Harvard Journal of Law & Technology
Volume 34, Number 1 Fall 2020

ACCURACY IS NOT ENOUGH: THE TASK MISMATCH EXPLANATION OF ALGORITHM AVERSION AND ITS POLICY IMPLICATIONS

*Ethan Lowens**

TABLE OF CONTENTS

I. INTRODUCTION

Humans are poor and inconsistent forecasters. We have limited memory and processing power and are deceived by cognitive defects and artifacts. It is not surprising that algorithms often do better.[1] What

* Harvard Law School, Candidate for J.D., 2021. Many thanks to Professor Cass Sunstein, who advised and encouraged on this project from start to end. I am also very grateful to Dr. Arevik Avedian for supporting my empirical analysis and to Sam Friedlander for her valuable notes and input.

1. *See infra* Part II.

is puzzling is that people prefer to rely on human forecasters even when they are given overwhelming evidence that an algorithm would be more accurate. That phenomenon has inspired a wave of research on the drivers of "algorithm aversion."

Existing research dismisses algorithm aversion as the irrational consequence of cognitive biases. It appends algorithm aversion to the ever-expanding list of documented human cognitive defects like the tendency to value an object more when it is in one's possession than when it is not,[2] or for local news viewers to believe that crime is more prevalent in their neighborhoods than it really is.[3] This view suggests that popular outcry against an algorithm deserves little, if any, deference.

I argue that the story is not so simple. Prior studies examined algorithm aversion in a situation designed so that a human and an algorithm perform exactly the same task. In reality, such a situation is rare, if it exists at all. Aversion to an algorithm replacing humans in the real world may result from an intuition or observation that the algorithm lacks important capabilities. The algorithm's shortcoming may be technical, failing to account for key variables or malfunctioning under certain conditions. Alternatively, it may be metaphysical: virtually every task performed by a human involves some element of discretion or human touch that an algorithm cannot emulate.

At the core of this article is an empirical study which finds that a perceived mismatch between the task performed by a human and the capability of an algorithm poised to replace her drives respondents' aversion to the algorithm. I call this the "task mismatch" explanation of algorithm aversion.

It follows from the results of this study that policymakers should not systematically dismiss algorithm aversion as irrational. Popular outcry against an algorithm, motivated by a perceived task mismatch, may signal that adopting the algorithm would have unintended consequences. This signal is especially valuable where policymakers do not have personal experience in the context they are regulating — for instance, navigating the immigration system or enrolling for state-sponsored nutrition, healthcare, or housing benefits. Then, such a popular response may be the *only* way to detect a task mismatch.

---

2. Daniel Kahneman, Jack L. Knetsch & Richard H. Thaler, *Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias*, 5 J. ECON. PERSP. 193, 194–97 (1991).

3. *See* F. James Davis, *Crime News in Colorado Newspapers*, 57 AM. J. SOCIO. 325, 330 (1952); Travis L. Dixon, *Crime News and Racialized Beliefs: Understanding the Relationship Between Local News Viewing and Perceptions of African Americans and Crime*, 58 J. COMM. 106, 108 (2008).

The paper proceeds as follows: In Part II, I review past research on algorithm aversion. In Part III, I introduce the task mismatch explanation of algorithm aversion. In Part IV, I report results from an empirical study. In the study, participants learn about an algorithm that predicts whether a criminal defendant will fail to appear for trial with far greater accuracy than human judges. They are then asked to decide whether judges or the algorithm should decide if criminal defendants should be released before trial. The study shows that a considerable portion of people who expressed algorithm aversion were wary of a task mismatch between judges and their potential algorithmic replacement. In Part V, I discuss two implications of my findings. First, the task mismatch explanation for algorithm aversion documented in this paper, along with past research, creates a roadmap for policymakers to interpret and respond to algorithm aversion. Second, advocates and detractors of algorithms can leverage task-mismatch-driven algorithm aversion to influence popular opinion toward an algorithm.

II. EXISTING EXPLANATIONS OF ALGORITHM AVERSION: THE INACCURACY EXPLANATION AND THE CONFUSION EXPLANATION

Algorithms outperform human forecasters in myriad contexts. A meta-analysis of 136 studies between 1944 and 1994 found that, with only eight exceptions, algorithmic forecasters were as accurate, or more accurate, than human forecasters.[4] Since then, computing power and machine learning have improved, increasing algorithms' sophistication and accuracy. Moreover, algorithmic forecasters have additional advantages over humans: They are often more economical[5] and exceedingly consistent.[6]

---

4. W. M. Grove et al., *Clinical Versus Mechanical Prediction: A Meta-Analysis*, 12 PSYCH. ASSESSMENT 19, 25 (2000) (finding that "the general superiority (or at least material equivalence) of mechanical prediction . . . holds in general medicine, in mental health, in personality, and in education and training settings"); *see also* Scott Highhouse, *Stubborn Reliance on Intuition and Subjectivity in Employee Selection*, 1 INDUS. & ORG. PSYCH. 333, 334 (2008) (discussing the superiority of algorithmic tools such as written tests over unstructured interviews at predicting a job candidate's future performance).

5. *See, e.g.*, Adnan Tufail et al., *Automated Diabetic Retinopathy Image Assessment Software: Diagnostic Accuracy and Cost-Effectiveness Compared with Human Graders*, 124 OPHTHALMOLOGY 343, 348–49 (2017); John Lightbourne, *Algorithms & Fiduciaries: Existing and Proposed Regulatory Approaches to Artificially Intelligent Financial Planners*, 67 DUKE L.J. 651, 676 (2017); Erik Brynjolfsson, Yu Hu & Duncan Simester, *Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sales*, 57 MGMT. SCI. 1373, 1373–76 (2011).

6. Daniel Kahneman, Andrew Rosenfield, Linnea Gandhi & Tom Blaser, *Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making*, HARV. BUS. REV., Oct. 2016, https://hbr.org/2016/10/noise [https://perma.cc/E8XH-QT3B].

Yet, given the choice, people often prefer to rely on human forecasters.[7] This observation gave birth to the term "algorithm aversion" and an academic quest to determine its underlying causes. Recent empirical studies have focused on scenarios where participants may choose to rely on an algorithmic or human forecaster to perform exactly the same task: making a prediction about a future event.[8] Participants then receive evidence that the algorithm is a more accurate forecaster, and yet, to their detriment, most opt against relying on the algorithm.

One explanation for this phenomenon is that people wrongly perceive algorithms as less accurate, in spite of evidence that they are more accurate (the "inaccuracy explanation"). Dietvorst et al. demonstrate a mechanism behind this explanation: People display greater intolerance for error from algorithms than from humans.[9] If people see an algorithm make mistakes, they dismiss the algorithm as flawed.[10] When they see a human err, they are willing to give it another chance, believing he or she will learn.[11] In the Dietvorst et al. study, participants chose whether to rely on a human forecaster (either themselves or an anonymous third party) or an algorithm to predict the academic performance of MBA students based on their admissions files.[12] They received cash compensation for accurate predictions.[13] After seeing the algorithm perform (and make some mistakes), the vast majority (74%) of participants chose to rely on a human forecaster.[14] They did so in spite of the fact that they also observed that the algorithm was, on the whole, more accurate than the human forecaster.[15] Their tactic was costly — for most participants, relying on the algorithm would have resulted in considerably higher payments.[16]

In a subsequent study, Yeomans et al. identified a different driver of algorithm aversion: People's distrust for algorithms may stem from a lack of understanding of how they work (the "confusion

---

7. *See, e.g.*, Dalia L. Diab et al., *Lay Perceptions of Selection Decision Aids in US and Non-US Samples*, 19 INT'L J. SELECTION & ASSESSMENT 209, 209 (2011); Joseph Eastwood et al., *What People Want from Their Professionals: Attitudes Toward Decision-Making Strategies*, 25 J. BEHAV. DECISION MAKING 458, 458 (2012).

8. *See generally* Berkeley J. Dietvorst et al., *Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err*, 144 J. EXPERIMENTAL PSYCH. 114 (2015); Michael Yeomans et al., *Making Sense of Recommendations*, 32 J. BEHAV. DECISION MAKING 403 (2019).

9. Dietvorst et al., *supra* note 8, at 119.

10. *Id.* at 124.

11. *Id.*

12. *Id.* at 115.

13. *Id.* at 117.

14. *Id.* at 120.

15. *Id.* at 119–20.

16. *Id.*

explanation").[17] In Phase 1 of the Yeomans et al. study, participants predicted how funny a counterpart (the "target") would find a set of jokes.[18] Before the participant made her predictions, the target had previously rated a list of twelve jokes.[19] The participant had the opportunity to calibrate her predictions by seeing the target's ratings of four of these jokes.[20] Then, the participant predicted the target's ratings on the other eight jokes.[21] Perhaps surprisingly, an algorithm's predictions were consistently more accurate than the human participants', *even when* the target and participant were close friends or relatives.[22] Subsequently, participants were told that they could rely on help from the exceedingly accurate algorithm and to their detriment, many refused the offer.[23] In a variation of the study, people accepted help at higher rates after they read an explanation of how the algorithm worked.[24]

The Dietvorst et al. and Yeomans et al. studies examine controlled scenarios where a human and an algorithm perform the exact same task — make a prediction — and participants have concrete evidence that the algorithm is the superior predictor. Choosing to rely on a human forecaster under these circumstances is irrational: It clearly conflicts with participants' interest in making the best predictions. And yet, most did so anyway. These studies provide convincing evidence that cognitive defects, presented as the inaccuracy explanation and the confusion explanation, contribute to irrational algorithm aversion in carefully controlled, laboratory settings.

III. THE TASK MISMATCH EXPLANATION OF ALGORITHM AVERSION

However, it would be a mistake to generalize from Dietvorst et al.'s and Yeomans et al.'s findings that *all* algorithm aversion is irrational or the product of cognitive defects. Unlike the contrived situations in these studies, in many, if not most instances where algorithms are poised to replace human actors, the humans are not merely prediction machines.

People may spurn an algorithm when they perceive that it does not perform the same function as the human it is poised to replace; in

17. Yeomans et al., *supra* note 8, at 403.
18. *Id.* at 404.
19. *Id.* at 405.
20. *Id.*
21. *Id.*
22. *Id.* at 405, 411.
23. *Id.* at 408.
24. *Id.* at 411–12.

other words, where they perceive a task mismatch. Consider the following hypothetical: Michelle must choose between an algorithm and a human to complete *Task X*. *Task X* is traditionally performed by humans. Predicting *Y* is a necessary component of *Task X*. Michelle knows that the algorithm is extremely accurate at predicting *Y* — considerably more accurate than any person. She also completely understands how the algorithm works. If Michelle perceives that there is more to *Task X* than predicting *Y*, it could be reasonable for her to pick the human over the accurate, but misplaced, algorithm.

We can construct a concrete illustration of a task mismatch by drawing on terminology in the Yeomans et al. article. The paper is titled "Making Sense of Recommendations."[25] The authors refer to the participants in their study as "recommenders" and the targets as recipients of "recommendations."[26] In their conclusion, they report that algorithms outperform human "recommenders."[27] However, participants in that study were never asked to *recommend* a joke. Rather, they were asked to predict how funny someone rated jokes on a scale from -10 to 10.[28] In the parlance of the preceding paragraph's hypothetical, recommending a joke is *Task X* and guessing how funny someone rates a joke is predicting *Y*. Participants were asked only to predict *Y*. In spite of the article's misleading statements to the contrary, the study *does not* examine recommendations (*Task X*).

Giving recommendations is a complex social practice for which Yeomans et al.'s funniness scale is a coarse proxy. A joke recommendation, especially among close friends, factors in more than just the biggest laugh (for example, a reference to a shared past experience or a subtle romantic advance), just as a restaurant recommendation factors in more than food quality (price, location, taste preferences, creating opportunity to compare experiences afterward, etc.). If Michelle's *Task X* were making joke recommendations, and she believes that a joke recommendation is more than just a funniness prediction (*Y*), it would be perfectly reasonable for her to pick a human over Yeomans et al.'s algorithm for the task. That is in spite of the fact that — indeed it is *because* of the fact that — she knew exactly how Yeomans et al.'s algorithm worked.

The rest of this essay explores the possibility that a meaningful amount of algorithm aversion is driven by perceived task mismatches, rather than cognitive defects. This observation has important implications for legislators and policymakers.

---

25. *Id.* at 403.
26. *Id.*
27. *Id.* at 411.
28. *Id.* at 405.

IV. EMPIRICAL STUDY ON THE TASK MISMATCH EXPLANATION OF ALGORITHM AVERSION

A. Background on Pre-Trial Detention Decisions[29]

In this Section, I focus on the decision to incarcerate or release criminal defendants before trial. Traditionally, a judge has decided whether to (a) require that a defendant be incarcerated prior to his or her trial date, (b) release the defendant unconditionally, or (c) release the defendant conditional on providing collateral (cash or the title to property), which would be returned if the defendant returns for trial ("money bail").

In a remarkable study, Kleinberg et al. demonstrated that an algorithm far outperforms New York City's judges at predicting whether a given defendant, if released before trial, would fail to appear for his or her court date.[30] Given a subset of the data available to the judge (past criminal history, current offense, age, and prior skipped court appearances), the algorithm was asked to make a binary prediction: will the defendant fail to appear for trial?[31] At the time of the study under New York City law, judges could consider only failure to appear ("FTA") when deciding whether to allow pre-trial

29. This article should not be understood to endorse the use of pre-trial detention, money bail, or risk assessment algorithms for deciding pre-trial release. To the contrary, there is ample evidence that pre-trial detention causes permanent and unnecessary harm to defendants, their communities, and families. *See, e.g.*, Paul Heaton, Sandra Mayson & Megan Stevenson, *The Downstream Consequences of Misdemeanor Pretrial Detention*, 69 STAN L. REV. 711, 711 (2017); Will Dobbie, Jacob Goldin & Crystal S. Yang, *The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges*, 108 AM. ECON. REV. 201, 201 (2018). Moreover, evidence shows that money bail is an ineffective means of ensuring that people return to court, Cynthia E. Jones, *Accused and Unconvicted: Fleeing from Wealth-Based Pretrial Detention*, 82 ALB. L. REV. 1063, 1092 (2018), and is arguably unconstitutional because of the disproportionate burden it places on low-income defendants, *see* Christine S. Scott-Hayward & Sarah Ottone, *Punishing Poverty: California's Unconstitutional Bail System*, 70 STAN. L. REV. ONLINE 167, 168 (2018). Furthermore, risk assessment tools often incorporate racism and classism through facially neutral inputs such as prior arrests. *See* Sarah Picard et al., *Beyond the Algorithm: Pretrial Reform, Risk Assessment, and Racial Fairness*, CTR. FOR CT. INNOVATION, July 2019, at 8, https://www.courtinnovation.org/sites/default/files/media/document/2019/Beyond_The_Algorithm.pdf [https://perma.cc/U45H-C4H5]. In sum, I use the example of a failure-to-appear prediction algorithm not because I support its use, but because it illustrates how algorithm aversion can play a central role in critical contemporary public policy decisions.

30. *See* Jon Kleinberg et al., *Human Decisions and Machine Predictions*, 133 Q. J. ECON. 237, 237 (2018).

31. A shortcoming of the study is that the algorithm is asked only whether or not to release the defendant, not how much, if any, money the defendant would need to post as collateral. Accordingly, the study's authors must assume that either (a) the bail amount has no influence on a defendant's decision to return to court or (b) the algorithm would set the same bail amounts as the judges did. *See id.* at 245.

release.[32] Accordingly, the judges, like the algorithm, were, on their face, making an FTA prediction. The algorithm's predictions were much more accurate than the judges who ruled in these actual cases.[33]

Kleinberg et al. calculated that, using their algorithm instead of judges, New York City could reduce pre-trial detention rates by as much as 41.9% with no increase in crime or FTA.[34] An obvious takeaway would be that these algorithms should immediately replace the judges who currently determine pre-trial release. But Kleinberg et al. are careful to avoid that suggestion. They refer to their algorithm as a "decision aid," and note explicitly, when reporting their most dramatic findings, "In practice algorithms would be decision aids, not decision-makers. Our calculations simply highlight the scope of the potential gains."[35] Why would Kleinberg et al. preemptively curb the implications of their findings? Perhaps the authors themselves are algorithm averse. Alternatively, they may anticipate that the public is algorithm averse, and accordingly, that replacing human judges with their FTA-prediction algorithm would be politically untenable.

In this context, algorithm aversion could be driven by the inaccuracy explanation. In spite of evidence to the contrary, people might believe the judges' predictions were more accurate, especially if they observed the algorithm make errors. Or, it could be the product of the confusion explanation: People might spurn the algorithm, not understanding how it works.

I propose that the task mismatch explanation of algorithm aversion plays a significant role. People perceive that a judge's role deciding pre-trial detention (*Task X*) is fundamentally distinct from a flight risk prediction (*Y*), no matter how accurate the prediction.

B. Survey Design[36]

I test this hypothesis empirically using participants from Amazon's Mechanical Turk ("MTurk"). In Part 1 of the survey, participants read a description of a pre-trial release hearing and its two

32. MARY T. PHILLIPS, A DECADE OF BAIL RESEARCH IN NEW YORK CITY 13 (2012), https://www.prisonpolicy.org/scans/DecadeBailResearch12.pdf [https://perma.cc/DWV7-TXD6]. It is worth noting, however, that although the letter of the law restricted judges' pre-trial release decisions to the defendant's likelihood of failing to appear, there is certainly doubt around whether they truly cabined their analysis to this factor. Kleinberg et al., *supra* note 30, at 241, 241 n.5, 243. If judges surreptitiously considered other factors, then their decisions would not be FTA predictions comparable to the algorithm's. There would be a task mismatch between the judges' decisions and the algorithm's predictions. *See infra* note 40.

33. Kleinberg et al., *supra* note 30, at 240–41.

34. *Id.*

35. *Id.* at 241 n.5.

36. The complete survey is reproduced in the Appendix.

possible outcomes (detention or release pending trial). Then, they answer whether, in their opinion, a judge or an algorithm should determine pre-trial release. In Part 2, participants read about an algorithm modeled after the one in Kleinberg et al.'s study. The blurb notes that this algorithm is better than judges at predicting flight risk. After reading about the algorithm, participants again answer whether, in their opinion, a judge or an algorithm should determine pre-trial release. In Part 3, participants estimate, in their opinion, how accurate judges and the algorithm are at predicting flight risk and identify the factors that they think should underlie the decision to release a defendant before trial.

Part 1 captures the participants' "baseline" preference for judges or algorithms in the context of pre-trial release. Part 2 begins to identify the source of participants' algorithm aversion. Participants who still prefer judges at this stage, after learning that the algorithm is more accurate than judges at predicting FTA, and understanding how it works, may perceive a task mismatch. Part 3 probes for further indications of the task mismatch explanation of algorithm aversion.

C. Hypotheses

(H1) General algorithm aversion: A substantial portion of participants at baseline will prefer that judges, rather than algorithms, make pre-trial release decisions.

(H2) Exposure to data and the confusion explanation: Of those who preferred judges at baseline, many will switch to preferring algorithms upon reading how an FTA-prediction algorithm works and that it is more accurate than judges at predicting flight risk.

(H3) The task mismatch explanation: Some participants will still prefer judges after learning that an algorithm more accurately predicts FTA, and how it works. The task mismatch explanation may explain this group's algorithm aversion, as evidenced by the following:

- (a) They indicate that a factor other than predicting flight risk should be the primary basis for determining pre-trial release, *and/or*
- (b) They acknowledge that algorithms are equally or more accurate than judges at predicting flight risk.

D. Limitations and Assumptions

First and most importantly, the survey design is underinclusive. Participants read a description of the algorithm indicating that it predicts flight risk, but it is up to the participant to realize that flight risk might not be the only, or the appropriate, criteria for determining pre-trial release. I expect that some participants will wrongly infer, on the basis of the algorithm's description, that flight risk is the only criteria they may consider. Second, participants may exhibit demand characteristics, inferring from the positive description of the algorithm that the researcher would like them to regard the algorithm favorably.[37] Third, some participants took the survey very quickly. A delay timer forced them to spend five seconds and eight seconds, respectively, on the pages defining pre-trial release and describing the algorithm. While that is enough time for a fast reader to read the words on the page, it is likely not sufficient for serious deliberation: 17% of participants spent less than ten seconds on both these key pages, and 18% spent less than 2 seconds answering one of the study's key questions. Fourth, in order to avoid the influence of status quo bias,[38] I do not indicate whether judges or algorithms currently decide pre-trial release. However, participants likely intuit that judges are currently in charge and this may bias their decision-making.

E. Results

A total of 220 participants completed the study.[39] Participants were largely white (82%), young (37% were 19 to 30 years old, 90% were below 50 years old), and educated (93% attended at least some college). Men were slightly overrepresented (58%), as were Democrats (45%, vs. 21% Independents and 34% Republicans).

Consistent with H1, 69% of participants at baseline indicated that they would prefer that judges, rather than algorithms, determine pre-trial release.

Consistent with H2, 56% of those who said they would prefer judges at baseline (38% of the total sample) switched to preferring the algorithm after they learned how the algorithm worked and that it was more accurate than judges at predicting flight risk.

---

37. Martin T. Orne, *Demand Characteristics and the Concept of Quasi-Controls*, *in* ARTIFACTS IN BEHAVIORAL RESEARCH: ROBERT ROSENTHAL AND RALPH L. ROSNOW'S CLASSIC BOOKS 110 (4th ed. 2009).

38. *See* Kahneman et al., *supra* note 2, at 198–200.

39. A total of 250 participants took the survey. However, thirty failed the attention check, answering that "Food" or "The Stock Market" were the main focus of the survey.

Consistent with H3, 44% of those who said they would prefer judges at baseline (30% of the total sample, or 67 participants) continued to prefer judges even after they learned how the algorithm worked and that it was more accurate than judges at predicting flight risk. Let us refer to these participants as "algorithm averse participants," or "AAPs." A substantial majority of the 67 AAPs, 84%, indicated that a factor other than flight risk should be the primary basis for determining pre-trial release. Of the 67 AAPs, 30% (9% of the total sample) demonstrated the clearest expression of task-mismatch-driven algorithm aversion: They both (a) indicated that a factor other than flight risk should be the primary basis for determining pre-trial release and (b) acknowledged that the algorithm was *more* accurate at predicting flight risk. An additional 19% of the 67 AAPs (6% of the total sample) favored judges while indicating that (a) a factor other than flight risk should be the primary basis for determining pre-trial release and (b) the algorithm and judges were *equally* accurate at predicting flight risk.

F. Analysis of Results

The results of this study help explain a previously unexplained segment of algorithm-averse people. Of participants who at baseline stated that judges should decide pre-trial release, nearly half, 44%, continued to prefer judges even after they were told how an FTA-prediction algorithm worked and that the algorithm was more accurate than judges who decided pre-trial release. This suggests that they were not driven by the confusion explanation. Nor does the inaccuracy explanation explain their algorithm aversion: Over half of these participants (57%) indicated that they believed the algorithm was equally or more accurate than judges at predicting FTA. The task mismatch explanation may help fill these cracks left by prior research. Among participants who continued to prefer that judges decide pre-trial release after learning about the algorithmic alternative, the vast majority (84%) indicated that decisions about pre-trial release should be based primarily on a factor *other than* flight risk.[40] In the end, 30%

40. The factor chosen most often was "the likelihood that the defendant will commit another crime before their trial if they are released" (27% of AAPs). I assume that participants were unaware that in a small minority of jurisdictions, pre-trial release decisions must be based exclusively on flight risk (if participants *were* aware of that nuance, they may have been expressing dissatisfaction with the law, not task-mismatch-driven algorithm aversion). Several factors support that assumption. First, only five states do not permit judges to consider public safety concerns in deciding pre-trial release. Phillips, *supra* note 32, at 25. Second, participants are probably not aware of the subtleties of state criminal procedure. Carissa Byrne Hessick, *The Myth of Common Law Crimes*, 105 VA. L. REV. 965, 997–1001 (2019). Third, participants may intuit that, in spite of the letter of the law, in jurisdictions where they are admonished to consider only FTA risk, judges nonetheless

of AAPs gave an especially clear indication that a perceived task mismatch drove their preference: They both acknowledged that the algorithm was more accurate than judges at predicting flight risk and chose a factor other than risk of failure to appear as the most important for determining pre-trial release.

The influence of the task mismatch may be even more pervasive, but because of the survey's design, that influence fails to permeate to the key outcome variables. Every participant indicated that a factor other than flight risk should inform pre-trial release decisions in some way.[41] Many participants may have been uncomfortable with the algorithm's limitations but not so uncomfortable as to spurn it in favor of a judge who is worse at predicting FTA. Though these participants would not have registered as algorithm averse in this study's analysis, they did express some skepticism of the algorithm based on a perceived task mismatch.

The task mismatch explanation is, in theory, fully consistent with other explanations discussed in this paper.[42] These forces may operate in concert to contribute to algorithm aversion. Further research is needed to understand whether and how the different explanations for algorithm aversion interact.

V. POLICY IMPLICATIONS

These findings generate two layers of policy implications. First, this paper may inform how policymakers interpret algorithm aversion.

---

consider other factors such as public safety risk. *See, e.g.*, Jack F. Williams, *Process and Prediction: A Return to a Fuzzy Model of Pretrial Detention*, 79 MINN. L. REV. 325, 332 (1994) (describing how, in a bail system that did not allow judges to consider dangerousness, judges "surreptitiously forc[ed] defendants" they perceived as dangerous into pre-trial detention by setting "unlawfully high bail"); Samuel R. Wiseman, *Fixing Bail*, 84 GEO. WASH. L. REV. 417, 422 (2016) (explaining the incentives for judges to scrutinize dangerousness that result in excessive pre-trial detention).

41. The factors available to choose from were "promoting justice," "promoting public safety," "the likelihood that the defendant will fail to appear for their trial if they are released," "the likelihood that the defendant will commit another crime before their trial if they are released," "whether the defendant deserves to be released before trial," and "punishing the defendant." These options were displayed in a random order. The algorithm used by Kleinberg et al. could be calibrated to predict the combined likelihood that a defendant would commit a crime or fail to appear. Kleinberg et al., *supra* note 30, at 240. If they were aware of that, participants who thought that the likelihood of committing a crime before trial was an important factor for determining pre-trial release might have viewed the algorithm more favorably. Nonetheless, many participants still indicated that factors the algorithm ignored should inform the flight risk decision: 88% of all participants indicated that a factor other than flight risk or likelihood of committing a crime or public safety should inform the pre-trial release decision, and 37% indicated that a factor other than flight risk or the likelihood of committing a crime or public safety should be the main consideration.

42. *See supra* Part II.

A strong reaction against a plan to substitute a human with an algorithm might signal that the algorithm is an imperfect replacement. In response, policymakers could undertake to determine whether the algorithm was missing key features, and if so, to redesign the algorithm or develop means to reintegrate elements that were lost by switching to the algorithm. Second, an understanding of task-mismatch-driven algorithm aversion could benefit both pro- and anti-algorithm advocates. By highlighting the ways in which a human actor and a prospective algorithmic replacement perform the same task, algorithm advocates may alleviate task-mismatch-driven algorithm aversion. By contrast, algorithm opponents could emphasize differences between the algorithm and a human actor.

A. Public Policy Responses to Algorithm Aversion

How should policymakers interpret and respond to algorithm aversion? The findings in this paper provide a roadmap. Popular opposition to replacing a human with an algorithm is a signal to analyze that algorithm's capabilities against the role it will assume. Where the capability and role are identical and the algorithm is a superior performer, algorithm aversion may well be the consequence of irrational cognitive bias. This situation reflects the experimentally produced conditions in the Dietvorst et al. and Yeomans et al. studies.

Alternatively, algorithm aversion may illuminate a previously overlooked task mismatch. It is difficult, even impossible, to think of a situation where an algorithm exactly replaces a human actor. These shortcomings fall into two broad categories. First, there are capabilities that a human actor has and that an algorithm lacks but could conceivably incorporate. For example, in the context of the pre-trial release study described above, many participants indicated that pre-trial release decisions should consider the defendant's likelihood to commit another crime if they were released pending trial (46%).[43] Kleinberg et al. created a variant of their algorithm that predicts re-arrest pending trial in addition to FTA.[44] Adopting this more comprehensive version of the algorithm may alleviate some task-mismatch-driven algorithm aversion.

Often, however, human actors incorporate a "human touch" in their work that an algorithm is constitutionally incapable of replicating. A human can demonstrate sympathy and empathy,

43. Echoing the comment above, *supra* note 29, there are severe ethical and constitutional concerns around using predictions of re-arrest to inform pre-trial release decisions, whether made by a judge or by an algorithm. The discussion in this paper should not be read as an endorsement of any particular law, policy, or practice.

44. Kleinberg et al., *supra* note 30, at 240.

articulate and analyze her thought processes, provide the reassurance of human presence, and make comforting physical contact. Where an algorithm's deficiency is its lack of human touch, tweaking the algorithm is simply not an option. If they are committed to adopting the algorithm, policymakers and system designers must reintegrate lost human elements elsewhere in the system. This is already done in some contexts. For example, doctors and nurses deliver the results of fully automated tests in person. Weather reporters deliver forecasts on television even though computer models generate the meteorological predictions they announce. Pilots (and increasingly, operators of autonomous cars) supervise the autopilot system and even have discretion to override it in case they perceive an unusual situation or malfunction.

This analysis is not merely an academic exercise. Algorithm advocates, including Dietvorst et al., imply that algorithm aversion deserves no deference: "[Algorithm aversion is] enormously problematic, as it is a barrier to adopting superior approaches to a wide range of important tasks."[45] In contrast, I argue that it would be irresponsible, even dangerous, for lawmakers and policymakers to systematically ignore popular outcries against algorithms. Such a reaction may highlight previously overlooked deficiencies in the algorithm. This is especially likely given that policymakers often do not have the same life experiences as their constituents. For example, many, if not most, elected legislators have no personal experience applying for or enrolling in SNAP benefits, public housing, Medicaid, or asylum. They may not appreciate the full effects of replacing a human with an algorithm in one of these contexts. Popular backlash to an algorithm may be the best or even the *only* way to signal an important overlooked task mismatch.

Ultimately, policymakers must weigh the value of what is lost by switching to an algorithm against the value of gains in forecast accuracy, speed, consistency, or cost savings. If the value of the loss is substantial relative to the gains, the best course of action may be to abandon or reconfigure the algorithm or the system in which it will operate.[46]

---

45. Dietvorst et al., *supra* note 8, at 124.

46. These system design elements could be subtle or vast. A healthcare provider may decide to continue delivering the results of an automated test through a doctor or nurse, even though they could be communicated automatically via email. Perhaps the provider could entirely reimagine how results get delivered to patients, employing "sympathy specialists" especially for this role. Insurance companies might employ communications specialists, rather than trained actuaries, to help customers feed information into a lengthy, complex risk profile algorithm. As described in the next Section, professional tennis crafted rules so that its automated "Hawk Eye" ball-tracking system is activated only in limited contexts.

B. Implications for Algorithm Advocates, Detractors, and Designers

The task mismatch explanation of algorithm aversion suggests that the way people view a task will shape their perception of whether an algorithm can perform it adequately. For example, does a radiologist (a) assist in patients' treatment or (b) predict the likelihood that spots on scans will be harmful? Does a college admissions officer (a) identify the most deserving candidates or (b) predict which candidates are most likely to graduate and get a job? People who view the task as (a) might experience task mismatch aversion to an algorithm that did only (b).[47] Consider how framing could influence the conversation about using a Kleinberg et al. style algorithm to decide pre-trial release: In a jurisdiction like New York City, algorithm advocates could emphasize the letter of the law, which states that flight risk is the only basis for determining pre-trial release. Framed that way, a flight risk algorithm would be a perfect substitute for a judge, potentially quashing task-mismatch-driven aversion. Meanwhile, algorithm opponents could emphasize the more abstract functions of a judge: her role interpreting the law, giving a face to the justice system, and administering justice. A flight risk prediction algorithm would perform none of these functions.

The role an algorithm plays within a system can affect popular perception of the system as a whole. Designers and policymakers considering replacing a human decision-maker with an algorithm would be wise to carve out a role for the algorithm precisely tailored to its capability. For example, in 2004, professional tennis began using "Hawk Eye," an algorithmic tool that determined whether a ball was "in" or "out" in real time.[48] Hawk Eye could have replaced the human line judges altogether. However, that might have generated task-mismatch-driven backlash. The human judges do more than just make in/out calls; they help pace the game, bring energy when they make calls, and are part of an aesthetic and tradition that fans find comforting. Instead of replacing line judges outright, Hawk Eye activated only to decide close calls upon request by one of the players — its role was crafted narrowly to align with its capability. In a similar vein, Kleinberg et al. envision their algorithm as a "decision aid" to judges making pre-trial release decisions.[49] The algorithm's task would be defined narrowly and in a manner consistent with its

47. People may use abstract terms to define a task done by a human because it is done by a human. Similarly, they may view an algorithm's task as concrete because an algorithm is doing it. Accordingly, we might expect a strong status quo bias in how people view the abstractness of a decision-maker's task.

48. Cindy Shmerler, *Tennis Moves Toward Taking the Human Element Out of Line Calls*, N.Y. TIMES (Mar. 1, 2018), https://nyti.ms/2t7BswX [https://perma.cc/W7YL-JDNY].

49. Kleinberg et al., *supra* note 30, at 241 n.5.

capability — predicting flight risk — and thus reduce the risk of a task mismatch.

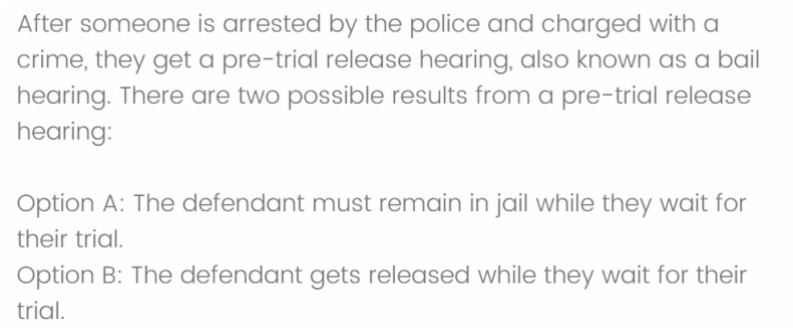
VI. CONCLUSION

It is tempting to dismiss algorithm aversion as a Luddite response or to disregard it as an artifact of human cognitive defects. After all, modern technology has successfully automated countless tasks that were previously performed by humans. Nonetheless, it is important to pay attention to algorithm aversion. Task-mismatch-driven algorithm aversion is a valuable, and maybe even the only, way to detect the loss of uniquely human elements when an algorithm replaces a human actor. These elements may be extremely meaningful, even if not measurable: a nurse's touch, a judge's sympathy, a referee's zeal. Algorithm aversion, handled properly, is not a barrier to technological progress but a tool to ensure that with the adoption of algorithms, humanity does not slip by, lost and unnoticed.

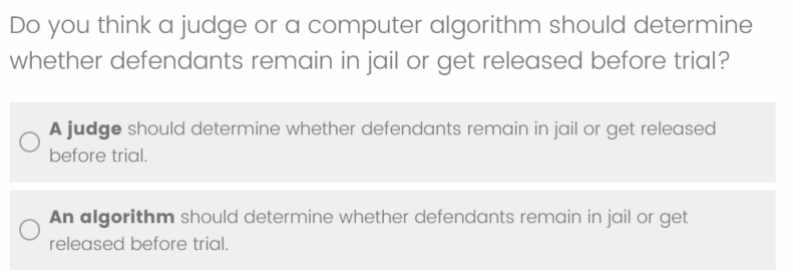
APPENDIX

Survey Questions

Part 1, Page 1



Part 1, Page 2



Part 2, Page 1

Here's some more information on a computer algorithm that predicts whether a defendant, if they are released, will fail to appear for their trial.

The algorithm predicts whether the defendant is likely to fail to appear based on their similarity to past defendants. Similarity is determined by:
-The crime the defendant is accused of committing
-The defendant's criminal history
-Whether the defendant has failed to appear in the past
-The defendant's age

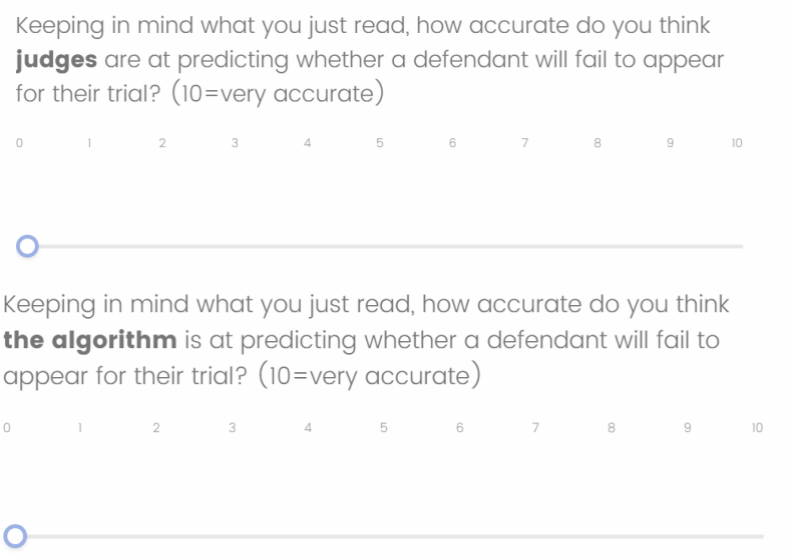
Research shows that this computer algorithm is better than judges at predicting whether a defendant will fail to appear for their trial.

Part 2, Page 2

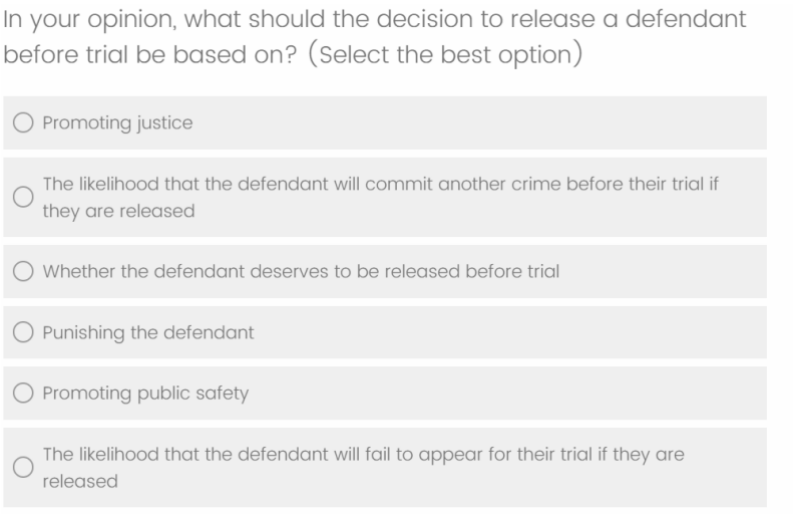
Keeping in mind what you just read, do you think a judge or the computer algorithm should determine whether defendants remain in jail or get released before trial?

- ○ **A judge** should determine whether defendants remain in jail or get released before trial.
- ○ **The algorithm** should determine whether defendants remain in jail or get released before trial.

Part 3, Page 1



Part 3, Page 2



Part 3, Page 3

In your opinion, what should the decision to release a defendant before trial be based on? (This time, select all that you think apply)

- ☐ Promoting justice
- ☐ Promoting public safety
- ☐ Punishing the defendant
- ☐ The likelihood that the defendant will commit another crime before their trial if they are released
- ☐ The likelihood that the defendant will fail to appear for their trial if they are released
- ☐ Whether the defendant deserves to be released before trial

Attention Check

Previously, we told you that: "Research shows that this computer algorithm is better than judges at predicting whether a defendant will fail to appear for their trial."

Do you believe that the computer algorithm is more accurate?

- ◯ Yes
- ◯ No