461\*\*\* (df = 7; 469) |
|  |
| *Note:* | \*p<0.1; >\*\*p<0.05; >\*\*\*p<0.01 |

We next move to analyzing the possible mediation effects in hypotheses 2 and 3. In the pre-registration, we proposed two potential mediators: expectations of efficacy and perceptions the judge abdicated their responsibility to make a judgment by using an algorithm. To measure the first, we asked respondents the extent to which they thought the judge should have made the correct decision. The question wording is as follows:

On the following scale, where 1 indicates "never" and 10 indicates "always", how often do you think the judge should be able to make the correct decision in this scenario?

o Never (1)

o (2)

o (3)

o (4)

o Sometimes (5)

o (6)

o (7)

o (8)

o (9)

o Always (10)

To measure the degree to which the respondent (retrospectively) views the judge as abdicating their responsibility to use their own judgment, we included two questions. The first asks the degree to which they believed the decision reflected the judge’s own judgment versus that of the algorithm.

In the scenario you read, to what degree do you think the decision reflected the judge's own evaluation of the case, or reflected the advice of another expert or algorithm?

On the scale below, where 1 indicates "completely the judge's own evaluation" and 10 indicates "completely the advice of another expert or algorithm," indicate your opinion.

o Completely judge's own evaluation (1)

o (2)

o (3)

o (4)

o (5)

o (6)

o (7)

o (8)

o (9)

o Completely the advice of another expert or algorithm (10)

The second asked about the degree to which they think the judge should use their own judgment versus relying on the advice of experts/algorithms.

In the scenario you read, some people argue that judges must use their own judgment in making sentencing decisions.

Others think that judges should defer to algorithms and/or experts in making sentencing decisions.

On the scale below, where 1 indicates "always defer to algorithms/experts" and 10 indicates "always use own judgment," where would you place your opinion?

o Always defer (1)

o (2)

o (3)

o (4)

o (5)

o (6)

o (7)

o (8)

o (9)

o Always use own judgment (10)

These questions were presented in a randomized order. In the pre-registration, we suggested combining these two questions into an index, since the first deals with the degree to which the judge is perceived as using their own judgment and the second is the degree to which the respondent thinks they should have used their own judgment. We found, however, that combining the two questions in an index created difficulties in interpreting the scales. As such, we will present these results separately, although they are quite similar in their conclusions.

Mediation analysis is conducted as described in the main paper. Some scholars have noted that there is the potential for inflation of the mediation effect in this type of analysis if $cov(e\_{1},e\_{3})\ne 0$. This is unlikely to be the case, since *M* is not randomly assigned, meaning that any unobserved variables that are correlated with both *M* and *Y* will bias the results [(Bullock, Green, and Ha 2010; Green, Ha, and Bullock 2010)](https://paperpile.com/c/yikrbz/ldES6%2BJ1gd7). This is an issue that is very difficult to address through design. Instead, for the situations in which we find a significant mediation effect, we follow the advice of [(Imai et al. 2011; Imai, Keele, and Tingley 2010; Imai, Keele, and Yamamoto 2010)](https://paperpile.com/c/yikrbz/8Pl30%2BYRNpK%2BCTt1u) and test the sensitivity of the results by simulating a potential confounder and determining the level of correlation necessary for the mediation effect observed to be zero. While this does not eliminate the problem – in particular, it only looks at pretreatment confounders and not the possible existence of confounders that are affected by the treatment and then confound the relationship between the mediator and outcome – it provides us with a sense of how sensitive the results are to confounding. Based on the limitations of this type of analysis, any results from mediation should be considered indicative and in need of further testing using alternative designs.

In what follows, we present each step in the mediation analysis for both treatments and both proposed mediators (as well as both measures of the second proposed mediator). Note that we also present a third regression equation for the effect of the treatment without the mediator. Although this is redundant, we include it for those who may be more familiar with, and have an easier time interpreting, techniques that include this third equation.

We start with agreement with the algorithm. The first proposed mediator is that use of an algorithm increases the expectations of accuracy. We start by looking at the effect of the treatment on these expectations. As Table A14 shows, there is no significant relationship between this treatment and the proposed mediator in the first model, meaning that we can stop the analysis at this point, since there will not be detected mediation without this link (and, indeed, carrying through the analysis, the ACME has a p-value = 0.32).

|  |
| --- |
|  | *Dependent variable:* |
|  |  |
|  | Expectations | Blame on judge |
|  | (1) | (2) | (3) |
|  |
| Expectations |  |  | 0.035\*\*\* |
|  |  |  | (0.005) |
|  |  |  |  |
| Agrees with algorithm | 0.139 | 0.026 | 0.022 |
|  | (0.135) | (0.021) | (0.021) |
|  |  |  |  |
| Female | -0.141 | 0.011 | 0.016 |
|  | (0.138) | (0.022) | (0.021) |
|  |  |  |  |
| Democrat | 0.359\* | 0.031 | 0.018 |
|  | (0.191) | (0.030) | (0.029) |
|  |  |  |  |
| Republican | 0.272 | 0.051 | 0.042 |
|  | (0.199) | (0.031) | (0.030) |
|  |  |  |  |
| Income | 0.017 | 0.003\*\* | 0.003\* |
|  | (0.011) | (0.002) | (0.002) |
|  |  |  |  |
| White | -0.031 | -0.037 | -0.036 |
|  | (0.173) | (0.027) | (0.026) |
|  |  |  |  |
| Education | -0.037 | 0.001 | 0.002 |
|  | (0.042) | (0.007) | (0.006) |
|  |  |  |  |
| Constant | 7.181\*\*\* | 0.580\*\*\* | 0.331\*\*\* |
|  | (0.281) | (0.044) | (0.057) |
|  |  |  |  |
|  |
| Observations | 883 | 883 | 883 |
| R2 | 0.010 | 0.013 | 0.061 |
| Adjusted R2 | 0.002 | 0.005 | 0.053 |
| Residual Std. Error | 1.999 (df = 875) | 0.313 (df = 875) | 0.305 (df = 874) |
| F Statistic | 1.291 (df = 7; 875) | 1.600 (df = 7; 875) | 7.141\*\*\* (df = 8; 874) |
|  |
| *Note:* | \*p<0.1\*\*p<0.05\*\*\*p<0.01 |

**Table A15: Regression test for mediation between agreeing with algorithm and expectations that the judge should have made the correct decision**

Moving to the second proposed mediator, there are two measures that we will evaluate separately. The first asked the degree to which the respondent believed the judge used their own judgment versus that of the algorithm. In Table A15, we see that the results suggest a pretty substantial level of mediation. Respondents who saw the agreement with algorithm treatment are significantly more likely to say that the judge relied more on the algorithm than on their own judgment (p < 0.01), and, in turn, perceptions that the judge relied on the algorithm increase blame on the judge (p<0.01). The total ACME is 0.037 (p<0.01), though the proportion mediated should give some pause (1.388).

|  |
| --- |
|  | *Dependent variable:* |
|  |  |
|  | Who evaluated | Blame on judge |
|  | (1) | (2) | (3) |
|  |
| Who evaluated |  |  | 0.013\*\*\* |
|  |  |  | (0.004) |
|  |  |  |  |
| Agrees with algorithm | 2.831\*\*\* | 0.026 | -0.010 |
|  | (0.179) | (0.021) | (0.024) |
|  |  |  |  |
| Female | -0.170 | 0.011 | 0.013 |
|  | (0.183) | (0.022) | (0.021) |
|  |  |  |  |
| Democrat | -0.009 | 0.031 | 0.031 |
|  | (0.254) | (0.030) | (0.030) |
|  |  |  |  |
| Republican | -0.115 | 0.051 | 0.053\* |
|  | (0.265) | (0.031) | (0.031) |
|  |  |  |  |
| Income | 0.027\* | 0.003\*\* | 0.003\* |
|  | (0.014) | (0.002) | (0.002) |
|  |  |  |  |
| White | -0.264 | -0.037 | -0.034 |
|  | (0.229) | (0.027) | (0.027) |
|  |  |  |  |
| Education | 0.015 | 0.001 | 0.001 |
|  | (0.056) | (0.007) | (0.007) |
|  |  |  |  |
| Constant | 3.962\*\*\* | 0.580\*\*\* | 0.529\*\*\* |
|  | (0.373) | (0.044) | (0.046) |
|  |  |  |  |
|  |
| Observations | 883 | 883 | 883 |
| R2 | 0.229 | 0.013 | 0.025 |
| Adjusted R2 | 0.223 | 0.005 | 0.016 |
| Residual Std. Error | 2.654 (df = 875) | 0.313 (df = 875) | 0.311 (df = 874) |
| F Statistic | 37.121\*\*\* (df = 7; 875) | 1.600 (df = 7; 875) | 2.756\*\*\* (df = 8; 874) |
|  |
| *Note:* | \*p<0.1\*\*p<0.05\*\*\*p<0.01 |

**Table A16: Regression test for mediation between agreeing with algorithm and assessment of who made the decision**

As noted earlier, there are substantial reasons to worry about confounding in mediation analysis. We tested the sensitivity of these results using the methods discussed above. Figure A5 shows the results of this sensitivity analysis. The correlation between the error terms ($ρ$) would need to be 0.1 in order for the 95% confidence intervals of the ACME to include 0. The substantive meaning of this is a little difficult to determine, and the use of cross-study comparisons is not available in this analysis. An equivalent, but more intuitive, manner of analyzing this is to look at the importance of a potential confounder in explaining the variation in the mediator and outcome variable. In this case, the $R^{2}$ value for such a confounder would need to be 0.01 for the residual variance and 0.0075 for the total variance. These values are not particularly high, which means we cannot be overly confident that these results are robust to confounding, and, as noted earlier, suggests the need for further analysis, possibly using alternative experiment designs.



**Figure A5. Sensitivity analysis for the AMCE of perceptions about whose judgment was used in the decision. (A) The relationship between the sensitivity parameter (**$ρ$**) and the AMCE. (B) Contour plot of the** $R^{2}$**on the mediator (*M*) and the outcome (*Y*).**

Moving to the second measure dealing with the degree to which the judge *should have* used their own judgment, we find similar results, with some relatively strong indication of mediation. There is a significant relationship between whether the judge agreed with the algorithm and perceptions that the judge should have used their own judgment (p<0.01) (Table A16). There is also a significant and positive relationship between the perception that the judge should have used their own judgment and the amount of blame placed on the judge (p<0.01). The estimated ACME is 0.017 (p < 0.01), with the proportion of the treatment effect mediated at 0.660.

|  |
| --- |
|  | *Dependent variable:* |
|  |  |
|  | Should evaluate | Blame on judge |
|  | (1) | (2) | (3) |
|  |
| Should evaluate |  |  | 0.019\*\*\* |
|  |  |  | (0.005) |
|  |  |  |  |
| Agrees with algorithm | 0.897\*\*\* | 0.026 | 0.009 |
|  | (0.140) | (0.021) | (0.021) |
|  |  |  |  |
| Female | -0.298\*\* | 0.011 | 0.017 |
|  | (0.143) | (0.022) | (0.021) |
|  |  |  |  |
| Democrat | 0.192 | 0.031 | 0.027 |
|  | (0.197) | (0.030) | (0.030) |
|  |  |  |  |
| Republican | 0.182 | 0.051 | 0.048 |
|  | (0.206) | (0.031) | (0.031) |
|  |  |  |  |
| Income | 0.005 | 0.003\*\* | 0.003\* |
|  | (0.011) | (0.002) | (0.002) |
|  |  |  |  |
| White | -0.090 | -0.037 | -0.036 |
|  | (0.179) | (0.027) | (0.027) |
|  |  |  |  |
| Education | 0.054 | 0.001 | 0.0001 |
|  | (0.043) | (0.007) | (0.007) |
|  |  |  |  |
| Constant | 6.444\*\*\* | 0.580\*\*\* | 0.455\*\*\* |
|  | (0.291) | (0.044) | (0.055) |
|  |  |  |  |
|  |
| Observations | 883 | 883 | 883 |
| R2 | 0.058 | 0.013 | 0.029 |
| Adjusted R2 | 0.050 | 0.005 | 0.020 |
| Residual Std. Error | 2.066 (df = 875) | 0.313 (df = 875) | 0.310 (df = 874) |
| F Statistic | 7.671\*\*\* (df = 7; 875) | 1.600 (df = 7; 875) | 3.259\*\*\* (df = 8; 874) |
|  |
| *Note:* | \*p<0.1\*\*p<0.05\*\*\*p<0.01 |

**Table A17: Regression test for mediation between agreeing with algorithm and assessment of the degree to which judge should have used own judgment**

Once again, these results are tempered by the sensitivity analysis. As above, the amount of correlation between the errors for the 95% confidence intervals of the AMCE to include 0 is 0.1. In terms of $R^{2}$ values, these would need to be 0.01 for the residual variance and 0.0091 for the total variance (Figure A6). Here again, the results should be taken with some caution, but they provide a nice starting place for further analysis and exploration.



**Figure A6. Sensitivity analysis for the AMCE of perceptions about whose judgment should have been used to make the decision. (A) The relationship between the sensitivity parameter (**$ρ$**) and the AMCE. (B) Contour plot of the** $R^{2}$**on the mediator (*M*) and the outcome (*Y*).**

Moving to discussion of where the judge disagreed with the algorithm, we find little evidence for any of the potential mediators. Table A17 shows the equations for the mediation of expectations of accuracy. The results show that the expectations of accuracy decline when a judge disagrees with the algorithm. However, as the third model shows, expectations are positively and significantly related to the level of blame placed on the judge. The ACME, while significant (p < 0.05), is in the wrong direction for explaining the main outcome. We conclude, therefore, that the increase in blame for the mistake in this treatment is not due to this change in expectations.

|  |
| --- |
|  | *Dependent variable:* |
|  |  |
|  | Expectations | Blame on judge |
|  | (1) | (2) | (3) |
|  |
| Expectations |  |  | 0.014\*\*\* |
|  |  |  | (0.005) |
|  |  |  |  |
| Disagrees with algorithm | -0.365\*\*\* | 0.018 | 0.023 |
|  | (0.132) | (0.020) | (0.020) |
|  |  |  |  |
| Female | -0.366\*\*\* | -0.006 | -0.001 |
|  | (0.134) | (0.021) | (0.021) |
|  |  |  |  |
| Democrat | 0.224 | 0.076\*\*\* | 0.073\*\* |
|  | (0.187) | (0.029) | (0.029) |
|  |  |  |  |
| Republican | 0.205 | 0.086\*\*\* | 0.083\*\*\* |
|  | (0.198) | (0.031) | (0.031) |
|  |  |  |  |
| Income | 0.021\*\* | 0.005\*\*\* | 0.005\*\*\* |
|  | (0.011) | (0.002) | (0.002) |
|  |  |  |  |
| White | 0.172 | -0.032 | -0.034 |
|  | (0.169) | (0.026) | (0.026) |
|  |  |  |  |
| Education | 0.045 | -0.007 | -0.007 |
|  | (0.041) | (0.006) | (0.006) |
|  |  |  |  |
| Constant | 6.812\*\*\* | 0.570\*\*\* | 0.474\*\*\* |
|  | (0.269) | (0.042) | (0.054) |
|  |  |  |  |
|  |
| Observations | 917 | 917 | 917 |
| R2 | 0.035 | 0.026 | 0.034 |
| Adjusted R2 | 0.028 | 0.018 | 0.025 |
| Residual Std. Error | 1.989 (df = 909) | 0.309 (df = 909) | 0.307 (df = 908) |
| F Statistic | 4.730\*\*\* (df = 7; 909) | 3.438\*\*\* (df = 7; 909) | 3.986\*\*\* (df = 8; 908) |
|  |
| *Note:* | \*p<0.1\*\*p<0.05\*\*\*p<0.01 |

**Table A18: Regression test for mediation between disagreeing with algorithm and expectations that judge should have made the correct decision**

In terms of the degree to which respondents thought the judgment reflected the opinion of the algorithm, we see an expected negative relationship, with people attributing less of the judgment to the algorithm when the judge disagrees, but there is no significant impact of this on blame placed on the judge (Table A18). Not surprisingly, the ACME estimate is also not statistically significant (p = 0.32).

|  |
| --- |
|  | *Dependent variable:* |
|  |  |
|  | Who evaluated | Blame on judge |
|  | (1) | (2) | (3) |
|  |
| Who evaluated |  |  | 0.004 |
|  |  |  | (0.004) |
|  |  |  |  |
| Disagrees with algorithm | -0.399\*\* | 0.018 | 0.020 |
|  | (0.178) | (0.020) | (0.020) |
|  |  |  |  |
| Female | -0.715\*\*\* | -0.006 | -0.003 |
|  | (0.180) | (0.021) | (0.021) |
|  |  |  |  |
| Democrat | 0.251 | 0.076\*\*\* | 0.075\*\* |
|  | (0.252) | (0.029) | (0.029) |
|  |  |  |  |
| Republican | -0.063 | 0.086\*\*\* | 0.086\*\*\* |
|  | (0.268) | (0.031) | (0.031) |
|  |  |  |  |
| Income | 0.028\* | 0.005\*\*\* | 0.005\*\*\* |
|  | (0.014) | (0.002) | (0.002) |
|  |  |  |  |
| White | -0.097 | -0.032 | -0.031 |
|  | (0.227) | (0.026) | (0.026) |
|  |  |  |  |
| Education | 0.051 | -0.007 | -0.007 |
|  | (0.055) | (0.006) | (0.006) |
|  |  |  |  |
| Constant | 3.821\*\*\* | 0.570\*\*\* | 0.555\*\*\* |
|  | (0.363) | (0.042) | (0.044) |
|  |  |  |  |
|  |
| Observations | 917 | 917 | 917 |
| R2 | 0.038 | 0.026 | 0.027 |
| Adjusted R2 | 0.031 | 0.018 | 0.018 |
| Residual Std. Error | 2.683 (df = 909) | 0.309 (df = 909) | 0.308 (df = 908) |
| F Statistic | 5.178\*\*\* (df = 7; 909) | 3.438\*\*\* (df = 7; 909) | 3.154\*\*\* (df = 8; 908) |
|  |
| *Note:* | \*p<0.1\*\*p<0.05\*\*\*p<0.01 |

**Table A19: Regression test for mediation between disagreeing with algorithm and assessment of who made the decision**

Finally, we looked for possible mediation based on the degree to which the respondent thought the judge should have used their own judgment. Here again, there is little support for any mediation. A judge disagreeing with the algorithm has no significant impact on the evaluation of respondents about the degree to which the judge should have used their own judgment. Not surprisingly, the ACME is also not significant (p = 0.22) (Table A19).

|  |
| --- |
|  | *Dependent variable:* |
|  |  |
|  | Should evaluate | Blame on judge |
|  | (1) | (2) | (3) |
|  |
| Should evaluate |  |  | -0.016\*\*\* |
|  |  |  | (0.005) |
|  |  |  |  |
| Disagrees with algorithm | -0.155 | 0.018 | 0.016 |
|  | (0.138) | (0.020) | (0.020) |
|  |  |  |  |
| Female | -0.489\*\*\* | -0.006 | -0.014 |
|  | (0.140) | (0.021) | (0.021) |
|  |  |  |  |
| Democrat | 0.522\*\*\* | 0.076\*\*\* | 0.084\*\*\* |
|  | (0.196) | (0.029) | (0.029) |
|  |  |  |  |
| Republican | 0.500\*\* | 0.086\*\*\* | 0.093\*\*\* |
|  | (0.208) | (0.031) | (0.031) |
|  |  |  |  |
| Income | 0.002 | 0.005\*\*\* | 0.005\*\*\* |
|  | (0.011) | (0.002) | (0.002) |
|  |  |  |  |
| White | -0.029 | -0.032 | -0.032 |
|  | (0.177) | (0.026) | (0.026) |
|  |  |  |  |
| Education | 0.061 | -0.007 | -0.006 |
|  | (0.043) | (0.006) | (0.006) |
|  |  |  |  |
| Constant | 6.234\*\*\* | 0.570\*\*\* | 0.668\*\*\* |
|  | (0.282) | (0.042) | (0.051) |
|  |  |  |  |
|  |
| Observations | 917 | 917 | 917 |
| R2 | 0.030 | 0.026 | 0.037 |
| Adjusted R2 | 0.022 | 0.018 | 0.028 |
| Residual Std. Error | 2.086 (df = 909) | 0.309 (df = 909) | 0.307 (df = 908) |
| F Statistic | 4.002\*\*\* (df = 7; 909) | 3.438\*\*\* (df = 7; 909) | 4.324\*\*\* (df = 8; 908) |
|  |
| *Note:* | \*p<0.1\*\*p<0.05\*\*\*p<0.01 |

**Table A20: Regression test for mediation between disagreeing with algorithm and assessment of the degree to which judge should have used own judgment**

Based on the above analysis, there is some indication of moderation and mediation in these results. As we have noted, however, there are reasons to be concerned about the degree to which we can definitely say that these are the only or the main reasons for the results. This was, obviously, also not helped by the aberrant main result in this sample. With these important caveats, the results do give some indications about the reasons why we observe the results that are at odds with what the ethics and legal literature predict and provide a clearer starting point for future analysis.

**SI.14. Comments from Respondents**

Another method for trying to understand what is happening in the experiment is to look at the comments left by participants. For these comments, participants were asked at the end of the survey whether they had anything to communicate with the research team. The question did not specify anything in terms of what they were trying to communicate, and we did not ask specifically about the treatment. Nevertheless, some participants offered up their opinion. These were asked on both Lucid surveys (but not on the NORC survey, since there was not space). A sample of these statements is shown in Table A20. It should also be noted that the initial Lucid survey had a second experiment included, so some of the comments may refer to issues outside of this experiment.

We tried to select quotations based on relevance, removing any identifying information, not based on our particular findings. We have also refrained from editing them for spelling or punctuation. Nevertheless, it is interesting to see how many of the comments are along similar lines to the factors we were testing in the moderation and mediation survey, with a number of respondents emphasizing the duty of judges to use their own judgment in deciding cases like this.

| **Initial Lucid Survey** |
| --- |
| “I still think that a human decision should be made after looking at the outcome of the algorithm” “Although algorithms are a good idea, in general, creating a good algorithm is entirely a different matter. Just because a decision is based on support of an algorithm doesn’t necessarily make that decision a good one.”“An algorithm will not reach human intuition”“You can just see how great algorithms worked in 2016 before & during the presidential and other elections. You can never nor should anyone try and predict what another human would do…”“I don't trust a computer to make life changing decisions”“Bottom line is while the public is involved in trials there are a lot of factors that make it difficult to come to a logical decision on many cases and there is probably times breaks should be given and the traits of each person tried should be considered as well which I a sure in many cases that is the case.”“Algorithms are fine, but they cannot be the sole factor in a decision. You cannot exclude the human factor”“When deciding on a court case, I would definitely like to have the human factor involved in the final decision. It is good to have every helpful tool possible but in the end, I want the human to decide.”“we are humans which use logic and gut feelings not algarithms”“Data is important, algorithms are important -- but human intuition is important also.”“Humans make better decisions than computers, because there are many nuances to life's problems.”“I would never be in favor of any such technology as a deciding factor for something this important.” “There is no substitute for human decision”“The judge be the one in charge of everything.” |
| **Moderation and Mediation Lucid Survey** |
| “Using a scientific formula for all people is dangerous because no two people are the same. I don't believe the judge should have not solely used this form of judgement”“Not every situation has a good answer. But computers do not have the ability to empathize with defendant's not victims. Computers should never decide somones fate entirely.”“Judges have some tuff dissection”“Yes I think next time the judge take a second opinion in most cases like these.”“The judge was not at fault. The judge should be able to look at all options, then use their best judgement. If we let the algorithm set judgement, then we have lost our humanity. That's what makes a good judge great, humanity. The defendant did not deserve prison for their crime. The Judge can't punish someone for something they haven't done yet. Even if we knew what the future held, because at the time the defendant still had not committed the 3rd deadly crime.”“Very interesting topic. I mostly disagree with using a computer based system to determine a person's level of punishment. each punishment should be individualized by another person in charge of making recommendations.”“I believe that the judge ,as always, should be the final decision maker in all cases”“I think the judge should have considered the defendant's previous infraction and the computer algorithm at the same time.”“Why didn't the judge trust the computer”“The defendant was 100% to blame, not the Judge.”“peoples lives should never be left to a machine”“Fascinating! Neither computer algorithms nor human judges can be right all the time. Neither were to blame, because only the drunk driver was responsible. He probably should have gone into rehab or AA.”“I would like to say that judges should take into consideration all the tools available to them including their own good judgement and that is the reason I answered as I did”“Although taking past information about the outcome of situations in aggregate is helpful, it doesn't take into account the person involved in a current case. I believe a judge should use the information provided but based on the dealing with the person the judge should form their decision for punishment.”“I would think that because of the mentioned so called standing of the defendent that the judge should have found other sources of info and not make a bad decision. Just because someone is rich and upstanding does not mean they can't make mistakes and should pay for them.”“The judge needs to talk to the defendant and make his choice based on his gut feelings, that how I make my decisions.”“Although algorithms may be pretty accurate, human behavior is not 100% predictable.”“As someone with a criminal justice background, I have become utterly contemptuous of "experts" and their algorithms. Far too often, these promise far more than they can deliver, and do so with an "authority" that does not reflect their reliability. Maybe someday that will change, but I doubt it'll happen in my lifetime.”“The Judge is in charge of his court and he/she should make the final choice as to what will happen”“My two cents - To be a judge is not an easy job. Although they are supposed to have the experience, knowledge, morals, ethics etc. they are human beings and therefore not perfect. Similarly, computers are programmed by humans and there is also room for error. Although it would seem that the algorithm was "correct" we should not allow this to determine someone's fate as there are many variables that should be considered.”“Yes researchers, I think in this case the Judge should go on trial for neglect. Just because this guy was an upstanding citizen doesn't mean he is not capable of drinking too much as obvisouly he did and killed a person. This is entirely son the Judges shoulder and conscious. At least it sure should be.”“The judge should have taken into account that this driver has a more serious drinking problem despite his or her college degree, background, or previous lack of a serious criminal background. Thanks for the thoughtful and thought provoking exercise.”“I, personally, believe that it is impossible for judges to eliminate their personal biases (positive or negative) while making decisions. There should be some checks and balances in place to show trends. If a judge is consistently ignoring all algorithms and expert advice; someone should ask "why".”“I can't believe an algrorithm would be used in the court of law.”“I think it's important for me to add that I in no way believe the judge or the algorithm is at fault for the outcome of the scenario posed. I believe it was the choices of the perpetrator that are at fault and that fault cannot be placed on anything or anyone else, nor is it appropriate for it to be.”“A judge is in place for a reason. If we relied on an algorithm or the feedback of others, not everything would be handled specifically to the case/person. I do believe that there is some blame on the judge in this situation, but in general, this person who was arrested is at fault for their own actions.”“I hope future decisions do take into consideration computer algorithms to ensure the law upholds justice for the safety and protection of the people.”“I think in this instance, the judge should use both logic and algorithms to make the most conscious decision. Not that the judge knew that the person was going to continue to make poor choices again that would in turn end the life of another person but should use also other scenarios from previous cases to help argue that the crime should have been more punishable.”“Judges are trained and highly educated, and have their own experience with the topic at hand. Depending on the judge to be totally acceptable can be weighed in his own experience and not a bunch of data. His own personal experiences are more important' God made him human just like the rest of us. Who are we to cast doubt.”“Algorithms are good for helping identify possible outcomes, but human nature can be very unpredictable.”“This survey is interesting but difficult to answer because I know both that judges are worse than average at predicting reoffense in the criminal justice system and that algorithms are often incredibly racist because of their training with racially biased data. I wish there was an option to indicate I trust neither of these actors.”“Where the algorithms fail, is Not having the ability to consider all the human factors that a judge may consider. The algorithms are without empathy and the ability to actually speak with the defendant to ask pertinent questions, eg. how often do you drink, have you ever had a psychological profile. Were you abused or are you being abused. The algorithm cannot look a defendant in the eye, experienced judges have a knack for character evaluations”“That was a really tough scenario with hindsight being 20/20 and all that. It is hard to know whether the computer algorithm would always be right and there is really no good (or ethical) way to test that theory in the real world. I will be thinking about this scenario for some time to come.”“On a first time offense algorithm would be logical to help decide, but this was a second offense so as a repeat offender a harsher penalty to include counseling would be warranted”“I think public evaluations of judicial decision-making can vary depending on the information they are given. What criteria did the judge use for their decision in this situation? How long as the prison term the algorithm suggested?”“Judges should never rely on computer. They should listen to the testimony and make there own decision.” “This was difficult . I truly believe that each situation is different . Itâ€™s most important to use critical thinking , common sense and professional opinions in each case . In that order !!!”“A judge should rely on his judgement the most”“A judge will always reflect on his own experience while on the bench. But sometimes they still get it wrong. It's because their human and humans always make mistakes. It's if they learn from those mistakes or not is what matters!!”“the judges should look at experts advice before making a decision on a case.”“Judges ultimately make the final decision which should be guided by a number of factors/influences.”“they hold the office”“Who is to blame? The DWI person, stop trying to break down the blame!”“This example brings up the use of advanced artificial intelligence & how it affects much of our lives today, Each event much be evaluated more on common sense than bars & graphs.”“Everyone blames everyone else except the main person -- the person himself took his own action to drink and drive....no one else is at fault”“You asked in the first question: "Who is to blame?" I wondered what you meant. Who is to blame FOR WHAT, exactly? The first DUI? (The driver.) The 2nd DUI? (The idiot judge and the driver.) Or who was to blame for killing someone in the 3rd offense? (Absolutely the judge that let the driver off for the first 2 DUIs, and ABSOLUTELY the driver!) Meh. The judge should not have let the driver off so easily after the FIRST offense.”“The Chevy was good I would just like to mention though that considering the guys good record on everything else other than the drinking and driving the judge did not have the responsibility of the man choosing to drink and drive again and killing someone this time judge was giving him a chance and I think everybody deserves a chance”“Trusting a computer algorithm to make choices? In this case, even though the driver is a hard worker, benevolent, etc. Driving drunk for a third time and killing somebody is definitely not deserving of a light sentence. The driver did that three times. If he never learned first time, he will never learn again.”“I truly hope we aren't headed towards using computer algorithms to guide/make such decisions going forward”“The question asking who was at fault I feel was worded strangely- I don't think an algorithm could be called 'at fault' for anything, and the driver was certainly the most at fault”“Why didn't you allow the computer scientists and the criminal justice experts to share the blame since they altogether created the algorithm?”“That is why I don't trust machines.”“Humans are far better at evaluating a person's future behavior than a computer.”“the judge was appointed or elected to make those decisions HE SOULD MAKE THEM”“I think a judge needs to look at more than just algorithms, and use their common sense. To me, it should be 50/50, with the judge being able to rule not solely on the rules or algorithms, but based on the defendants appearance in court, as well as their support systems. We are smarter than a computer, you have to use common sense sometimes. Being a judge should mean that you are able to judge one's character, as well as the normal rules, not just one or the other and to go based off of your brain and common sense. Not all situations are the same, something a computer or algorithm would not be able to figure out.”“The judge should not tolerate drunk driving and give stiff penalty. Until that happens people will flaunt the rules as they know they will get off with a light sentence”“I do not think that the judge was necessarily wrong. There are many additional details that the judge may have had that we don't.”“Algorithms, computer scientist and expert can give some aide; however, the judge and person have to used Godly wisdom. Along with the judge sensitivity level and moral compass. Algorithms aren't human and experts also have to have wisdom, high moral compass, and will to listen to truth beyond their intelligent.”“In this new era of corruption, it is scary to think any judge especially a liberal judge, has enough common sense to make any worthwhile decision!”“Yeah I do I support the fact and hardwork of the researchers but I Also support the judges if he or she uses this method it no longer his or her court”“I think there is a level to decision making that we can't program a computer to do yet. We can only input so many variables into an algorithm.”“If it were up yo a computer, why have judges?”“Leaving decisions to a computer leaves far to many variables left unconsidured” “What about the guys who wrote the algorithm?”“Ultimately it is up to the judge to weigh the advice of the algorithms imput with the circumstances of the individual to make a ruling.”“Judge must be the one to judge by the letter of the law if he can’t then we do need them”“Judges are there to make a human decision and uphold justice, an algorythm cant do that. Algorythms are easily manipulated, and since outsourcing our prison system is becoming common, an entire industry has a vested interest in imprisoning people...and we already imprison more people per capita than any other nation on earth.”“The mitigating factors in the case seemed to support the judge simply going along with the computer algorithms recommended sentence. It was the offenders second offense. The judge could have been harsher, but that was his/her judgment to make. The judge could not have known there would be a third offense resulting in the death of someone.”“Critiques of criminal justice algorithms and algorithmic decisionmaking are central to my work, so I feel like I was kind of tripped up by this. Criminal justice systems are historically biased, algorithms bake that bias in, and the absence of demographic information about the defendent seemed really problematic in your study design here, because I know that such algorithms overpredict lack of re-offense in white offenders and underpredict it on offenders of color.”“Sentencing guide lines are in place everywhere. The judge has the final say”“I thought the judge should have gone beyond probation with a temporary driving suspension and driving restrictions during the time of probation. It may not have prevented the 3rd DUI, but I think it would lowered the likelihood of it happening. The judge should have been more concerned about the 2nd DUI closely following the 1st.” “I feel there should be multiple supports, sources and professionals to evaluate and judge a person in seeking the appropriate charges and treatments.”“The algorithm could have reprimanded the accused with a suspended liscense and the judge might have not reacted in a completely opposite way.”“A person should be HELD to their responsibilities or choices that THEY alone make. Using the law or making a deal to get a lighter sentence. Should Not even be on the books. I do not care who it is. We are supposed to be equal.”“Not one question concerning the drunk drivers choices or responsibility. Nothing is cool proof. You could take away a drivers license on first offence and they could drive anyway. That has become the direction of this country. Blame it on the computer, the judge of the gun and never on the actions of behavior of the indiviual”“In my opinion the judge did exactly right because the law of averages will insure that every person will not make the same mistakes so the judge riding towards the middle of both recommendations is probably the safest route for his position!”“"Computer algorithms" get a bad rap. However, I deal all the time with sentencing assessment reports in my work (I'm a local news reporter) and I'm not sure, in the real world, whether there is a whole lot of difference between a SAR produced by human beings who use standard methods of evaluating people's past offenses and likelihood to reoffend and computers that do the same thing.”“I would have liked to have known the judge's and the offender's race, religion, sex and orientation to factor into my assessment as to the judge's mindset; meaning considering possible contributing factors such as bias or prejudice. Ex: If the judge is a straight white educated male, they might relate to and be more lenient towards a straight white educated male offender.”“I don't know how much this has to do with the survey, but both judges and algorithms can be biased when using evidence to draw conclusions, so it's important to identify and address these biases before making decisions.” |

**Table A20: Sample of comments from respondents in Lucid samples**

**SI.15. IRB Review and Ethics Compliance**

In response to the American Political Science Association’s (APSA) *A Guide to Professional Ethics in Political Science* (2012) and *Guidance for Human Subjects Research* (2020), this section lays out the review, consent, and other ethics procedures for this project.

First, as with any research involving human subjects, this project was reviewed by the [institution omitted for blind review] Committee for the Projection of Human Subject (CPHS, the university’s name for the Institutional Review Board - IRB). Study 1 underwent review and approval (STUDY00001247) and Study 2 was approved as a modification (MOD00004154). This included making sure that participation was voluntary and that informed consent was acquired prior to their participation.

In the Lucid experiments,, the first screen seen by the participant is the informed consent agreement. It read as follows, and included the study information, details about risks and rewards, as well as contact information for reporting issues. Respondents were required to click “yes” prior to starting the survey. Those who clicked “no” were immediately returned to the Lucid service. Following the advice of scholars who have criticized the standard informed consent wording as incomprehensible to most participants [(Kadam 2017)](https://paperpile.com/c/yikrbz/5Iib), we started with a simplified version of the consent form to make it more likely for participants to read and comprehend the informed consent, followed by a “more information” option that took them to the formal legal version of the informed consent. The survey was listed as taking 15 minutes, though in our pre-testing average completion time was around 10 minutes.

Consent:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

This experiment is being conducted by [omitted for peer review] at [omitted for peer review] and [omitted for peer review] at [omitted for peer review], who are interested in how people utilize information in making predictions about the likelihood of future events.

**Before you participate, we need to let you know a few things about participating in academic research like this.**

Here goes:

You are being invited to participate in an online survey about assigning blame to individual decision makers. This session is being facilitated by [omitted for peer review at [omitted for peer review]. If you are interested in participating, you can click on the link below and you will be taken directly to the survey. The survey will last approximately 15 minutes.

Your participation will be used in research to better understand how people utilize information in making forecasts. Your individual identity will be kept confidential and your responses will be stored in a secure data server. Any report of this research that is made available to the public will not include your name or any other individual information by which you could be identified. There are no known risks or discomforts associated with participating in this session. Taking part in this study is completely voluntary. If you choose to be in the study you can withdraw at any time without any consequences.

Participants may be compensated for their participation. To receive compensation, however, the participant must correctly answer questions that check whether they are paying attention during the survey, have read proscribed materials, or have acquired the knowledge needed to complete the task.

If you would like to know more about the project or have any questions, you can find a more detailed description by choosing the "more information" option below or you can contact [omitted for peer review] at [omitted for peer review]. This project has been reviewed by the [omitted for peer review] Committee for the Protection of Human Subjects [CPHS phone number]. If you have any questions about your rights as a research participant or if you have any complaints about the conduct of this research, contact the [omitted for peer review] Committee for the Protection of Human Subjects. All research projects that are carried out by Investigators at [omitted for peer review] are governed by requirements of the University and the federal government.

If you agree to participate, please click "yes" below and continue to the next screen to take the survey.

\_\_ yes

\_\_ no [if selected, takes back to Lucid homepage]

\_\_ more information [if selected, taken to next screen below]

[If more information selected]

WELCOME

THIS PAGE PROVIDES DETAILED INFORMATION ABOUT THE CONSENT TO PARTICIPATE IN RESEARCH AT [INSTITUTION OMITTED FOR PEER REVIEW].

[INSTITUTION OMITTED FOR PEER REVIEW]

CONSENT TO PARTICIPATE IN RESEARCH

Title of research study: Blame Attribution in Criminal Sentencing

Investigator: [omitted for peer review] ([omitted for peer review]) and [omitted for peer review] ([omitted for peer review])

Why am I being invited to take part in a research study?

We invite you to take part in a research study because you have verified that you are eligible by your survey provider and clicked the link to our survey.

This research is being funded by the Intelligence Advanced Research Projects Activity (IARPA) and the National Science Foundation (NSF).

What should I know about a research study?

· Someone will explain this research study to you.

· Whether or not you take part is up to you.

· You can choose not to take part.

· You can agree to take part and later change your mind.

· Your decision will not be held against you.

· You can ask all the questions you want before you decide, and can ask questions at any time during the study.

Why is this research being done?

This study attempts to collect information about how people use information presented to them in an online system to make decisions, in this case about assigning blame in cases of criminal recidivism.

How long will the research last?

We expect that you will be in this research study for 15 minutes.

How many people will be studied?

We expect to enroll up to 8,000 people in this research study.

What happens if I say yes, I want to be in this research?

You will be taken to the survey.

What happens if I do not want to be in this research?

You can click “no” and you will be taken back to your survey provider’s website.

What happens if I say yes, but I change my mind later?

You can leave the research at any time it will not be held against you.

Is there any way being in this study could be bad for me?

There are no foreseeable risks related to the procedures conducted as part of this study. If you choose to take part and undergo a negative event you feel is related to the study, please inform your study team.

Will I get anything for being in this study?

You will receive compensation in accordance with your agreement with the survey provider.

Will being in this study help me in any way?

There are no known benefits to you from your taking part in this research. However, possible benefits to others include a better understanding of how advice is used in public-policy decision-making and potential policies to improve the use of information.

What happens to the information collected for the research?

Your taking part in this project is anonymous, and information you provide cannot be linked to your identity.

Who can I talk to?

If you have questions, concerns, or complaints, or think the research has hurt you, you should talk to the research team at [email of PI].

This research has been reviewed and approved by the [omitted for peer review] Institutional Review Board (IRB). You may also talk to them at [phone number]] or[email]:

· Your questions, concerns, or complaints are not being answered by the research team.

· You cannot reach the research team.

· You want to talk to someone besides the research team.

· You have questions about your rights as a research subject.

· You want to get information or provide input about this research.

If you agree to participate, please click "yes" below and continue to the survey.

\_\_ yes

\_\_ no [if selected, taken back to the Lucid homepage]

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Lucid charges a flat fee per respondent, and utilizes respondents drawn from a range of panels. For each respondent, Lucid charges $1 per response (<https://luc.id/academic-solutions/>). Respondents are remunerated for their participation in a variety of different ways, including direct rewards, loyalty points, and online games, depending on what the respondent has chosen. Unfortunately, Lucid does not disclose what each participant receives as compensation for each of their respondents, whether that be monetary, loyalty points, online games, etc. Respondents were provided with details of their reward prior to choosing to participate in the survey. We rely on the individual’s judgment on whether this compensation is worth their time. We did note one participant who asked us to talk with the survey provider about increasing payments for participation in these surveys and one complaint that it was too long, but also noted 152 comments that they enjoyed participating in the survey and would be interested in doing another like it in the future..

Since our surveys involved attention checks, we attempted to make these as non-onerous as possible. In Study 1, all respondents who completed the survey were allowed to receive the reward, and we excluded them afterwards. For Study 2, a standard attention check [(Berinsky, Margolis, and Sances 2014)](https://paperpile.com/c/yikrbz/0uxa0) was used with some modification. It was the second question after the consent, and it returned the respondent to their survey provider’s page without receiving their reward. We included it early soas to avoid taking too much of a respondent’s time without remuneration.

For the Time-sharing Experiments for the Social Science (TESS) study, which is done through the University of Chicago’s NORC Amerispeaks Panel, the full IRB information can be found here (<https://amerispeak.norc.org/SiteAssets/supportingDocs/NORC%20AmeriSpeak%20Information%20for%20IRBs%202016%2010%2018.pdf>). Participants in the NORC panel are consented into participation in the studies involved, and are only required to undergo further consent if the questions involve particularly sensitive topics. Our question was included as one of many in the time-sharing experiment. As a general rule, AmeriSpeak Panel members receive modest incentives, in the form of “AmeriPoints,” for participation in surveys. For the majority of surveys, a respondent will earn between 2,000 and 10,000 AmeriPoints (1,000 points = $1) for completing a survey. If a survey is very long (e.g., 30 minutes or longer), or we ask a respondent to participate in other custom research such as an indepth telephone interview, then they may receive an incentive of 10,000 points or more. A panel member is credited with earned points for a survey once they complete the survey.

Our survey did not involve explicit deception. Respondents were told the following before responding to the scenario:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

On the next screen, you will be given a scenario dealing with a situation in the criminal justice system.

Please read the scenario carefully and answer the following questions about your reaction to that scenario.

Click the button below to proceed to the scenario.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

It was determined by the IRB that there was no need for debriefing, given that there was no attempt to portray the scenario as an actual event, nor to provide actual names of any involved that might result in retaliation on the part of the respondents.

The main risk identified by the IRB was the potential for a loss of confidentiality by the researchers. This risk was mitigated by the survey recruitment process. We did not receive any PID from the respondents on the Lucid panel, nor did we receive PID from those participating in the NORC AmeriSpeaks panel.

None of this research involved public actors or vulnerable communities. As we noted in our IRB review, the only vulnerable community included was, potentially, pregnant women, as we had no way to screen for pregnancy in the survey and it would be discriminatory to exclude them from the potential rewards for their participation by barring them from participating in an extremely low risk online activity.

**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 University Press.](http://paperpile.com/b/yikrbz/oXTH6)

[Angwin, Julia, Jeff Larson, Lauren Kirchner, and Surya Mattu. 2016. “Machine Bias — ProPublica.” *ProPublica*.](http://paperpile.com/b/yikrbz/uU13Q) <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> [(September 15, 2018).](http://paperpile.com/b/yikrbz/uU13Q)

[Arceneaux, Kevin, and Ryan J. Vander Wielen. 2017. *Taming Intuition: How Reflection Minimizes Partisan Reasoning and Promotes Democratic Accountability*. Cambridge University Press.](http://paperpile.com/b/yikrbz/yt8HB)

[Aronow, Peter Michael, Joshua Kalla, Lilla Orr, and John Ternovski. 2020. “Evidence of Rising Rates of Inattentiveness on Lucid in 2020.”](http://paperpile.com/b/yikrbz/TUW0i)

[Baron, Reuben M., and David A. Kenny. 1986. “The Moderator–mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations.” *Journal of personality and social psychology* 51(6): 1173–82.](http://paperpile.com/b/yikrbz/GYRu3)

[Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. “Evaluating Online Labor Markets for Experimental Research: Amazon.com’s Mechanical Turk.” *Political analysis: an annual publication of the Methodology Section of the American Political Science Association* 20(3): 351–68.](http://paperpile.com/b/yikrbz/7MDEw)

[Berinsky, Adam J., Michele F. Margolis, and Michael W. Sances. 2014. “Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys.” *American journal of political science* 58(3): 739–53.](http://paperpile.com/b/yikrbz/0uxa0)

[Bertsou, Eri. 2021. “Bring in the Experts? Citizen Preferences for Independent Experts in Political Decision‐making Processes.” *European journal of political research* (1475-6765.12448).](http://paperpile.com/b/yikrbz/4vHhH) <https://onlinelibrary.wiley.com/doi/10.1111/1475-6765.12448>[.](http://paperpile.com/b/yikrbz/4vHhH)

[Bollen. 1989. *Structural Equations with Latent Variables*. 1st ed. Wiley-Interscience.](http://paperpile.com/b/yikrbz/OAK3t)

[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): 550–58.](http://paperpile.com/b/yikrbz/J1gd7)

[Chesterman, Simon. 2021. *We, the Robots?: Regulating Artificial Intelligence and the Limits of the Law*. Cambridge University Press.](http://paperpile.com/b/yikrbz/cuLHi)

[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.](http://paperpile.com/b/yikrbz/Wv7aE)

[Dressel, Julia, and Hany Farid. 2018. “The Accuracy, Fairness, and Limits of Predicting Recidivism.” *Science Advances* 4(1): eaao5580.](http://paperpile.com/b/yikrbz/Enw8e)

[Faul, Franz, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. “Statistical Power Analyses Using G\*Power 3.1: Tests for Correlation and Regression Analyses.” *Behavior research methods* 41(4): 1149–60.](http://paperpile.com/b/yikrbz/SUDX4)

[Fritz, Matthew S., and David P. Mackinnon. 2007. “Required Sample Size to Detect the Mediated Effect.” *Psychological science* 18(3): 233–39.](http://paperpile.com/b/yikrbz/Zqftj)

[Gerber, Alan et al. 2014. “Reporting Guidelines for Experimental Research: A Report from the Experimental Research Section Standards Committee.” *Journal of Experimental Political Science* 1(1): 81–98.](http://paperpile.com/b/yikrbz/SIJOz)

[Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. Illustrated edition. W. W. Norton & Company.](http://paperpile.com/b/yikrbz/s89ia)

[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.](http://paperpile.com/b/yikrbz/ldES6)

[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. The Guilford Press.](http://paperpile.com/b/yikrbz/4k6rW)

[Hoff, Kevin Anthony, and Masooda Bashir. 2015. “Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust.” *Human factors* 57(3): 407–34.](http://paperpile.com/b/yikrbz/GXuuG)

[Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. “A General Approach to Causal Mediation Analysis.” *Psychological methods* 15(4): 309–34.](http://paperpile.com/b/yikrbz/YRNpK)

[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): 765–89.](http://paperpile.com/b/yikrbz/8Pl30)

[Imai, Kosuke, Luke Keele, and Teppei Yamamoto. 2010. “Identification, Inference and Sensitivity Analysis for Causal Mediation Effects.” *Schweizerische Monatsschrift fur Zahnheilkunde = Revue mensuelle suisse d’odonto-stomatologie / SSO* 25(1): 51–71.](http://paperpile.com/b/yikrbz/CTt1u)

[Kadam, Rashmi Ashish. 2017. “Informed Consent Process: A Step Further towards Making It Meaningful!” *Perspectives in clinical research* 8(3): 107–12.](http://paperpile.com/b/yikrbz/5Iib)

[Kennedy, Ryan, Philip D. Waggoner, and Matthew Ward. 2022. “Trust in Public Policy Algorithms.” *The journal of politics*.](http://paperpile.com/b/yikrbz/zsXoH)

[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): 544–64.](http://paperpile.com/b/yikrbz/cYLW2)

[Logg, Jennifer M. 2016. “When Do People Rely on Algorithms?” University of California, Berkeley.](http://paperpile.com/b/yikrbz/4yv3N)

[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: 90–103.](http://paperpile.com/b/yikrbz/tMf6Z)

[Merkley, Eric. 2021. “Anti-Intellectualism, Populism, and Motivated Resistance to Expert Consensus.” *Public opinion quarterly* 84(1): 24–48.](http://paperpile.com/b/yikrbz/9YvWT)

[Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. 2018. “How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It.” *American journal of political science* 62(3): 760–75.](http://paperpile.com/b/yikrbz/QAm3j)

[Mutz, Diana C. 2011. *Population-Based Survey Experiments*. Princeton University Press.](http://paperpile.com/b/yikrbz/vDQeR)

[Mutz, Diana C., and Robin Pemantle. 2015. “Standards for Experimental Research: Encouraging a Better Understanding of Experimental Methods.” *Journal of Experimental Political Science* 2(2): 192–215.](http://paperpile.com/b/yikrbz/VO7b9)

[Open Science Collaboration. 2015. “Estimating the Reproducibility of Psychological Science.” *Science* 349(6251): aac4716.](http://paperpile.com/b/yikrbz/jVVn2)

[Soderberg, Courtney K. et al. 2021. “Initial Evidence of Research Quality of Registered Reports Compared with the Standard Publishing Model.” *Nature human behaviour* 5(8): 990–97.](http://paperpile.com/b/yikrbz/5rhUH)

[Surden, Harry. 2021. “Ethics of AI in Law: Basic Questions.” In *The Oxford Handbook of Ethics of AI*, eds. Markus D. Dubber, Frank Pasquale, and Sunit Das. Oxford University Press, 719–36.](http://paperpile.com/b/yikrbz/ZQnKK)

1. For the Lucid sample, the pseudo-random number generator is the built-in randomizer in Qualtrics’ online survey software. For the TESS sample, we utilized the randomization protocol for NORC’s AmeriSpeak panel. [↑](#footnote-ref-0)
2. By behavioral trust, we are referring to the amount of weight respondents give different sources when making their own forecasts [(Logg 2016; Logg, Minson, and Moore 2019)](https://paperpile.com/c/yikrbz/tMf6Z%2B4yv3N). [↑](#footnote-ref-1)
3. This wording for describing the algorithm is taken from [(Kennedy, Waggoner, and Ward 2022)](https://paperpile.com/c/yikrbz/zsXoH). Some previous commentators have expressed concerns that respondents may not understand what “algorithms” are. However, several studies have asked respondents to describe what they think of when they hear about an “algorithm.” These studies have generally found that respondents understand what an algorithm is quite well [(Logg 2016; Logg, Minson, and Moore 2019)](https://paperpile.com/c/yikrbz/4yv3N%2BtMf6Z). The wording is also chosen with specific reference to how both ethicists and politicians usually describe their decision aids. [↑](#footnote-ref-2)
4. <https://osf.io/f32p6/?view_only=4c4ce9959d784dadaa5211a03ce16575> [↑](#footnote-ref-3)
5. <https://osf.io/nz74r/?view_only=e3457d1558f74eaa82be0c0f1e17defd> [↑](#footnote-ref-4)