Algorithm Aversion

Berkeley Jay Dietvorst

University of Pennsylvania, bdietvo@gmail.com

Follow this and additional works at: http://repository.upenn.edu/edissertations

Part of the Advertising and Promotion Management Commons, Business Administration, Management, and Operations Commons, Management Sciences and Quantitative Methods Commons, Marketing Commons, and the Psychology Commons

Recommended Citation


This paper is posted at ScholarlyCommons. http://repository.upenn.edu/edissertations/1686
For more information, please contact repository@pobox.upenn.edu.
Algorithm Aversion

Abstract
Research shows that evidence-based algorithms more accurately predict the future than do human forecasters. Yet when forecasters are deciding whether to use a human forecaster or a statistical algorithm, they often choose the human forecaster. This phenomenon, which we call algorithm aversion, is costly, and it is important to understand its causes. In Chapter 1, we show that people are especially averse to algorithmic forecasters after seeing them perform, even when they see them outperform a human forecaster. This is because people more quickly lose confidence in algorithmic than human forecasters after seeing them make the same mistake. In Chapter 2, we investigate how aversion to imperfect algorithms might be overcome. We find that people are considerably more likely to choose to use an imperfect algorithm, and thus perform better, when they can modify its forecasts. Importantly, this is true even when they are severely restricted in the modifications they can make. Moreover, we find that people's decision to use a modifiable algorithm is relatively insensitive to the magnitude of the modifications they are able to make. Finally, we find that giving people the freedom to modify an imperfect algorithm makes them feel more satisfied with the forecasting process, more likely to believe that the algorithm is superior, and more likely to choose to use an algorithm to make subsequent forecasts.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Operations & Information Management

First Advisor
Joseph P. Simmons

Keywords
Confidence, Decision aids, Decision making, Forecasting, Heuristics and biases, Prediction

Subject Categories
Advertising and Promotion Management | Business Administration, Management, and Operations | Management Sciences and Quantitative Methods | Marketing | Psychology

This dissertation is available at ScholarlyCommons: http://repository.upenn.edu/edissertations/1686
ALGORITHM AVERSION

Berkeley J. Dietvorst

A DISSERTATION

in

Operations, Information and Decisions

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2016

Supervisor of Dissertation

_____________________
Joseph P. Simmons
Associate Professor of Operations, Information and Decisions

Graduate Group Chairperson

_____________________
Eric Bradlow, K.P. Chao Professor; Professor of Marketing, Statistics, and Education

Dissertation Committee:

Alex Rees-Jones, Assistant Professor of Operations, Information and Decisions
Cade Massey, Practice Professor of Operations, Information and Decisions
Deborah A. Small, Laura and John J. Pomerantz Professor of Marketing
ALGORITHM AVERSION

COPYRIGHT

2016

Berkeley Jay Dietvorst

This work is licensed under the
Creative Commons Attribution-
NonCommercial-ShareAlike 3.0
License

To view a copy of this license, visit

https://creativecommons.org/licenses/by-nc-sa/3.0/us/
ACKNOWLEDGMENTS

I am indebted to my coauthors Joseph P. Simmons and Cade Massey for their thoughtful work on this research. I thank my dissertation committee members Alex Rees-Jones and Deborah Small for their time and feedback. I am grateful to Uri Simonsohn and the members of The Wharton Decision Processes Lab for their helpful feedback throughout the research process. I thank the Operations, Information and Decisions department, the Wharton Behavioral Laboratory, and the Wharton Risk Center Ackoff Doctoral Student Fellowship for financial assistance. Finally, I would like to thank my spouse, Laura Nitowski, and my parents, Virginia Berkeley and Tom Dietvorst, for their unwavering support.
ABSTRACT

ALGORITHM AVERSION

Berkeley J. Dietvorst
Joseph P. Simmons

Research shows that evidence-based algorithms more accurately predict the future than do human forecasters. Yet when forecasters are deciding whether to use a human forecaster or a statistical algorithm, they often choose the human forecaster. This phenomenon, which we call algorithm aversion, is costly, and it is important to understand its causes. In Chapter 1, we show that people are especially averse to algorithmic forecasters after seeing them perform, even when they see them outperform a human forecaster. This is because people more quickly lose confidence in algorithmic than human forecasters after seeing them make the same mistake. In Chapter 2, we investigate how aversion to imperfect algorithms might be overcome. We find that people are considerably more likely to choose to use an imperfect algorithm, and thus perform better, when they can modify its forecasts. Importantly, this is true even when they are severely restricted in the modifications they can make. Moreover, we find that people’s decision to use a modifiable algorithm is relatively insensitive to the magnitude of the modifications they are able to make. Finally, we find that giving people the freedom to modify an imperfect algorithm makes them feel more satisfied with the forecasting process, more likely to believe that the algorithm is superior, and more likely to choose to use an algorithm to make subsequent forecasts.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>.......................... III</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>................................ IV</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>................................ VI</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>................................ VII</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>........................................ 1</td>
</tr>
<tr>
<td>CHAPTER 1</td>
<td>........................................ 4</td>
</tr>
<tr>
<td>Abstract</td>
<td>.......................................................... 5</td>
</tr>
<tr>
<td>Introduction</td>
<td>......................................................... 6</td>
</tr>
<tr>
<td>A Cause of Algorithm Aversion</td>
<td>................................................ 8</td>
</tr>
<tr>
<td>Overview of Studies</td>
<td>..................................................... 10</td>
</tr>
<tr>
<td>Methods</td>
<td>.......................................................... 11</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>.................................................. 21</td>
</tr>
<tr>
<td>General Discussion</td>
<td>..................................................... 33</td>
</tr>
<tr>
<td>CHAPTER 2</td>
<td>........................................ 37</td>
</tr>
<tr>
<td>Abstract</td>
<td>.......................................................... 38</td>
</tr>
<tr>
<td>Introduction</td>
<td>......................................................... 39</td>
</tr>
<tr>
<td>Overcoming Algorithm Aversion</td>
<td>................................................ 41</td>
</tr>
<tr>
<td>Study 1</td>
<td>.......................................................... 43</td>
</tr>
<tr>
<td>Study 2</td>
<td>.......................................................... 52</td>
</tr>
<tr>
<td>Study 3</td>
<td>.......................................................... 58</td>
</tr>
<tr>
<td>General Discussion</td>
<td>..................................................... 72</td>
</tr>
<tr>
<td>APPENDICIES</td>
<td>........................................ 77</td>
</tr>
<tr>
<td>BIBLIOGRAPHY</td>
<td>............................................. 78</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Tables in Chapter 1

Table 1. Studies 1, 2, and 4: Belief and Confidence Measures............................................................. 16
Table 2. Studies 3a and 3b: Belief and Confidence Measures................................................................. 19
Table 3. Studies 1–4: Forecasting Performance of Model vs. Human. ....................................................... 22
Table 4. Confidence in Model’s and Human’s Forecasts: Means (and Standard Deviations)...... 27
Table 5. Estimates of Model’s and Human’s Performance: Means (and Standard Deviations). ... 31
Table 6. Beliefs about the Model and Human Forecaster: Means (and Standard Deviations). ..... 32
Table 7. Participants’ Perceptions of the Model vs. Human Forecaster on Specific Attributes: Means (and Standard Deviations). ................................................................. 33
LIST OF FIGURES

Figures in Chapter 1

Figure 1. Example of forecasting task stimuli presented in Studies 1, 2, and 4......................... 13
Figure 2. Example of forecasting task stimuli presented in Studies 3a and 3b......................... 18
Figure 3. Studies 1-4: Participants who saw the statistical model’s results were less likely to choose it........................................................................................................................................... 24
Figure 4. Most People Do Not Choose The Statistical Model Unless They Are More Confident In The Model’s Forecasts Than In The Human’s Forecasts......................................................... 29

Figures in Chapter 2

Figure 5. Example of task stimuli used in all studies................................................................. 45
Figure 6. Study 1: Participants who could restrictively modify the model’s forecasts were more likely to choose to use the model, and performed better as a result........................................... 50
Figure 7. Study 1: The distribution of participants’ average absolute errors by condition and whether or not they chose to use the model’s forecasts. ............................................................. 51
Figure 8. Study 2: Participants who could restrictively modify the model’s forecasts were more likely to choose to use the model, and performed better as a result.......................................... 56
Figure 9. Study 2: The distribution of participants’ average absolute errors by condition and whether or not they chose to use the model’s forecasts. ......................................................... 57
Figure 10. Study 3’s Procedure..................................................................................................... 62
Figure 11. Study 3: Participants who could modify the model’s forecasts were more satisfied.. 65
Figure 12. Study 3: Participants in each condition chose to modify the model’s forecasts most often, instead of only using their own forecasts or only using the model’s forecasts. .......... 67
Figure 13. Study 3: Participants who could modify the model’s forecasts were more confident in the model’s forecasts and thought that the model performed better relative to themselves... 68
Figure 14. Study 3: Participants who had the option to adjust the model restrictively in the second stage of forecasts performed better and earned more money........................................... 69
INTRODUCTION

An abundance of research has shown that algorithms (e.g. statistical models, decision rules, actuarial tables) produce more accurate predictions than human experts. However, consumers, managers, and organizations often elect to use human forecasters instead of algorithms. This *algorithm aversion* is costly, as it leads to suboptimal outcomes in important domains including: forecasting demand for products, making hiring decisions, making financial investments, choosing which products to consume, making medical diagnoses, et cetera.

Algorithm aversion is an important and timely problem to address because advances in technology and the increased collection of data are constantly providing opportunities to capitalize on algorithmic prediction. However, very little research has investigated how people interact with algorithmic forecasters, and why people frequently fail to use algorithms for forecasting. In this research, we investigate when and why people fail to use algorithms to make predictions, and explore prescriptions that increase consumers’ and managers’ willingness to use algorithmic forecasters. Our objective in this research is to gain a better understanding of algorithm aversion, discover ways to increase consumers’ and managers’ use of algorithmic forecasters, and learn when people are most likely to avoid algorithms.

In the first essay, “Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err”, we investigate how people’s decision of whether to use an algorithm or a human forecaster is affected by their experience with those forecasters. In a series of
experiments, we demonstrate that people are substantially less likely to choose to use an algorithm in favor of a human forecaster after seeing the algorithm perform, even if they have also seen it outperform the human forecaster. This is because people lose considerably more confidence in algorithmic forecasters than human forecasters after seeing them make similar mistakes. People’s intolerance of imperfect algorithms in these experiments is very problematic; anyone who has experience with an algorithm will almost inevitably learn that it is imperfect because of the probabilistic nature of forecasting. Thus, people making consequential decisions may fail to use algorithms after learning that they are imperfect through experience.

In the second essay, “Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them.”, we explore prescriptions that increase the likelihood that people will choose to use an imperfect algorithm. We find that people are substantially more likely to use imperfect algorithms when they can use their own judgment to modify the algorithm’s forecasts. Interestingly, we also find people are relatively insensitive to the amount that they can modify an algorithm’s forecasts, which suggests that giving people even a little freedom to modify an algorithm’s forecasts will greatly increase their willingness to use the algorithm. This finding has important implications, as it is usually optimal to keep forecasts as close as possible to an algorithm’s forecasts. We also find that people who have modified an algorithm’s forecasts have more confidence in the algorithm’s forecasts and are more likely to defer to the algorithm than people who have used an algorithm’s forecasts without modifying them.
Overall, this dissertation makes five important contributions. First, Chapter 1 identifies a cause of algorithm aversion. We show that people lose confidence in algorithms after learning that they are imperfect, and fail to use them as a result. Second, Chapter 2 reports a prescription to reduce algorithm aversion. We find that people are willing to use an imperfect algorithm to make forecasts if they are able to slightly adjust that algorithm’s forecasts. Third, Chapter 2 shows that people are relatively insensitive to the magnitude by which they can adjust an algorithm’s forecasts when deciding whether or not to use it, and that people are not less satisfied modifying an algorithm by a limited amount versus an unlimited amount. Fourth, Chapter 2 demonstrates that people who are able to modify an imperfect algorithm’s forecasts are more satisfied with their forecasting process, and think that the algorithm performs better relative to themselves compared to people who are not able to modify the algorithm’s forecasts. Fifth, Chapter 2 shows that restricting the amount by which people can modify an algorithm’s forecasts leads them to deviate from the algorithm less and perform better as a result.
Algorithm Aversion:
People Erroneously Avoid Algorithms After Seeing Them Err

Berkeley J. Dietvorst       Joseph P. Simmons       Cade Massey

University of Pennsylvania
Abstract

Research shows that evidence-based algorithms more accurately predict the future than do human forecasters. Yet when forecasters are deciding whether to use a human forecaster or a statistical algorithm, they often choose the human forecaster. This phenomenon, which we call *algorithm aversion*, is costly, and it is important to understand its causes. We show that people are especially averse to algorithmic forecasters after seeing them perform, even when they see them outperform a human forecaster. This is because people more quickly lose confidence in algorithmic than human forecasters after seeing them make the same mistake. In five studies, participants either saw an algorithm make forecasts, a human make forecasts, both, or neither. They then decided whether to tie their incentives to the future predictions of the algorithm or the human. Participants who saw the algorithm perform were less confident in it, and less likely to choose it over an inferior human forecaster. This was true even among those who saw the algorithm outperform the human.

Keywords: Decision making, Decision aids, Heuristics and biases, Forecasting, Confidence
**Introduction**

Imagine that as an admissions officer for a university, it is your job to decide which student applicants to admit to your institution. Because your goal is to admit the applicants who will be most likely to succeed, this decision requires you to forecast students’ success using the information in their applications. There are at least two ways to make these forecasts. The more traditional way is for you to review each application yourself and make a forecast about each one. We refer to this as the *human method*. Alternatively, you could rely on an evidence-based algorithm\(^1\) to make these forecasts. For example, you might use the data of past students to construct a statistical model that provides a formula for combining each piece of information in the students’ applications. We refer to this as the *algorithm method*.

Research comparing the effectiveness of algorithmic and human forecasts shows that algorithms consistently outperform humans. In his 1954 book, *Clinical versus Statistical Prediction: A Theoretical Analysis and Review of the Evidence*, Paul Meehl reviewed results from 20 forecasting studies across diverse domains, including academic performance and parole violations, and showed that algorithms outperformed their human counterparts. Dawes subsequently gathered a large body of evidence showing that human experts did not perform as well as simple linear models at clinical diagnosis, forecasting graduate students’ success, and other prediction tasks (Dawes, 1979; Dawes, Faust, & Meehl, 1989). Following this work, Grove et al. (2000) meta-analyzed 136 studies investigating the prediction of human health and behavior. They found that algorithms

---

\(^1\) We use the term “algorithm” to encompass any evidence-based forecasting formula or rule. Thus, the term includes statistical models, decision rules, and all other mechanical procedures that can be used for forecasting.
outperformed human forecasters by 10% on average, and that it was far more common for algorithms to outperform human judges than the opposite. Thus, across the vast majority of forecasting tasks, algorithmic forecasts are more accurate than human forecasts (see also Silver, 2012).

If algorithms are better forecasters than humans, then people should choose algorithmic forecasts over human forecasts. However, they often don’t. In a wide variety of forecasting domains, experts and laypeople remain resistant to using algorithms, often opting to use forecasts made by an inferior human rather than forecasts made by a superior algorithm. Indeed, research shows that people often prefer humans’ forecasts to algorithms’ forecasts (Diab, Pui, Yankelvich, & Highhouse, 2011; Eastwood, Snook, & Luther, 2012), more strongly weigh human input than algorithmic input (Önkal et al., 2009; Promberger & Baron, 2006), and more harshly judge professionals who seek out advice from an algorithm rather than from a human (Shaffer et al., 2013).

This body of research indicates that people often exhibit what we refer to as algorithm aversion. However, it does not explain when people use human forecasters instead of superior algorithms, or why people fail to use algorithms for forecasting. In fact, we know very little about when and why people exhibit algorithm aversion.

Although scholars have written about this question, most of the writings are based on anecdotal experience rather than empirical evidence.² Some of the cited reasons for the cause of algorithm aversion include: the desire for perfect forecasts (Dawes, 1979; 

² One exception is the work of Arkes, Dawes, and Christensen (1986), who found that domain expertise diminished people’s reliance on algorithmic forecasts (and led to worse performance).
Einhorn, 1986; Highhouse, 2008), the inability of algorithms to learn (Dawes, 1979), the
presumed ability of human forecasters to improve through experience (Highhouse, 2008),
the notion that algorithms are dehumanizing (Dawes, 1979; Grove & Meehl, 1996), the
notion that algorithms cannot properly consider individual targets (Grove & Meehl,
1996), concerns about the ethicality of relying on algorithms to make important decisions
(Dawes, 1979), and the presumed inability of algorithms to incorporate qualitative data
(Grove & Meehl, 1996). On the one hand, these writings offer thoughtful and potentially
viable hypotheses about why algorithm aversion occurs. On the other hand, the absence
of empirical evidence means that we lack real insight into which of these (or other)
reasons actually drive algorithm aversion, and thus when people are most likely to exhibit
algorithm aversion. By identifying an important driver of algorithm aversion, our
research begins to provide this insight.

A Cause of Algorithm Aversion

Imagine that you are driving to work via your normal route. You run into traffic and
you predict that a different route will be faster. You get to work 20 minutes later than
usual, and you learn from a co-worker that your decision to abandon your route was
costly; the traffic was not as bad as it seemed. Many of us have made mistakes like this
one, and most would shrug it off. Very few people would decide to never again trust their
own judgment in such situations.

Now imagine the same scenario, but instead of you having wrongly decided to
abandon your route, your traffic-sensitive GPS made the error. Upon learning that the
GPS made a mistake, many of us would lose confidence in the machine, becoming
reluctant to use it again in a similar situation. It seems that the errors that we tolerate in humans become less tolerable when machines make them.

We believe that this example highlights a general tendency for people to more quickly lose confidence in algorithmic than human forecasters after seeing them make the same mistake. We propose that this tendency plays an important role in algorithm aversion. If this is true, then algorithm aversion should (partially) hinge on people’s experience with the algorithm. Although people may be willing to trust an algorithm in the absence of experience with it, seeing it perform – and almost inevitably err – will cause them to abandon it in favor of a human judge. This may occur even when people see the algorithm outperform the human.

We test this in five studies. In these studies, we asked participants to predict real outcomes from real data, and they had to decide whether to bet on the accuracy of human forecasts or the accuracy of forecasts made by a statistical model. We manipulated participants’ experience with the two forecasting methods prior to making this decision. In the control condition, they had no experience with either the human or the model. In the human condition, they saw the results of human forecasts but not model forecasts. In the model condition, they saw the results of model forecasts but not human forecasts. Finally, in the model-and-human condition, they saw the results of both the human and model forecasts.

Even though the model is superior to the humans – it outperforms the humans in all of the studies – experience reveals that it is not perfect and therefore makes mistakes. Because we expected people to lose confidence in the model after seeing it make mistakes, we expected them to choose the model much less often in the conditions
which they saw the model perform (the model and model-and-human conditions) than in those in which they did not (the control and human conditions). In sum, we predicted that people’s aversion to algorithms would be increased by seeing them perform (and therefore err), even when they saw the algorithms make less severe errors than a human forecaster.

Overview of Studies

In this article, we show that people’s use of an algorithmic vs. a human forecaster hinges on their experience with those two forecasters. In five studies, we demonstrate that seeing an algorithm perform (and therefore err) makes people less likely to use it instead of a human forecaster. We show that this occurs even for those who have seen the algorithm outperform the human, and regardless of whether the human forecaster is the participant herself or another, anonymous participant.

In all of our studies, participants were asked to use real data to forecast real outcomes. For example, in Studies 1, 2, and 4, participants were given MBA admissions data from past students and asked to predict how well the students had performed in the MBA program. Near the end of the experiment we asked them to choose which of two forecasting methods to rely on to make incentivized forecasts – a human judge (either themselves, in Studies 1-3, or another participant, in Study 4) or a statistical model that we built using the same data given to participants. Prior to making this decision, we manipulated whether participants witnessed the algorithm’s performance, the human’s performance, both, or neither.

Because the methods and results of these five studies are similar, we first describe the methods of all five studies and then reveal the results. For each study, we report how we
determined our sample size, all data exclusions (if any), all manipulations, and all measures. The exact materials and data are available in the Supplementary Online Materials here: https://osf.io/yc9ea/.

**Methods**

**Participants**

We conducted Studies 1, 2, and 4 in the Wharton School’s Behavioral Lab. Participants received a $10 show-up fee for an hour-long session of experiments, of which ours was a 20-minute component, and they could earn up to an additional $10 for accurate forecasting performance. In Study 1, we recruited as many participants as we could in two weeks; in Study 2 we recruited as many as we could in one week; and in Study 4, each participant was yoked to a different participant in Study 1, and so we decided to recruit exactly as many participants as had fully completed every question in Study 1. Eight, 4, and 0 participants from Studies 1, 2, and 4, respectively, exited the survey before completing the study’s key dependent measure, leaving us with final samples of 361, 206, and 354. These samples averaged 21-24 years of age and were 58-62% female.

We conducted Studies 3a and 3b using participants from amazon.com’s Mechanical Turk (MTurk) website. Participants received $1 for completing the study and they could earn up to an additional $1 for accurate forecasting performance. In Study 3a we decided in advance to recruit 400 participants (100 per condition) and in Study 3b we decided to recruit 1,000 participants (250 per condition). In both studies, participants who responded to the MTurk posting completed a question before they started the survey design to ensure that they were reading instructions. We programmed the survey to exclude any
participants who failed this check (77 in Study 3a and 217 in Study 3b), and some participants did not complete the key dependent measure (70 in Study 3a and 187 in Study 3b). This left us with final samples of 410 in Study 3a and 1,036 in Study 3b. These samples averaged 33-34 years of age and were 46-53% female.

**Procedures**

**Overview.** This section describes the procedures of each of the five studies, beginning with a detailed description of Study 1, and then briefer descriptions of the ways in which Studies 2-4 differed from Study 1. For ease of presentation, Tables 1, 2, and 7 list all measures we collected across the five studies.

**Study 1.** This experiment was administered as an online survey. After giving their consent and entering their Wharton Behavioral Lab ID number, participants were introduced to the experimental judgment task. Participants were told that they would play the part of an MBA admissions officer and that they would evaluate real MBA applicants using their application information. Specifically, they were told that it was their job to forecast the actual success of each applicant, where success was defined as an equal weighting of GPA, respect of fellow students (assessed via a survey), prestige of employer upon graduation (as measured in an annual poll of MBA students around the U.S.), and job success 2 years after graduation (measured by promotions and raises).

Participants were then told that the admissions office had created a statistical model that was designed to forecast student performance. They were told that the model was based on hundreds of past students, using the same data that the participants would
receive, and that the model was sophisticated, “put together by thoughtful analysts.”

Participants were further told that the model was designed to predict each applicant’s percentile among his/her classmates according to the success criteria described above, and a brief explanation of percentiles was provided to ensure that participants understood the prediction task. Finally, participants received detailed descriptions of the eight variables that they would receive about each applicant (undergraduate degree, GMAT scores, interview quality, essay quality, work experience, average salary, and parents’ education) before making their forecasts. Figure 1 shows an example of what participants saw when making their forecasts.

Figure 1. Example of forecasting task stimuli presented in Studies 1, 2, and 4.

<table>
<thead>
<tr>
<th>Undergraduate Degree</th>
<th>Business</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMAT - Verbal</td>
<td>43/60</td>
</tr>
<tr>
<td>GMAT - Quantitative</td>
<td>47/60</td>
</tr>
<tr>
<td>Essay Score</td>
<td>Good</td>
</tr>
<tr>
<td>Interview Score</td>
<td>Good</td>
</tr>
<tr>
<td>Work Experience (years)</td>
<td>5</td>
</tr>
<tr>
<td>Average Salary</td>
<td>$55,333</td>
</tr>
<tr>
<td>Average of Parents’ Education</td>
<td>Undergraduate degree(s)</td>
</tr>
</tbody>
</table>

The rest of the study proceeded in two stages. In the first stage, participants were randomly assigned to one of four conditions, which either gave them experience with the forecasting performance of the model (model condition), themselves (human condition), both the model and themselves (model-and-human condition), or neither (control condition). The three treatment conditions (human, model, model-and-human) were informed that they would next make (or see) 15 forecasts, and the control condition ($n = 3$).

---

3 The statistical model was built using the same data provided to participants and is described in the Supplement.
91) skipped this stage of the survey altogether. Participants in the model-and-human condition \( (n = 90) \) learned that, for each of the 15 applicants, they would make their own forecast, and then get feedback showing their own prediction, the model’s prediction, and the applicant’s true percentile. Participants in the human condition \( (n = 90) \) learned that they would make a forecast and then get feedback showing their own prediction and the applicant’s true percentile. Participants in the model condition \( (n = 90) \) learned that they would get feedback showing the model’s prediction and the applicant’s true percentile. After receiving these instructions, these participants proceeded through the 15 forecasts, receiving feedback after each one. They were not incentivized for accurately making these forecasts. The 15 forecasted applicants were randomly selected (without replacement) from a pool of 115 applicants, and thus varied across participants.

Next, in the second stage of the survey, all participants learned that they would make 10 “official” incentivized estimates, earning an extra $1 each time the forecast they used was within 5 percentiles of an MBA student’s realized percentile. To be sure they understood this instruction, participants were required to type the following sentence into a text box before proceeding: “You will receive a $1 bonus for each of your 10 estimates that is within 5 percentiles of a student’s true percentile. Therefore, you can earn an extra $0 to $10 depending on your performance.”

We then administered the study’s key dependent measure. Participants were told that they could choose to have either their own forecasts or the model’s forecasts determine their bonuses for the 10 incentivized estimates. They were then asked to choose between the two methods, by answering, “Would you like your estimates or the model’s estimates to determine your bonuses for all 10 rounds?” The two response options were “Use only
the statistical model’s estimates to determine my bonuses for all 10 rounds” and “Use only my estimates to determine my bonuses for all 10 rounds.” We made it very clear to participants that their choice of selecting either the model or themselves would apply to all 10 of the forecasts they were about to make.

After choosing between themselves and the algorithm, participants forecasted the success of 10 randomly chosen applicants (excluding those they were exposed to in the first stage, if any). All participants made a forecast and then saw the model’s forecast for 10 randomly selected MBA applicants. They received no feedback about their own or the model’s performance while completing these forecasts.

After making these forecasts, participants answered questions designed to assess their confidence in, and beliefs about, the model and themselves (see Table 1 for the list of questions). Finally, participants learned their bonus and reported their age, gender, and highest level of education.

---

4 For all five studies, after each stage 2 trial participants guessed if their estimate or the model’s was closer to the true value after seeing the model’s forecast. This measure was exploratory and we do not discuss it further.
Table 1. Studies 1, 2, and 4: Belief and Confidence Measures.

<table>
<thead>
<tr>
<th>Study 1</th>
<th>Study 2</th>
<th>Study 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>How much bonus money do you think you would earn if your own estimates determined your bonus? (0–10)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>How much bonus money do you think you would earn if the model’s estimates determined your bonus? (0–10)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>What percent of the time do you think the model’s estimates are within 5 percentiles of a student’s true score? (0–100)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>What percent of the time do you think your estimates are within 5 percentiles of a student’s true score? (0–100)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>On average, how many percentiles do you think the model’s estimates are away from students’ actual percentiles? (0–100)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>On average, how many percentiles do you think your estimates are away from students’ actual percentiles? (0–100)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>How much confidence do you have in the statistical model’s estimates? (1 = none; 5 = a lot)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>How much confidence do you have in your estimates? (1 = none; 5 = a lot)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>How well did the statistical model perform in comparison to your expectations? (1 = much worse; 5 = much better)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Why did you choose to have your bonus be determined by your [the statistical model’s] estimates instead of the statistical model’s [your] estimates? (open-ended)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>What are your thoughts and feelings about the statistical model? (open-ended)</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Note. All measures were collected after participants decided whether to tie their bonuses to the model or the human. Check marks with “a” subscripts indicate that the measure was collected after participants completed the stage 2 forecasts. Check marks with “b” subscripts indicate that the measure was collected before participants completed the stage 2 forecasts. Questions are listed in the order in which they were asked. In Study 4, all questions asking about “your estimates” instead asked about “the lab participant’s estimates.”

Study 2. In Study 2, we conducted a closer examination of our most interesting experimental condition – the “model-and-human” condition in which participants saw both the human and the model perform before deciding which forecasting method to bet on. We wanted to see the model-and-human condition’s tendency to tie their incentives to their own forecasts would replicate in a larger sample. We also wanted to see whether it would be robust to changes in the incentive structure, and to knowing during the first stage, when getting feedback on both the model’s and their own performance, what the incentive structure would be.

This study’s procedure was the same as Study 1’s except for five changes. First, all participants were assigned to the model-and-human condition. Second, participants were randomly assigned to one of three types of bonuses in the experiment’s second stage. Participants were either paid $1 each time their forecast was within 5 percentiles of an MBA student’s realized percentile (5-percentile condition; \( n = 70 \)), paid $1 each time their forecast was within 20 percentiles of an MBA student’s realized percentile (20-
percentile condition; \( n = 69 \), or paid based on their average absolute error (AAE condition; \( n = 67 \)). Participants who were paid based on average absolute error earned $10 if their average absolute error was less than or equal to four, and this bonus decreased by $1 for each four additional units of average error. This payment rule is reproduced in Appendix A.

Third, unlike in Study 1, participants learned this payment rule just before making the 15 unincentivized forecasts in the first stage. Thus, they were fully informed about the payment rule while encoding their own and the model’s performance during the first 15 trials. We implemented this design feature in Studies 3a and 3b as well.

Fourth, participants completed a few additional confidence and belief measures, some of which were asked immediately before they completed their stage 2 forecasts (see Table 1). Participants also answered an exploratory block of questions asking them to rate the relative competencies of the model and themselves on a number of specific attributes. This block of questions, which was also included in the remaining studies, is listed in Table 7.

**Study 3a.** Study 3a examined whether the results of Study 1 would replicate in a different forecasting domain, and when the model outperformed participants’ forecasts by a much wider margin.

As in Study 1, participants were randomly assigned to one of four conditions – model \( (n = 101) \), human \( (n = 105) \), model-and-human \( (n = 99) \), and control \( (n = 105) \) – which determined whether, in the first stage of the experiment, they saw the model’s forecasts, made their own forecasts, both, or neither.
Study 3a’s procedure was the same as Study 1 except for a few changes. Most notably, the forecasting task was different. Study 3a’s forecasting task involved predicting the rank (1 to 50) of individual U.S. states in terms of the number of airline passengers that departed from that state in 2011. A rank of 1 indicates that the state had the most departing airline passengers and a rank of 50 indicates that it had the least departing airline passengers.

To make each forecast, participants received the following pieces of information about the state: its number of major airports (as defined by the Bureau of Transportation), its 2010 census population rank (1 to 50), its total number of counties rank (1 to 50), its 2008 median household income rank (1 to 50), and its 2009 domestic travel expenditure rank (1 to 50). Figure 2 shows an example of the stimuli used in this study. All of the stimuli that participants saw during the experiment were randomly selected without replacement from a pool of the 50 U.S. states. The statistical model was built using airline passenger data from 2006-2010 and the same variables provided to participants, and is described in more detail in the Supplement.

Figure 2. Example of forecasting task stimuli presented in Studies 3a and 3b.

<table>
<thead>
<tr>
<th>Number of Major Airports</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Census Population Rank - 2010</td>
<td>9</td>
</tr>
<tr>
<td>Number of Counties Rank</td>
<td>2</td>
</tr>
<tr>
<td>Median Household Income Rank - 2008</td>
<td>23</td>
</tr>
<tr>
<td>Domestic Travel Expenditure Rank - 2009</td>
<td>9</td>
</tr>
</tbody>
</table>

There were five other procedural differences between Study 3a and Study 1. First, participants who were not in the control condition completed 10 unincentivized forecasts
instead of 15 in the first stage of the experiment. Second, in the second stage of the study, all participants completed one incentivized forecast instead of 10. Thus, their decision about whether to bet on the model’s forecast or their own pertained to the judgment of a single state.

Third, we used a different payment rule to determine participants’ bonuses for that forecast. Participants were paid $1 if they made a perfect forecast. This bonus decreased by $0.15 for each additional unit of error associated with their estimate. This payment rule is reproduced in Appendix B. Fourth, as in Study 2, participants learned this payment rule before starting the first stage of unincentivized forecasts instead of after that stage. Finally, as shown in Tables 2 and 7, the measures that we asked participants to complete were slightly different.

**Study 3b.** Study 3b was a higher-powered direct replication of Study 3a.⁵ Except for some differences in the measures that we collected, and in the timing of those measures (see Table 2), the procedures of Studies 3a and 3b were identical.

**Table 2.** Studies 3a and 3b: Belief and Confidence Measures.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Study 3a</th>
<th>Study 3b</th>
</tr>
</thead>
<tbody>
<tr>
<td>On average, how many ranks do you think the model’s estimates are away from states’ actual ranks? (0–50)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>On average, how many ranks do you think your estimates are away from states’ actual ranks? (0–50)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>How much confidence do you have in the statistical model’s estimates? (1 = none; 5 = a lot)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>How much confidence do you have in your estimates? (1 = none; 5 = a lot)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>How likely is it that the model will predict a state’s rank almost perfectly? (1 = certainly not true; 9 = certainly true)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>How likely is it that you will predict a state’s rank almost perfectly? (1 = certainly not true; 9 = certainly true)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>How many of the 50 states do you think the model would estimate perfectly? (0–50)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>How many of the 50 states do you think you would estimate perfectly? (0–50)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>How likely is the model to make a really bad estimate? (1 = extremely unlikely; 9 = extremely likely)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>How well did the statistical model perform in comparison to your expectations? (1 = much worse; 5 = much better)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Why did you choose to have your bonus be determined by your [the statistical model’s] estimates instead of the statistical model’s [your] estimates? (open-ended)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>What are your thoughts and feelings about the statistical model? (open-ended)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* All measures were collected after participants decided whether to tie their bonuses to the model or the human. Check marks with “a” subscripts indicate that the measure was collected after participants completed the stage 2 forecasts. Check marks with “b” subscripts indicate that the measure was collected before participants completed the stage 2 forecasts. Questions are listed in the order in which they were asked.

⁵ As described in the Supplement, Study 3b’s replication attempt was motivated by having observed some weaker results in similar studies run prior to Study 3a. This study ensured that Study 3a’s findings were not due to chance.
Study 4. The previous studies investigated whether people are more likely to use their *own* forecasts after seeing an algorithm perform. In Study 4, we investigated whether this effect extends to choices between an algorithm’s forecasts and the forecasts of a different person.

The procedure for this experiment was identical to Study 1’s, except that participants chose between a past participant’s forecasts and the model’s instead of between their own forecasts and the model’s. Each participant was yoked to a unique participant from Study 1, and thus assigned to the same condition as that participant: either control ($n = 88$), human ($n = 87$), model ($n = 90$), or model-and-human ($n = 89$). Study 4 participants saw exactly the same sequence of information that the matched participant had seen, including the exact same 15 forecasting outcomes in stage 1. For example, Study 4 participants who were matched with a Study 1 participant who was in the model-and-human condition saw that participant’s stage 1 forecasts and saw exactly the same model forecasts that that participant had seen. Following stage 1, all participants decided whether to tie their stage 2 forecasting bonuses to the model’s forecasts or to the forecasts of the Study 1 participant they were matched with.

As shown in Table 1, Study 4 participants completed the same measures asked in Study 1. In addition, as in Studies 2, 3a, and 3b, they also answered the block of questions asking them to compare the human forecaster to the model; though in this study the questions required a comparison between the model and the participant they were matched with, rather than a comparison between the model and themselves (see Table 7).
Results and Discussion

Forecasting Performance

As expected, the model outperformed participants in all five studies. As shown in Table 3, participants would have earned significantly larger bonuses if they had tied their bonuses to the statistical model’s forecasts than if they had tied their bonuses to the human’s forecasts. Moreover, the model’s forecasts were much more highly correlated with realized outcomes than were humans’ forecasts (r = .53 vs. r = .16 in the MBA student forecasting task; r = .92 vs. r = .69 in the airline passenger forecasting task). In terms of average absolute error, the human forecasters produced 15-29% more error than the model in the MBA student forecasting task of Studies 1, 2, and 4, and 90-97% more error than the model in the airline passenger forecasting task of Studies 3a and 3b (see Table 3). This was true in both the stage 1 and stage 2 forecasts. Thus, participants in the model-and-human condition, who saw both the model and the human perform in stage 1, were much more likely to see the model outperform the human than to see the opposite, and this was especially true in Studies 3a and 3b. Participants in every condition in every study were better off choosing the model over the human.
Table 3. Studies 1-4: Forecasting Performance of Model vs. Human.

<table>
<thead>
<tr>
<th></th>
<th>Model</th>
<th>Human</th>
<th>Difference</th>
<th>Paired t test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bonus if chose model vs. human</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 1</td>
<td>$1.78 (1.17)</td>
<td>$1.38 (1.04)</td>
<td>$0.40 (1.52)</td>
<td>(t(360) = 4.98, p &lt; .001)</td>
</tr>
<tr>
<td>Study 2</td>
<td>$1.77 (1.13)</td>
<td>$1.09 (0.99)</td>
<td>$0.68 (1.56)</td>
<td>(t(204) = 6.26, p &lt; .001)</td>
</tr>
<tr>
<td>Study 3a</td>
<td>$0.48 (0.37)</td>
<td>$0.31 (0.36)</td>
<td>$0.17 (0.45)</td>
<td>(t(405) = 7.73, p &lt; .001)</td>
</tr>
<tr>
<td>Study 3b</td>
<td>$0.49 (0.36)</td>
<td>$0.30 (0.34)</td>
<td>$0.20 (0.44)</td>
<td>(t(1,028) = 14.40, p &lt; .001)</td>
</tr>
<tr>
<td>Study 4</td>
<td>$1.79 (1.17)</td>
<td>$1.38 (1.05)</td>
<td>$0.41 (1.52)</td>
<td>(t(353) = 5.11, p &lt; .001)</td>
</tr>
</tbody>
</table>

AAE in model-and-human condition (Stage 1 unincentivized forecasts)

| Study 1       | 23.13 (4.39)| 26.67 (5.48)| −3.53 (6.08)| \(t(89) = −5.52, p < .001\) |
| Study 2       | 22.57 (4.08)| 29.12 (7.30)| −6.54 (7.60)| \(t(205) = −12.37, p < .001\) |
| Study 3a      | 4.28 (1.10)| 8.45 (3.52)| −4.17 (3.57)| \(t(98) = −11.63, p < .001\) |
| Study 3b      | 4.39 (1.19)| 8.32 (3.52)| −3.93 (3.64)| \(t(256) = −17.30, p < .001\) |
| Study 4       | 23.11 (4.41)| 26.68 (5.51)| −3.56 (6.10)| \(t(88) = −5.51, p < .001\) |

AAE (Stage 2 incentivized forecasts)

| Study 1       | 22.07 (4.98)| 26.61 (6.45)| −4.54 (7.50)| \(t(360) = −11.52, p < .001\) |
| Study 2       | 22.61 (5.10)| 28.64 (7.30)| −6.03 (7.50)| \(t(204) = −9.39, p < .001\) |
| Study 3a      | 4.54 (4.37)| 8.89 (8.99)| −4.35 (9.52)| \(t(405) = −9.21, p < .001\) |
| Study 3b      | 4.32 (4.23)| 8.34 (8.16)| −4.03 (8.36)| \(t(1,028) = −15.44, p < .001\) |
| Study 4       | 22.02 (4.98)| 26.64 (6.45)| −4.62 (7.44)| \(t(353) = −11.68, p < .001\) |

Note. AAE = average absolute error.

Main Analyses

We hypothesized that seeing the model perform, and therefore err, would decrease participants’ tendency to bet on it rather the human forecaster, despite the fact that the model was more accurate than the human. As shown in Figure 3, this effect was observed, and highly significant, in all four studies in which we manipulated experience with the model. In Study 1, we observed this effect in lab participants’ forecasts of MBA students’ performance. In Studies 3a and 3b, we learned that this effect generalizes to a different forecasting task – predicting states’ ranks in number of departing airline passengers – and, importantly, to a context in which the model dramatically outperforms the human forecasters, producing about half as much error in these two studies. Although some magnitude of advantage must lead participants who see the algorithm perform to be more likely to choose it – for example, if they were to see the algorithm predict all
outcomes exactly right – the model’s large advantage in these studies was not large enough to get them to do so. Finally, Study 4 teaches us that the effect extends to choices between the model and a different human judge. In sum, the results consistently support the hypothesis that seeing an algorithm perform makes people less likely to choose it.

Interestingly, participants in the model-and-human conditions, most of whom saw the model outperform the human in the first stage of the experiment (610 of 741 [83%] across the five studies), were, across all studies, among those least likely to choose the model. In every experiment, participants in the model-and-human condition were significantly less likely to tie their bonuses to the model than were participants who did not see the model perform. This result is not limited to the minority who saw the human outperform the model, as even those who saw the model outperform the human were less likely to choose the model than were participants who did not see the model perform. In addition, the results of Study 2, in which all participants were assigned to the model-and-human condition, teach us that the aversion to the model in this condition persists within a large sample, and is not contingent on the incentive structure.

Figure 3 also shows that although seeing the model perform, and therefore err, decreased the tendency to choose the model, seeing the human perform, and therefore err, did not significantly decrease the tendency to choose the human. This suggests, as hypothesized, that people are quicker to abandon algorithms that make mistakes than to

6 When we say that the model “outperformed” the human, we mean that, across the trials in the first stage of the experiment, the average absolute deviation between the model’s forecasts and the true percentiles was smaller than the average absolute deviation between the human’s forecasts and the true percentiles.
7 With all model-and-human condition participants included, the statistical tests are: Study 1, $\chi^2(1, N = 271) = 39.94, p < .001$; Study 3a, $\chi^2(1, N = 309) = 4.72, p = .030$; Study 3b, $\chi^2(1, N = 783) = 16.83, p < .001$; Study 4, $\chi^2(1, N = 264) = 13.84, p < .001$. Considering only the model-and-human condition participants who saw the model outperform the human during stage 1, the statistical tests are: Study 1, $\chi^2(1, N = 242) = 20.07, p < .001$; Study 3a, $\chi^2(1, N = 302) = 2.54, p = .111$; Study 3b, $\chi^2(1, N = 758) = 9.92, p = .002$; Study 4, $\chi^2(1, N = 235) = 5.24, p = .022$. 

23
abandon humans that make mistakes, even though, as is often the case, the humans’ mistakes were larger.

**Figure 3.** Studies 1-4: Participants who saw the statistical model’s results were less likely to choose it.

<table>
<thead>
<tr>
<th>% Choosing Statistical Model to Forecast MBA Students’ Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Graph" /></td>
</tr>
<tr>
<td><strong>Study 1</strong></td>
</tr>
<tr>
<td>Control</td>
</tr>
<tr>
<td>Saw Results of statistical model’s forecasts</td>
</tr>
<tr>
<td>Saw Results of human’s forecasts</td>
</tr>
<tr>
<td><strong>Study 2</strong></td>
</tr>
<tr>
<td>Control</td>
</tr>
<tr>
<td>Saw Results of statistical model’s forecasts</td>
</tr>
<tr>
<td>Saw Results of human’s forecasts</td>
</tr>
<tr>
<td><strong>Study 4</strong></td>
</tr>
<tr>
<td>Control</td>
</tr>
<tr>
<td>Saw Results of statistical model’s forecasts</td>
</tr>
<tr>
<td>Saw Results of human’s forecasts</td>
</tr>
</tbody>
</table>

Note. Errors bars indicate ±1 standard error. In Study 2, “AAE”, “5-Pct”, and “20-Pct” signify conditions in which participants were incentivized either for minimizing average absolute error, for getting within 5 percentiles of the correct answer, or for getting within 20 percentiles of the correct answer, respectively.

![Graph](image2)  
**Study 3a**  
<table>
<thead>
<tr>
<th>Control</th>
<th>Saw Results of statistical model’s forecasts</th>
<th>Saw Results of human’s forecasts</th>
</tr>
</thead>
<tbody>
<tr>
<td>57%</td>
<td>42%</td>
<td>43%</td>
</tr>
</tbody>
</table>

Note. Errors bars indicate ±1 standard error. In Study 2, “AAE”, “5-Pct”, and “20-Pct” signify conditions in which participants were incentivized either for minimizing average absolute error, for getting within 5 percentiles of the correct answer, or for getting within 20 percentiles of the correct answer, respectively.

Figure 3 reveals additional findings of interest. First, although all three of Study 2’s incentive conditions showed the hypothesized aversion to using the algorithm, the magnitude of this aversion did differ across conditions, $\chi^2(2, N = 206) = 8.50, p = .014$. Participants who were paid for providing forecasts within 20 percentiles of the correct
answer were less likely to choose the model than were participants who were paid for providing forecasts within 5 percentiles of the correct answer, $\chi^2(1, N = 139) = 3.56, p = .059$, as well as participants whose payment was based on the average absolute error, $\chi^2(1, N = 136) = 8.56, p = .003$. As correct predictions were easier to obtain in the 20-percentile condition, this effect likely reflects people’s greater relative confidence in their own forecasts when forecasts are easy than when they are difficult (e.g., Heath & Tversky, 1991; Kruger, 1999; Moore & Healy, 2008). In support of this claim, although participants’ confidence in the model’s forecasting ability did not differ between the 20-percentile ($M = 2.84, SD = 0.80$) and other payment conditions ($M = 2.73, SD = 0.85$), $t(203) = 0.92, p = .360$, they were significantly more confident in their own forecasting ability in the 20-percentile condition ($M = 3.20, SD = 0.85$) than in the other payment conditions ($M = 2.67, SD = 0.85$), $t(203) = 4.24, p < .001$. Moreover, follow-up analyses revealed that the effect of the 20-percentile incentive on preference for the model was mediated by confidence in their own forecasts, but not by confidence in the model’s forecasts.

Additionally, although one must be cautious about making comparisons across experiments, Figure 3 also shows that, across conditions, participants were more likely to bet on the model against another participant (Study 4) than against themselves (Study 1).

---

8 The 5-percentile and AAE conditions did not differ, $\chi^2(1, N = 137) = 1.21, p = .271$.

9 We conducted a binary mediation analysis, where the dependent variable was choice of the model or human, the mediators were confidence in their own forecasts and confidence in the model’s forecasts, and the independent variable was whether or not participants were in the 20-percentile condition. We then used Preacher and Hayes’s bootstrapping procedure to obtain unbiased 95% confidence intervals around the mediated effects. Confidence in their own forecasts significantly mediated the effect of incentive condition on choice of the model ($-.057, -.190$), but confidence in the model’s forecasts did not ($-.036, .095$).
This suggests that algorithm aversion may be more pronounced among those whose forecasts the algorithm threatens to replace.

**Confidence**

Participants’ confidence ratings show an interesting pattern, one that suggests that participants “learned” more from the model’s mistakes than from the human’s (see Table 4). Whereas seeing the human perform did not consistently decrease confidence in the human’s forecasts – it did so significantly only in Study 4 – seeing the model perform significantly decreased participants’ confidence in the model’s forecasts in all four studies.\(^\text{10}\) Thus, seeing a model make relatively small mistakes consistently decreased confidence in the model, whereas seeing a human make relatively large mistakes did not consistently decrease confidence in the human.

We tested whether confidence in the model’s or human’s forecasts significantly mediated the effect of seeing the model perform on participants’ likelihood of choosing the model over the human. We conducted binary mediation analyses, setting choice of the model or the human as the dependent variable (0 = chose to tie their bonus to the human; 1 = chose to tie their bonus to the model), whether or not participants saw the model perform as the independent variable (0 = control or human condition; 1 = model or model-and-human condition), and confidence in the human’s forecasts and confidence in the model’s forecasts as mediators. We used Preacher and Hayes’s (2008) bootstrapping procedure to obtain unbiased 95% confidence intervals around the mediated effects. In all

\(^\text{10}\) Seeing the model perform significantly decreased confidence in the model’s forecasts in every study: Study 1, \(t(358) = 6.69, p < .001\); Study 3a, \(t(403) = 2.19, p = .029\); Study 3b, \(t(1,032) = 7.16, p < .001\); Study 4, \(t(351) = 5.12, p < .001\). Seeing the human perform significantly decreased confidence in the human’s forecasts in only one of the four studies: Study 1, \(t(358) = 1.12, p = .262\); Study 3a, \(t(403) = -0.06, p = .952\); Study 3b, \(t(1,031) = 0.756, p = .450\); Study 4, \(t(351) = 2.28, p = .023\).
in the model’s forecasts significantly mediated the effect whereas confidence in the human did not.\textsuperscript{11}

\textit{Table 4. Confidence in Model’s and Human’s Forecasts: Means (and Standard Deviations).}

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Human</th>
<th>Model</th>
<th>Model-and-human</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confidence in model’s forecasts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 1</td>
<td>3.04\textsubscript{a} (0.86)</td>
<td>3.17\textsubscript{a} (0.82)</td>
<td>2.49\textsubscript{b} (0.71)</td>
<td>2.63\textsubscript{b} (0.68)</td>
</tr>
<tr>
<td>Study 2</td>
<td>3.40\textsubscript{a} (0.83)</td>
<td>3.57\textsubscript{a} (0.73)</td>
<td>3.34\textsubscript{b} (0.79)</td>
<td>3.29\textsubscript{a} (0.79)</td>
</tr>
<tr>
<td>Study 3a</td>
<td>3.75\textsubscript{a} (0.75)</td>
<td>3.61\textsubscript{a} (0.76)</td>
<td>3.34\textsubscript{b} (0.74)</td>
<td>3.36\textsubscript{b} (0.69)</td>
</tr>
<tr>
<td>Study 3b</td>
<td>3.30\textsubscript{a} (0.80)</td>
<td>3.28\textsubscript{a} (0.75)</td>
<td>2.86\textsubscript{b} (0.73)</td>
<td>2.87\textsubscript{a} (0.86)</td>
</tr>
</tbody>
</table>

| Confidence in human’s forecasts |         |         |         |                |
| Study 1                     | 2.70\textsubscript{a} (0.80) | 2.47\textsubscript{a} (0.69) | 2.60\textsubscript{a} (0.75) | 2.66\textsubscript{a} (0.75) |
| Study 2                     | 2.85\textsubscript{a} (0.83) | 2.90\textsubscript{a} (0.95) | 3.07\textsubscript{a} (1.01) | 3.03\textsubscript{a} (0.90) |
| Study 3a                    | 2.92\textsubscript{a} (0.85) | 2.78\textsubscript{a} (0.78) | 2.83\textsubscript{a} (0.81) | 2.90\textsubscript{a} (0.80) |
| Study 3b                    | 3.11\textsubscript{a} (0.73) | 2.79\textsubscript{a} (0.69) | 3.01\textsubscript{ab} (0.73) | 2.97\textsubscript{ab} (0.83) |

\textit{Note.} Within each row, means with different subscripts differ at $p < .05$ using Tukey’s test.

It is interesting that reducing confidence in the model’s forecasts seems to have led participants to abandon it, because participants who saw the model perform were \textbf{not} more confident in the human’s forecasts than in the model’s. Whereas participants in the control and human conditions were more confident in the model’s forecasts than in the human’s, participants in the model and model-and-human conditions were about equally confident in the model’s and human’s forecasts (see Table 4). Yet, in our studies, they chose to tie their forecasts to the human most of the time.

Figure 4 explores this further, plotting the relationship between choosing the statistical model and differences in confidence in the model’s forecasts versus the human’s forecasts. There are a few things to note. First, passing the sanity check, people

\textsuperscript{11} For confidence in the model’s forecasts, the 95\% confidence intervals were: Study 1, (-.165, -.070); Study 3a, (-.071, -.004); Study 3b, (-.112, -.060); Study 4, (-.174, -.068). For confidence in the human’s forecasts, the 95\% confidence intervals were: Study 1, (-.029, .013); Study 3a, (-.073, .004); Study 3b, (-.033, .027); Study 4, (-.043, .026).
who were more confident in the model’s forecasts than in the human’s forecasts were more likely to tie their bonuses to the model’s forecasts, whereas people who were more confident in the human’s forecasts than in the model’s forecasts were more likely to tie their bonuses to the human’s forecasts. More interestingly, the majority of people who were equally confident in the model’s and human’s forecasts chose to tie their bonuses to the human’s forecasts, particularly when they had seen the model perform. It seems that most people will choose the statistical model over the human only when they are more confident in the model than in the human.

Finally, the divergent lines in Figure 4 show that the effect of seeing the model perform on participant’s choice of the model is not fully accounted for by differences in confidence. Participants who expressed less confidence in the model’s forecasts than in the human’s forecasts were, unsurprisingly, relatively unlikely to tie their bonuses to the model, but this was more pronounced for those saw the model perform. This difference may occur because expressions of confidence in the model’s forecasts are less meaningful without seeing the model perform, or because the confidence measure may fail to fully capture people’s disdain for a model that they see err. Whatever the cause, it is clear that seeing the model perform reduces the likelihood of choosing the model, over and above the effect it has on reducing confidence.
Figure 4. Most People Do Not Choose The Statistical Model Unless They Are More Confident In The Model’s Forecasts Than In The Human’s Forecasts.

% Choosing Statistical Model

Note. Errors bars indicate ±1 standard error. The “Did Not See Model Perform” line represents results from participants in the control and human conditions. The “Saw Model Perform” line represents results from participants in the model and model-and-human conditions. Differences in confidence between the model’s and human’s forecasts were computed by subtracting participants’ ratings of confidence in the human forecasts from their ratings of confidence in the model’s forecasts (i.e., by subtracting one 5-point scale from the other). From left to right, the five x-axis categories reflect difference scores of: <-1, -1, 0, +1, and >1. The figure includes results from all five studies.

Beliefs

In addition to measuring confidence in the model’s and human’s forecasts, we also measured beliefs about the model’s and human’s forecasts. As shown in Tables 5 and 6, the results of these belief measures are similar to the results of the confidence measures: With few exceptions, seeing the model perform made participants less optimistic about the model. For example, Study 1 participants who saw the model perform were significantly less likely to believe that the model would be within five percentiles of the right answer than were participants who did not see the model perform. And Study 3b
participants who saw the model perform thought it would make fewer perfect predictions than participants who did not see the model perform.

Table 6 reveals other interesting results. One alternative account for why people find algorithms so distasteful may rest on people’s desire for perfect predictions. Specifically, people may choose human over algorithmic forecasts because, although they expect algorithms to outperform humans on average, they expect a human forecast to have a greater chance of being perfect. However, the data in Table 6 fail to support this. In every condition — even those in which people were unlikely to choose the model — participants believed the algorithm to be more likely than the human to yield a perfect prediction. Moreover, in Study 3b, the effect of seeing the model err on the likelihood of betting on the model persisted even among those who thought the model was more perfect than themselves (63% vs. 55%), $\chi^2(1, N = 795) = 4.92, p = .027$. Thus, the algorithm aversion that arises from experience with the model seems not entirely driven by a belief that the model is less likely to be perfect. Rather, it seems driven more by people being more likely to learn that the model is bad when they see the model make (smaller) mistakes than they are to learn that the human is bad when they see the human make (larger) mistakes.

Finally, it is also interesting to consider the responses of participants in the control condition in Study 3b, who did not see either the model or themselves make forecasts before making their judgments. These participants expected a superhuman performance from the human – to perfectly predict 16.7 of 50 (33%) ranks – and a supermodel$^{12}$

$^{12}$ Sorry.
performance from the model – to perfectly predict 30.4 of 50 (61%) ranks. In reality, the humans and the model perfectly predicted 2.2 (4%) and 6.0 (12%) ranks, respectively. Although one may forgive this optimism in light of the control condition’s unfamiliarity with the task, those with experience, including those who saw both the model and the human perform, also expressed dramatically unrealistic expectations, predicting the model and human to perfectly forecast many more ranks than was possible (see Table 6). Even those with experience may expect forecasters to perform at an impossibly high level.

*Table 5. Estimates of Model’s and Human’s Performance: Means (and Standard Deviations).*

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Human</th>
<th>Model</th>
<th>Model-and-human</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated % of model’s</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estimates within 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>percentiles</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 1</td>
<td>46.52&lt;sub&gt;a&lt;/sub&gt; (22.48)</td>
<td>47.63&lt;sub&gt;a&lt;/sub&gt; (23.48)</td>
<td>28.24&lt;sub&gt;b&lt;/sub&gt; (18.78)</td>
<td>36.73&lt;sub&gt;b&lt;/sub&gt; (22.61)</td>
</tr>
<tr>
<td>Study 4</td>
<td>52.89&lt;sub&gt;a&lt;/sub&gt; (17.50)</td>
<td>50.64&lt;sub&gt;ab&lt;/sub&gt; (20.28)</td>
<td>37.51&lt;sub&gt;c&lt;/sub&gt; (19.88)</td>
<td>43.47&lt;sub&gt;c&lt;/sub&gt; (20.83)</td>
</tr>
<tr>
<td>Estimated % of human’s</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estimates within 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>percentiles</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 1</td>
<td>37.02&lt;sub&gt;a&lt;/sub&gt; (19.35)</td>
<td>27.19&lt;sub&gt;b&lt;/sub&gt; (18.84)</td>
<td>32.67&lt;sub&gt;ab&lt;/sub&gt; (21.25)</td>
<td>31.63&lt;sub&gt;ab&lt;/sub&gt; (19.90)</td>
</tr>
<tr>
<td>Study 4</td>
<td>45.22&lt;sub&gt;a&lt;/sub&gt; (18.76)</td>
<td>36.80&lt;sub&gt;b&lt;/sub&gt; (19.62)</td>
<td>40.63&lt;sub&gt;ab&lt;/sub&gt; (21.22)</td>
<td>40.12&lt;sub&gt;ab&lt;/sub&gt; (18.70)</td>
</tr>
<tr>
<td>Estimated average absolute</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>deviation of model’s</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 3a</td>
<td>7.51&lt;sub&gt;a&lt;/sub&gt; (8.19)</td>
<td>5.08&lt;sub&gt;a&lt;/sub&gt; (5.75)</td>
<td>6.18&lt;sub&gt;ab&lt;/sub&gt; (6.06)</td>
<td>6.13&lt;sub&gt;ab&lt;/sub&gt; (6.30)</td>
</tr>
<tr>
<td>Study 3b</td>
<td>5.09&lt;sub&gt;b&lt;/sub&gt; (6.84)</td>
<td>4.87&lt;sub&gt;b&lt;/sub&gt; (4.29)</td>
<td>5.75&lt;sub&gt;ab&lt;/sub&gt; (4.39)</td>
<td>6.53&lt;sub&gt;c&lt;/sub&gt; (5.43)</td>
</tr>
<tr>
<td>Estimated average absolute</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>deviation of human’s</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 3a</td>
<td>8.56&lt;sub&gt;a&lt;/sub&gt; (8.51)</td>
<td>7.44&lt;sub&gt;a&lt;/sub&gt; (7.51)</td>
<td>7.36&lt;sub&gt;a&lt;/sub&gt; (8.46)</td>
<td>7.39&lt;sub&gt;a&lt;/sub&gt; (6.87)</td>
</tr>
<tr>
<td>Study 3b</td>
<td>8.11&lt;sub&gt;a&lt;/sub&gt; (8.38)</td>
<td>8.73&lt;sub&gt;a&lt;/sub&gt; (7.40)</td>
<td>7.29&lt;sub&gt;a&lt;/sub&gt; (6.36)</td>
<td>8.28&lt;sub&gt;a&lt;/sub&gt; (6.71)</td>
</tr>
</tbody>
</table>

*Note. Within each row, means with different subscripts differ at p < .05 using Tukey’s test.*
Table 6. Beliefs about the Model and Human Forecaster: Means (and Standard Deviations).

<table>
<thead>
<tr>
<th>Test</th>
<th>Control</th>
<th>Human</th>
<th>Model</th>
<th>Model-and-human</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood the model will make a perfect prediction (9-point scale)</td>
<td>Study 3a</td>
<td>Study 3b</td>
<td>Study 3b</td>
<td>Study 3b</td>
</tr>
<tr>
<td></td>
<td>5.35 (1.61)</td>
<td>5.59 (1.50)</td>
<td>4.80 (1.71)</td>
<td>4.60 (1.57)</td>
</tr>
<tr>
<td></td>
<td>6.14 (1.54)</td>
<td>5.72 (1.59)</td>
<td>4.89 (1.55)</td>
<td>4.94 (1.61)</td>
</tr>
<tr>
<td>Likelihood the human will make a perfect prediction (9-point scale)</td>
<td>Study 3b</td>
<td>Study 3b</td>
<td>Study 3b</td>
<td>Study 3b</td>
</tr>
<tr>
<td></td>
<td>4.30 (1.84)</td>
<td>3.64 (1.62)</td>
<td>3.73 (1.61)</td>
<td>3.89 (1.63)</td>
</tr>
<tr>
<td>Number of states the model will predict perfectly (0–50)</td>
<td>Study 3b</td>
<td>Study 3b</td>
<td>Study 3b</td>
<td>Study 3b</td>
</tr>
<tr>
<td></td>
<td>30.36 (14.01)</td>
<td>25.16 (14.57)</td>
<td>15.20 (11.83)</td>
<td>15.84 (12.35)</td>
</tr>
<tr>
<td>Number of states the human will predict perfectly (0–50)</td>
<td>Study 3b</td>
<td>Study 3b</td>
<td>Study 3b</td>
<td>Study 3b</td>
</tr>
<tr>
<td></td>
<td>16.70 (13.14)</td>
<td>8.43 (8.37)</td>
<td>9.11 (9.16)</td>
<td>8.60 (9.08)</td>
</tr>
<tr>
<td>Likelihood the model will make a really bad estimate (9-point scale)</td>
<td>Study 3b</td>
<td>Study 3b</td>
<td>Study 3b</td>
<td>Study 3b</td>
</tr>
<tr>
<td></td>
<td>3.78 (1.55)</td>
<td>3.80 (1.44)</td>
<td>4.41 (1.52)</td>
<td>4.36 (1.47)</td>
</tr>
<tr>
<td>Performance of model relative to expectations (5-point scale)</td>
<td>Study 3a</td>
<td>Study 3a</td>
<td>Study 3a</td>
<td>Study 3a</td>
</tr>
<tr>
<td></td>
<td>3.12 (0.73)</td>
<td>3.32 (0.69)</td>
<td>2.99 (0.83)</td>
<td>3.11 (0.78)</td>
</tr>
</tbody>
</table>

Note. Within each row, means with different subscripts differ at p < .05 using Tukey’s test.

Comparing The Model and Human on Specific Attributes

For purely exploratory purposes, we asked participants in Studies 2-4 to rate how the human and the model compared at different aspects of forecasting. These scale items were inspired by observations made by Einhorn (1986), Dawes (1979), and Grove and Meehl (1996), who articulated the various ways in which humans and algorithms may be perceived to differ. Our aim was to measure these perceived differences, in the hope of understanding what advantages people believe humans to have over models (and vice versa), which could inform future attempts to reduce algorithm aversion.

Table 7 shows the results of these measures. Participants rightfully thought that the model was better than human forecasters at avoiding obvious mistakes, appropriately weighing various attributes, and consistently weighing information. Consistent with research on the adoption of decision aids (see Highhouse, 2008), participants thought that
the human forecasters were better than the model at getting better with practice, learning from mistakes, and finding underappreciated candidates. These data suggest that one may attempt to reduce algorithm aversion by either educating people of the importance of providing consistent and appropriate weights, or by convincing them that models can learn or that humans cannot. We look forward to future research that builds on these preliminary findings.

Table 7. Participants’ Perceptions of the Model vs. Human Forecaster on Specific Attributes: Means (and Standard Deviations).

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Study 2</th>
<th>Study 3a</th>
<th>Study 3b</th>
<th>Study 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detecting exceptions</td>
<td>3.55 (0.99)</td>
<td>3.02 (1.08)</td>
<td>2.98 (1.08)</td>
<td>3.91 (0.97)</td>
</tr>
<tr>
<td>Finding underappreciated candidates</td>
<td>3.74 (0.96)</td>
<td>3.74 (0.92)</td>
<td>3.67 (0.95)</td>
<td>3.81 (0.99)</td>
</tr>
<tr>
<td>Avoiding obvious mistakes</td>
<td>2.68 (1.10)</td>
<td>2.64 (1.03)</td>
<td>2.62 (1.02)</td>
<td>2.55 (1.13)</td>
</tr>
<tr>
<td>Learning from mistakes</td>
<td>5.91 (0.81)</td>
<td>5.91 (0.82)</td>
<td>5.91 (0.82)</td>
<td>5.91 (0.82)</td>
</tr>
<tr>
<td>Appropriately weighing a candidate’s qualities (task’s attributes)</td>
<td>2.98 (1.09)</td>
<td>2.50 (0.92)</td>
<td>2.34 (0.93)</td>
<td>2.81 (1.11)</td>
</tr>
<tr>
<td>Consistently weighing information</td>
<td>2.33 (1.10)</td>
<td>2.49 (1.00)</td>
<td>2.29 (0.98)</td>
<td>2.05 (1.02)</td>
</tr>
<tr>
<td>Treating each student (state) individually</td>
<td>3.60 (1.02)</td>
<td>2.94 (1.02)</td>
<td>2.89 (1.02)</td>
<td>3.48 (1.25)</td>
</tr>
<tr>
<td>Getting better with practice</td>
<td>3.85 (0.82)</td>
<td>3.66 (0.96)</td>
<td>3.63 (0.98)</td>
<td>3.77 (1.08)</td>
</tr>
</tbody>
</table>

Note. In Studies 2-3b, participants were asked to, “Please indicate how you and the model compare on the following attributes.” In Study 4, participants were asked to, “Please indicate how the lab participant and the model compare on the following attributes.” All answers were given on 5-point scales, where 1 = “Model is much better” and 5 = “I am [The participant is] much better”. Each mean significantly below the scale midpoint is denoted with an “m” subscript, indicating that the model is significantly better than the human; each mean significantly above the scale midpoint is denoted with an “h” subscript, indicating that the human is significantly better than the model.

General Discussion

The results of five studies show that seeing algorithms err makes people less confident in them and less likely to choose them over an inferior human forecaster. This effect was evident in two distinct domains of judgment, including one in which the human forecasters produced nearly twice as much error as the algorithm. It arose regardless of whether the participant was choosing between the algorithm and her own forecasts or between the algorithm and the forecasts of a different participant. And it even
arose among the (vast majority of) participants who saw the algorithm outperform the human forecaster.

The aversion to algorithms is costly, not only for the participants in our studies who lost money when they chose not to tie their bonuses to the algorithm, but for society at large. Many decisions require a forecast, and algorithms are almost always better forecasters than humans (Dawes, 1979; Meehl, 1954; Grove et al., 2000). The ubiquity of computers and the growth of the “big data” movement (Davenport & Harris, 2007) have encouraged the growth of algorithms, but many remain resistant to using them. Our studies show that this resistance at least partially arises from greater intolerance for error from algorithms than from humans. People are more likely to abandon an algorithm than a human judge for making the same mistake. This is enormously problematic, as it is a barrier to adopting superior approaches to a wide range of important tasks. It means, for example, that people will more likely forgive an admissions committee than an admissions algorithm for making an error, even when on average the algorithm makes fewer such errors. In short, whenever prediction errors are likely – as they are in virtually all forecasting tasks – people will be biased against algorithms.

More optimistically, our findings do suggest that people will be much more willing to use algorithms when they do not see algorithms err, as will be the case when errors are unseen, the algorithm is unseen – as it often is for patients in doctors’ offices – or when predictions are nearly perfect. The 2012 U.S. Presidential election season saw people embracing a perfectly performing algorithm. Nate Silver’s New York Times blog, Five Thirty Eight: Nate Silver’s Political Calculus, presented an algorithm for forecasting that election. Though the site had its critics before the votes were in – one Washington Post
writer criticized Silver for “doing little more than weighting and aggregating state polls and combining them with various historical assumptions to project a future outcome with exaggerated, attention-grabbing exactitude” (Gerson, 2012, para. 2) – those critics were soon silenced: Silver’s model correctly predicted the presidential election results in all 50 states. Live on MSNBC Rachel Maddow proclaimed, “You know who won the election tonight? Nate Silver,” (Noveck, 2012, para. 21) and headlines like, “Nate Silver gets a big boost from the election” (Isidore, 2012) and, “How Nate Silver Won the 2012 Presidential Election” (Clark, 2012) followed. Many journalists and popular bloggers declared Silver’s success a great boost for “big data” and statistical prediction (Honan, 2012; McDermott, 2012; Taylor, 2012; Tiku, 2012).

However, we worry that this is not such a generalizable victory. People may rally around an algorithm touted as perfect, but we doubt that this enthusiasm will generalize to algorithms that are shown to be less perfect, as they inevitably will be much of the time.

Limitations and Future Directions

Our studies leave some open questions. First, we did not explore all of the boundaries of our effect. For example, we found that participants were significantly more likely to use humans that produced 13-97% more error than algorithms after seeing those algorithms err. However, we do not know if this effect would persist if the algorithms in question were many times more accurate than the human forecasters. Presumably there is some level of performance advantage that algorithms could exhibit over humans that would lead forecasters to use the algorithms even after seeing them err. However, in practice, algorithms’ advantage over human forecasters is rarely larger than the advantage
they had in our studies (Grove et al., 2000), and so the question of whether our effects generalize to algorithms that have an even larger advantage may not be an urgent one to answer. Also, although we found this effect on two distinct forecasting tasks, it is possible that our effect is contingent on features that these tasks had in common.

Second, our studies did not explore the many ways in which algorithms may vary, and how those variations may affect algorithm aversion. For example, algorithms can differ in their complexity, the degree to which they are transparent to forecasters, the degree to which forecasters are involved in their construction, and the algorithm designer’s expertise, all of which may affect forecasters’ likelihood of using an algorithm. For example, it is likely that forecasters would be more willing to use algorithms built by experts than algorithms built by amateurs. Additionally, people may be more or less likely to use algorithms that are simple and transparent – more likely if they feel more comfortable with transparent algorithms, but less likely if that transparency makes it obvious that the algorithm will err. We look forward to future research investigating how algorithms’ attributes affect algorithm aversion.

Third, our results show that algorithm aversion is not entirely driven by seeing algorithms err. In the studies presented in this paper, nontrivial percentages of participants continued to use an algorithm after they had seen it err and failed to use an algorithm when they had not seen it err. This suggests that there are other important drivers of algorithm aversion that we have not uncovered. Finally, our research has little to say about how best to reduce algorithm aversion among those who have seen the algorithm err. This is the next (and great) challenge for future research.
CHAPTER 2

Overcoming Algorithm Aversion:

People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them

Berkeley J. Dietvorst                Joseph P. Simmons                Cade Massey

University of Pennsylvania
Abstract

Although evidence-based algorithms consistently outperform human forecasters, people consistently fail to use them after learning that they are imperfect. In this paper, we investigate how such algorithm aversion might be overcome. In incentivized forecasting tasks, we give people the choice between using their own forecasts or those of an algorithm that was built by experts. We find that people are considerably more likely to choose to use an imperfect algorithm, and thus perform better, when they can modify its forecasts. Importantly, this is true even when they are severely restricted in the modifications they can make. Moreover, we find that people’s decision to use a modifiable algorithm is relatively insensitive to the magnitude of the modifications they are able to make. Additionally, we find that giving people the freedom to modify an imperfect algorithm makes them feel more satisfied with the forecasting process, more likely to believe that the algorithm is superior, and more likely to choose to use an algorithm to make subsequent forecasts. This research suggests that one may be able to overcome algorithm aversion by giving people just a slight amount of control over an imperfect algorithm’s forecasts.

Keywords: Decision making, Decision aids, Heuristics and biases, Forecasting, Confidence
Introduction

Forecasts made by evidence-based algorithms are more accurate than forecasts made by humans. This empirical regularity, supported by decades of research, has been observed in many different domains, including forecasts of employee performance (see Highhouse, 2008), academic performance (Dawes, 1971; Dawes, 1979), prisoners’ likelihood of recidivism (Thompson, 1952; Wormith & Goldstone, 1984), medical diagnoses (Adams et al., 1986; Beck et al., 2011; Dawes, Faust, & Meehl, 1989; Grove et al., 2000), demand for products (Schweitzer & Cachon, 2000), and so on (see Dawes, Faust, & Meehl, 1989; Grove et al., 2000; Meehl, 1954). When choosing between the judgments of an evidence-based algorithm and a human, it is wise to opt for the algorithm.

Despite the preponderance of evidence demonstrating the superiority of algorithmic judgment, decision makers are often averse to using algorithms, opting instead for the less accurate judgments of humans. Fildes and Goodwin (2007) conducted a survey of 149 professional forecasters from a wide variety of domains (e.g., cosmetics, banking, and manufacturing) and found that many professionals either did not use algorithms in their forecasting process or failed to give them sufficient weight. Sanders and Manrodt (2003) surveyed 240 firms and found that many did not use algorithms for forecasting, and that firms that did use algorithms made fewer forecasting errors. This failure to use forecasting algorithms extends beyond corporations. Vrieze and Grove (2009) surveyed

---

13 In this paper, the term “algorithm” describes any evidence-based forecasting formula, including statistical models, decision rules, and all other mechanical procedures used for forecasting.
183 clinical psychologists and found that only 31% of them used algorithms when making clinical predictions.

Although many professional forecasters fail to use algorithms in practice, recent research has shown that people are not always averse to using algorithms to make predictions. Dietvorst, Simmons, and Massey (2015) gave participants the choice of either exclusively using an algorithm’s forecasts or exclusively using their own forecasts during an incentivized forecasting task, and manipulated whether participants had experience with the algorithm prior to making this choice. They found that most participants chose to use the algorithm exclusively when they had no information about the algorithm’s performance, suggesting that people are not always averse to exclusive reliance on algorithms. However, participants were much more likely to choose to use human rather than algorithmic forecasts once they had seen the algorithm perform and learned that it was imperfect. Participants’ failure to use the imperfect algorithm persisted even when they had explicitly seen the algorithm outperform the human forecasts, and even when they recognized that the algorithm performed better than they did on average. This suggests that people are reluctant to use superior algorithms that they know to be imperfect, a tendency that Dietvorst et al. called algorithm aversion.

Forecasters’ reluctance to use superior but imperfect algorithms instead of inferior human forecasters represents a major challenge for any organization interested in making more accurate forecasts and better decisions, and for any organization that would benefit from persuading their customers to use algorithms. Because many real-world outcomes are far from perfectly predictable, many of even the best forecasting algorithms cannot
possibly produce forecasts that are nearly perfect. As a result, reluctance to use an imperfect algorithm effectively results in a reluctance to use any algorithm after receiving performance feedback. In this article, we offer an approach for overcoming people’s aversion to using imperfect algorithms.

**Overcoming Algorithm Aversion**

Multiple scholars have theorized that people’s reluctance to use algorithms for forecasting stems from an intolerance of error. Einhorn (1986) proposed that forecasters’ intolerance of algorithms arises because although people believe that algorithms will necessarily err, they believe that humans are capable of perfection (also see Highhouse, 2008). Moreover, Dietvorst et al. (2015) found that people were less tolerant of the algorithms’ (smaller) mistakes than of the humans’ (larger) mistakes. These findings do not invite optimism, as they suggest that people will avoid any algorithm that they recognize to be imperfect, even when it is less imperfect than its human counterpart.

Fortunately, people’s distaste for imperfect algorithms may be rooted in more than just an intolerance of error, but also in their beliefs about the qualities of human vs. algorithmic forecasts. Dietvorst et al. (2015) found that although people tend to think that algorithms are better than humans at avoiding obvious mistakes, appropriately weighing attributes, and consistently weighing information, they tend to think that humans are better than algorithms at learning from mistakes, getting better with practice, finding diamonds in the rough, and detecting exceptions to the rule. Indeed, people seem to believe that although imperfect algorithms are better than humans on average, the rigidity of algorithms means they may predictably misfire in ways that humans would not.
This suggests that what people may find especially distasteful about using imperfect algorithms is the lack of flexibility, the inability to intervene when they suspect that the imperfect algorithm has it wrong. If this is true, then people may be more open to using an imperfect algorithm if they are allowed to slightly or occasionally alter its judgments. Although people’s attempts to adjust algorithmic forecasts often make them worse (e.g. Carbone, Andersen, Corriiveau, & Corson, 1983; Goodwin & Fildes, 1999; Hogarth & Makridakis, 1981; Lim & O'Connor, 1995; Willemain, 1991), the benefits associated with getting people to use the algorithm may outweigh the costs associated with making the algorithm’s forecasts slightly worse. This is especially likely to be true if there is a limit on how much people can adjust the algorithm. If allowing people to adjust an imperfect algorithm by only a tiny amount dramatically increases their willingness to use it, then people’s judgments will be much more reliant on the algorithm, and much more accurate as a result.

In this article, we explore when and how forecasters choose to use imperfect algorithms. In three studies, we make five contributions to the literature on algorithm aversion. In Studies 1 and 2, we find that people will choose to use an imperfect algorithm's forecasts substantially more often when they can modify those forecasts, even if they are able to make only small adjustments to those forecasts. In Study 2, we find that people’s choices are surprisingly insensitive to how much they are allowed to adjust an imperfect algorithm's forecasts. In Study 3, we find that people are similarly satisfied with adjusting an algorithm's forecasts in a constrained vs. unconstrained manner, that forecasters who have the ability to adjust an algorithm’s forecasts believe it performs
better than those who do not, and that constraining the amount by which people can adjust an algorithm’s forecasts leads to better performance in the long run (after feedback).

For each study, we report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures. The exact materials and data from each study are available as Online Supplementary Materials here: https://osf.io/5nz9c/.

**Study 1**

**Methods**

**Overview.** In Study 1, we asked participants to forecast students’ scores on a standardized test from nine variables. All participants had the option of using a statistical model to make their forecasts. Participants were informed that the model was imperfect, and off by 17.5 percentiles on average. We manipulated whether or not participants had the option to modify the model’s forecasts. Participants were assigned to one of four conditions. Specifically, they were assigned to either to a condition in which they chose between using the model’s forecasts exclusively or not at all, to one of two conditions in which they were restricted in how much or how frequently they could modify the model’s forecasts if they chose to use them, or to a condition in which they received the model’s forecasts and could use them as much as they wanted. Compared to those who had to choose whether or not to use the model’s forecasts exclusively or not at all, we expected participants who were restrictively able to modify the model’s forecasts to be much more open to using an imperfect algorithm, and to perform better as a result. We were also
curious to see how much weight participants would give to the model’s forecasts when they were shown the model’s forecasts before forming their own opinion, but allowed to use the model’s forecasts as much as they wanted.

**Participants.** This study was conducted in our university’s behavioral lab. Participants received $10 for completing one hour of experiments, of which ours was a 20-minute portion. Participants could earn up to a $5 bonus from our study depending on their forecasting performance. We aimed to recruit over 300 participants for this study, so we ran it in two concurrent lab sessions (the lab at our university has two separate locations) and collected as many participants as we could. The behavioral lab failed to stop 19 participants who had already taken the study from taking it again. We dropped these participants’ second set of responses from our data. Also, 4 participants exited the study before completing their forecasts, leaving us with a sample of 288 participants who completed their forecasts.\(^{14}\) The final sample averaged 22 years of age and was \(66\%\) female.\(^{15}\)

**Procedures.** This experiment was administered as an online survey. Participants began by giving consent and entering their lab identification number. Next, they learned that they would estimate the percentiles of 20 real high school seniors on a standardized math test. They also received a brief explanation of percentiles to ensure that they understood the task. Participants were ensured that the data described real high school

\(^{14}\) Attrition did not differ across conditions in Study 1, Study 2, or Study 3.

\(^{15}\) For Study 1, these demographics exclude 8 participants who did not report their gender and age. The demographics that we report in Studies 2, and 3, exclude 7, and 8 participants for the same reason.
students. Participants then read detailed descriptions of the nine variables that they would receive to make forecasts. Figure 5 shows an example of the stimuli and variables.

**Figure 5.** Example of task stimuli used in all studies.

<table>
<thead>
<tr>
<th>Race</th>
<th>White, non-Hispanic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Socioeconomic status</td>
<td>Fifth quintile (highest)</td>
</tr>
<tr>
<td>Desired occupation at age 30</td>
<td>Healthcare Practitioners and Technical Occupations</td>
</tr>
<tr>
<td>Predicted highest degree</td>
<td>Complete Bachelor’s degree</td>
</tr>
<tr>
<td>Region of country</td>
<td>South</td>
</tr>
<tr>
<td>Times taken PSAT</td>
<td>Twice</td>
</tr>
<tr>
<td>How many friends are not going to college</td>
<td>None of them</td>
</tr>
<tr>
<td>Favorite school subject</td>
<td>Social studies/history/government/civics</td>
</tr>
<tr>
<td>Taken any AP test</td>
<td>No</td>
</tr>
</tbody>
</table>

Participants then learned that analysts had designed a statistical model to forecast students’ percentiles. They (truthfully) learned that the model was based on data from thousands of high school seniors, that the model used the same variables that they would receive, that the model did not have any further information, and that it was “a sophisticated model, put together by thoughtful analysts.” On the next page, participants learned that the model’s estimates for each student were off by 17.5 percentiles on average (i.e., that the model was imperfect). Additionally, they were informed that the model may be off by more or less than 17.5 percentiles for the 20 students that they would be assessing.

Next, participants learned about their incentives. Participants were paid a $5 bonus if their forecasts were within 5 percentiles of students’ actual percentiles on average, and this bonus decreased by $1 for each additional 5 percentiles of average error in participants’ forecasts (this payment rule is reproduced in Appendix C). Thus,

---

16 See the supplement for a more detailed description of this data and the statistical model.
participants whose forecasts were off by more than 25 percentiles received no bonus at all. Participants were required to type the following sentences to ensure that they understood the incentives: “During the official round, you will receive additional bonus money based on the accuracy of the official estimates. You can earn $0 to $5 depending on how close the official estimates are to the actual ranks.”

Next, participants were assigned to one of four conditions. In the can’t-change condition, participants learned that they would choose between exclusively using their own forecasts and exclusively using the model’s forecasts. In the adjust-by-10 condition, participants learned that they would choose between exclusively using their own forecasts and using the model’s forecasts, but that they could adjust all of the model’s forecasts by up to 10 percentiles if they chose to use the model. In the change-10 condition, participants learned that they would choose between exclusively using their own forecasts and using the model’s forecasts, but that they could adjust 10 of the model’s 20 forecasts by any amount if they chose to use the model. Participants in the use-freely condition learned that they would receive the model’s forecasts and could use them as much as they wanted when making their 20 forecasts. Participants were required to type a sentence that described their condition to ensure that they understood the procedures.  

17 Can’t-change: “If you choose to use the statistical model's estimates, you will not be able to change the model's estimates.” Adjust-by-10: “If you choose to use the statistical model's estimates, you will be able adjust the model's estimate for each student by up to 10 percentiles.” Change-10: “If you choose to use the statistical model's estimates, you will be able to overrule 10 of the model's estimates and use your own estimates instead.” Use-freely: “For the 20 official estimates, you can choose to use the model's estimated percentiles as much as you would like to.”
Finally, participants in the can’t-change, adjust-by-10, and change-10 conditions decided whether or not to use the statistical model’s forecasts. After making this choice, participants made 20 incentivized forecasts. The 20 students that participants judged were randomly drawn, without replacement, from a pool of 50 randomly selected high school seniors. Each high school student was presented on an individual page of the survey. Participants in the use-freely condition saw the information describing a student (see Figure 5), saw the model’s forecast for that student, and entered their forecast for that student. Participants who chose not to use the model in the can’t-change, adjust-by-10, and change-10 conditions made their forecasts without seeing the model’s forecasts. Participants in these conditions who chose to use the model entered their own forecasts anyway. In the can’t-change conditions, their own forecasts did not determine their payment; in the adjust-by-10 condition, these forecasts were used to determine their payment, and were required to be within 10 percentiles of the model’s forecasts; and, in the change-10 condition, these forecasts were used to determine their payment, but could not differ from the model for more than 10 of the forecasts.

After completing the forecasts, participants estimated their own average error and the model’s average error, reported their confidence in the model’s forecasts and their own forecasts on 5-point scales (1=none; 5=a lot), and answered two open-ended questions.

---

18 The first option was “Use only the statistical model’s estimated percentiles to determine my bonus” for the can’t-change condition, “Use the statistical model’s estimated percentiles to determine my bonus, adjusting them up to 10 percentiles if need be” for the adjust-by-10 condition, and “Use the statistical model’s estimated percentiles to determine my bonus, overruling up to 10 of them if need be” for the change-10 condition. The second option was “Use only my estimated percentiles to determine my bonus” for all three conditions.

19 We did not find interesting differences between conditions for the performance estimates and confidence measures in Studies 1 and 2. Thus, we report the results of these measures in the Online Supplement.
The first open-ended question asked participants in the can’t-change, adjust-by-10, and change-10 conditions to report why they chose to have their bonus determined by the model’s forecasts or their own forecast, depending on which they had chosen; participants in the use-freely condition reported how much they had used the model’s forecasts. The second question asked all participants to report their thoughts and feelings about the statistical model. After completing these questions, participants learned their bonus and reported it to a lab manager. Finally, participants reported their age, gender, and highest completed level of education.

**Results**

**Choosing to use the model.** As predicted, participants in the adjust-by-10 and change-10 conditions, who were restrictively able to modify the model’s forecasts, chose to use the model’s imperfect forecasts much more often than participants in the can’t-change condition, who could not modify the model’s forecasts (see Figure 6). Whereas only 32% of participants in the can’t-change condition chose to use the model’s forecasts, 73% of participants in the change-10 condition, \( \chi^2(1, N = 145) = 24.19, p < .001 \), and 76% of participants in the adjust-by-10 condition, \( \chi^2(1, N = 146) = 28.40, p < .001 \), chose to use the model. (See Study S2 in the supplement for a replication of this result using a different forecasting task).

---

20 Participants in the use-freely and can’t-change conditions also learned how they performed compared to participants from the same condition in a previous study (Study S1 in the Supplement), reported their confidence in the model’s forecasts and their own forecasts on 5-point scales, and reported their likelihood of using the model to complete this task in the future on 5-point scales. These questions were exploratory. We did not include them in any other study and we do not discuss them further.
As a result of their infrequent use of the model, participants in the can’t-change condition provided forecasts that were further from the model’s than participants in the adjust-by-10, \( t(144) = 3.24, p = .002 \), change-10, \( t(143) = 2.98, p = .003 \), and use-freely, \( t(143) = 3.47, p < .001 \), conditions (see Figure 6). Participants who chose to use the model in the adjust-by-10 and change-10 conditions deviated from the model less than participants in the use-freely condition. Whereas participants in the use-freely condition adjusted the model’s forecasts by 8.18 percentiles on average, participants who chose to use the model in the adjust-by-10 and change-10 conditions adjusted the model’s forecasts by an average of 4.71 percentiles, \( t(124) = -6.17, p < .001 \), and 5.29 percentiles, \( t(121) = -4.79, p < .001 \), respectively. While this comparison should be taken with a grain of salt because of potential selection concerns, it suggests that restricting the amount by which people can adjust a model’s forecasts does result in forecasts that are closer to the model’s. We cleanly test this hypothesis in Study 3.

We also found evidence that participants in the use-freely condition voluntarily integrated the model’s forecasts into their own forecasts. Participants in the use-freely condition provided forecasts that deviated from the model’s forecasts less than half as much (\( M = 8.18 \)) as participants in the can’t-change condition (\( M = 18.66 \)), \( t(143) = 12.52, p < .001 \), who all made their own forecasts without seeing the model’s (regardless of their choice). This suggests that providing forecasters with a model’s forecast before they have a chance to form their own opinion may lead them to anchor on the model’s forecast and, as a result, rely more heavily on the algorithm.
Figure 6. Study 1: Participants who could restrictively modify the model’s forecasts were more likely to choose to use the model, and performed better as a result.

Forecasting performance. As shown in Figure 6, participants who had the option to adjust the model’s forecasts outperformed those who did not. Participants’ forecasts in the can’t-change condition were less accurate, and earned them smaller bonuses, than the forecasts of participants in the adjust-by-10, change-10, and use-freely conditions.\(^1\)

Figure 7 displays the distribution of participants’ performance by condition. Three things are apparent from the figure. First, although the model’s estimates were far from perfect, reliance on the model was strongly associated with better performance. Indeed,

\(^1\)Participants in the can’t-change condition made larger errors on average than participants in the adjust-by-10, \(t(144) = 3.40, p < .001\), change-10, \(t(143) = 3.09, p = .002\), and use-freely, \(t(143) = 4.01, p < .001\), conditions. This translated into participants in the can’t-change condition earning smaller bonuses than participants in the adjust-by-10, \(t(144) = -2.90, p = .004\), change-10, \(t(143) = -2.53, p = .013\), and use-freely, \(t(143) = -2.88, p = .005\), conditions.
failing to choose to use the model was much more likely to result in very large average errors (and bonuses of $0). Second, participants in the can’t-change condition performed worse precisely because they were less likely to use the model, and not because their forecasting ability was worse. Third, exposing all participants in the use-freely condition to the model’s forecasts seems to have prevented them from making very large errors, as no participant erred by more than 28 percentiles on average.

Figure 7. Study 1: The distribution of participants’ average absolute errors by condition and whether or not they chose to use the model’s forecasts.

Discussion. In sum, participants who could restrictively modify the model’s imperfect forecasts were more likely to choose to use the model’s forecasts than those who could not. As a result, they performed better and earned more money. Additionally,
participants who could use the model’s forecasts freely seemed to anchor on the model’s forecasts, which improved their performance by reducing their chances of making large errors.

Study 2

Methods

Overview. Study 1 (and Study S2 in the Supplement) showed that people were more likely to choose to use an imperfect algorithm if they were given the option to restrictively adjust its forecasts. In Study 2, we explored people’s sensitivity to the restriction on their adjustments. Would further restricting the amount by which people can adjust their forecasts diminish their willingness to use the algorithm’s imperfect forecasts, or would people be willing to commit to using an imperfect algorithm as long as they are given even a modicum of control over its forecasts?

To answer this question, we asked participants to engage in the same student forecasting task as in Study 1, and we randomly assigned them to one of four experimental conditions: a can’t-change condition that was unable to modify the algorithm’s forecasts, or one of three conditions in which they could adjust the model’s forecasts by either 10, 5, or 2 percentiles. If participants’ use of the model depends on how much control they have over its imperfect forecasts, then they should be more likely to choose to use the model when they can adjust it by a larger amount (10 percentiles) than by a smaller amount (2 percentiles). However, if participants simply need to have some control over the model’s imperfect forecasts in order to choose it, then they should
be equally likely to choose to use the model no matter whether they can adjust the model by 10, 5, or even 2 percentiles.

**Participants.** MTurk participants earned $1 for completing the study and could earn up to an additional $0.50 depending on their forecasting performance. We decided in advance to recruit 800 participants (200 per condition). Participants began the study by answering a question designed to check whether they were carefully reading instructions. We prevented the 107 participants who failed this check from participating and 131 additional participants quit the survey before completing their forecasts. We replaced these participants, and our final sample consisted of 816 participants who completed their forecasts. The final sample averaged 34 years of age and was 48% female.

**Procedure.** This study used the same forecasting task as in Study 1: Participants predicted the percentiles of high school students on a standardized math test. The procedure was the same as Study 1’s except for five changes. First, the four experimental conditions were different. Participants were randomly assigned to either a can’t-change condition, an adjust-by-10 condition, an adjust-by-5 condition, or an adjust-by-2 condition. In the can’t-change condition, participants who chose to use the model could not modify its forecasts, whereas in the adjust-by-X conditions, participants who chose to use the model could adjust it by X percentiles. For example, in the adjust-by-2 condition, participants who decided to use the model’s forecasts could adjust its forecasts by up to 2 percentiles.
Second, we recruited participants from Amazon Mechanical Turk instead of the laboratory. Third, as previously mentioned, we added a reading check to the beginning of the survey to identify and remove participants who were not reading instructions. Fourth, we used a different payment rule. Participants were paid a $0.50 bonus if their official forecasts were within five percentiles of students’ actual percentiles. This bonus decreased by $0.10 for each additional five percentiles of error in participants’ forecasts (this payment rule is reproduced in Appendix D). As a result, participants whose forecasts were off by more than 25 percentiles received no bonus. Fifth, at the end of the survey we asked participants to recall the model’s average error.

Results

Choosing to use the model. Consistent with the results of Study 1, participants who had the option to adjust the model’s imperfect forecasts chose to use the model more often than participants who could not modify its forecasts (see Figure 8). Whereas only 47% of participants in the can’t-change condition chose to use the model’s forecasts, 70% of participants in the adjust-by-X conditions chose to the model, $\chi^2(1, N = 834) = 36.46$, $p < .001$. Additionally, and somewhat surprisingly, we found that participants’ decision to use the model in the adjust-by-X conditions did not depend on how much they were able to adjust the model: 71%, 71%, and 68% chose to the model in the adjust-by-10, adjust-by-5, and adjust-by-2 conditions. These three conditions did not differ significantly, $\chi^2(2, N = 623) = 0.42$, $p = .809$. Although we cannot reject the possibility that participants may have been slightly sensitive to the amount by which they could adjust the model, we can conclude that their willingness to use the model was not
detectably altered by imposing a fivefold restriction on the amount by which they could adjust. (See Study S3 in the supplement for a replication of this insensitivity using the change-X forecasting process).

Compared to participants in the can’t-change condition, participants in the adjust-by-10 condition deviated from the model directionally less, \( t(407) = 0.71, p = .475 \), and participants in the adjust-by-5, \( t(409) = 2.61, p = .010 \), and adjust-by-2, \( t(406) = 2.74, p = .006 \), conditions deviated significantly less (see Figure 8). Participants who chose to use the model in the adjust-by-X conditions did not deviate from the model as much as they could have, regardless of whether they were in the adjust-by-10 (\( M = 5.00 \)), adjust-by-5 (\( M = 2.61 \)), or adjust-by-2 condition (\( M = 1.33 \)). Given the desire of participants in the adjust-by-10 condition to adjust by 5 percentiles on average, it is surprising that those in the adjust-by-5 and adjust-by-2 conditions did not adjust by close to the maximum amount.
Figure 8. Study 2: Participants who could restrictively modify the model’s forecasts were more likely to choose to use the model, and performed better as a result.

![Figure 8: Graphs showing % of Participants Who Chose To Use Model and Average Absolute Error by condition.]

Note: Errors bars indicate ±1 standard error.

Forecasting performance. As in Study 1, participants who were given the option to adjust the model’s imperfect forecasts performed better than those who were not (see Figure 8). Participants in the can’t-change condition made significantly larger errors than participants in each of the adjust-by-X conditions and earned smaller bonuses as a result.\(^{22}\)

Figure 9 displays the distribution of participants’ performance by condition. We again see that reliance on the model was strongly associated with better performance, even though its forecasts were far from perfect. Also, participants in the can’t-change

\(^{22}\) Participants in the can’t-change condition made significantly larger errors on average than participants in the adjust-by-10, \(t(407) = 2.64, p = .009\), adjust-by-5, \(t(409) = 4.02, p < .001\), and adjust-by-2, \(t(406) = 2.85, p = .005\), conditions. As a result, participants in the can’t-change condition earned significantly smaller bonuses than participants in the adjust-by-10, \(t(407) = -2.08, p = .039\), adjust-by-5, \(t(409) = -3.67, p < .001\), and adjust-by-2, \(t(406) = -2.04, p = .042\), conditions.
condition performed worse precisely because they were less likely to use the model, and not because their forecasting ability was worse.

**Figure 9.** Study 2: The distribution of participants’ average absolute errors by condition and whether or not they chose to use the model’s forecasts.

**Discussion.** In Study 2, participants were once again more likely to choose to use an imperfect algorithm’s forecasts if they could modify those forecasts. Moreover, they were relatively insensitive to the amount by which they could adjust the model’s forecasts. This finding suggests that, while it is beneficial to give people *some* control over an imperfect algorithm’s forecasts, giving them additional control may not further reduce algorithm aversion.
Study 3

In Studies 1 and 2, we found that people were much more likely to choose to use an imperfect algorithm if they were allowed to adjust its forecasts by even a small amount (see also Studies S2 and S3 in the Supplement). However, whereas in each of these studies the decision to use the algorithm was made before participants experienced what it was like to use it and before receiving performance feedback, overcoming algorithm aversion over the long term requires a willingness to use the algorithm even after using it and making errors. This is no small feat, as prior research shows that people punish algorithms more than humans for making the same mistake, rendering them especially reluctant to choose to use algorithms after seeing them err (Dietvorst et al., 2015).

In Study 3, we investigated how people’s experience with a forecasting process in which they either can or cannot modify an imperfect algorithm’s forecasts affects their judgments and subsequent decisions. Using the same forecasting task as Studies 1 and 2, we conducted this experiment in two stages. In the first stage of 10 forecasts, participants were randomly assigned to adhere to one of three forecasting methods. In the model-only condition, participants were forced to use only the model’s estimates for each forecast. In the adjust-by-10 condition, participants could adjust the model’s forecasts by up to 10 percentiles. In the use-freely condition, participants were given the model’s forecasts and could adjust them as much as they wanted. After completing a round of forecasts, participants were asked to indicate their satisfaction with, and confidence in, the forecasting process they just used. Then participants learned their performance for their first round of forecasts.
Next, participants answered questions about three forecasting processes, and then chose which of them to use for a second round of 10 forecasts. All participants chose among using the model exclusively (model-only), using their own forecasts exclusively (human-only), and either adjusting the model’s forecasts in a constrained (adjust-by-10) or unconstrained (use-freely) manner. Participants in the use-freely condition and half of participants in the model-only condition had the option of adjusting the model in an unconstrained manner (use-freely). Participants in the adjust-by-10 condition and the other half of participants in the model-only condition had the option of adjusting the model in a constrained manner (adjust-by-10).

This design allowed us to answer four open questions. First, we examined how experience with different forecasting processes translates into satisfaction with, and confidence in, those processes. After one round of forecasts, are forecasters more satisfied with, and confident in, a process in which they (1) use an imperfect algorithm’s forecasts exclusively, (2) adjust an imperfect algorithm’s forecasts in a constrained manner, or (3) adjust an imperfect algorithm’s forecasts in an unconstrained manner? Studies 1 and 2 suggest that people do prefer adjusting an imperfect algorithm’s forecasts over using them exclusively; moreover, these studies also suggest that people may not object to being partially constrained to the imperfect algorithm.

Second, we examined whether people are willing to continue adjusting an imperfect algorithm’s forecasts in a constrained or unconstrained manner after receiving performance feedback. If allowing people to adjust an imperfect algorithm’s forecasts is an effective long-term prescription for algorithm aversion, people will have to be willing
to stick with this forecasting process after learning that it produces errors. It is not at all obvious that this is the case, as past research (Dietvorst et al., 2015) has found that people fail to use an algorithm exclusively after learning that it produces imperfect but superior forecasts.

Third, we examined whether people’s perceptions of the accuracy of the model’s forecasts relative to their own differs between the use-freely, adjust-by-10, and model-only forecasting processes. Participants who have the ability to adjust the model’s forecasts may subsequently hold the model in higher regard, perhaps because their increased satisfaction with the forecasting process bleeds into their feelings about the model. Fourth, we examined whether adjusting an imperfect algorithm’s forecasts in a constrained versus an unconstrained manner in Round 1 leads to better forecasting performance in Round 2.

Methods

Participants. MTurk participants earned $1 for completing the study and could earn up to an additional $1 depending on their forecasting performance. We decided in advance to recruit 800 participants (200 per condition). Participants began the study by answering a question designed to check whether they were carefully reading instructions. We prevented the 206 participants who failed this check from participating and 208 additional participants quit the survey before completing their forecasts. We replaced these participants, and had a sample of 818 participants who completed their forecasts. The final sample averaged 33 years of age and was 49% female.
**Procedure.** This study was administered as an online survey. Participants began the survey by providing informed consent and entering their Mechanical Turk ID Number. They then completed a question designed to ensure that they were reading the instructions. Only those who answered this question correctly proceeded to the remainder of the survey, which introduced participants to the forecasting task (predicting students’ performance on a standardized test), introduced participants to the statistical model, and informed participants that the model was off by 17.5 percentiles on average. This part of the survey was identical to Studies 1 and 2.

Figure 10 shows the rest of the procedure of Study 3. After reading about the forecasting task, participants were told that they would make 10 forecasts and that their performance would be incentivized. Just as in Study 2, they learned that they would be paid a $0.50 bonus if their official forecasts were within five percentiles of students’ actual percentiles on average, and that this bonus decreased by $0.10 for each additional five percentiles of average error. Participants were then assigned to one of three conditions. One-half of the participants were assigned to the model-only condition, in which they were forced to use the model’s forecasts without being able to adjust them. One-quarter of the participants were assigned to the use-freely condition, in which they received the model’s forecasts and could adjust them as much as they wanted. And the remaining one-quarter of participants were assigned to the adjust-by-10 condition, in which they received the model’s forecasts and could adjust them by up to 10 percentiles. Participants were not given the option to only use their own forecasts, as they were in the adjust-by-10 conditions used in Studies 1 and 2. Participants were required to type two
sentences describing their condition’s forecasting procedure to ensure that they understood the instructions.\textsuperscript{23}

\textit{Figure 10. Study 3’s Procedure.}

Next, participants completed their first set of 10 forecasts.\textsuperscript{24} After participants completed these forecasts, they were reminded of the forecasting process that they had used and asked to rate how satisfied they were with that process on a 5-point scale (1 =

\textsuperscript{23}Model-only: “For the following 10 estimates, you will use the model's estimates. You will not be able to change the model's estimates.” Use-freely: “For the following 10 estimates, you can use the model's estimates as much as you would like to. You will see the model's estimate and you can use it to form your estimate.” Adjust-by-10: “For the following 10 estimates, you will use the model's estimates. You will be able adjust the model's estimate for each student by up to 10 percentiles.”

\textsuperscript{24}Unlike in Studies 1 and 2, participants who could not change the model’s forecasts did not make their own forecasts. Instead, they simply viewed the model’s forecast for each student.
very dissatisfied; 5 = very satisfied), and how much confidence they had that the process performed well (1 = none; 5 = a lot). On the next page, participants learned how much their first 10 forecasts had erred on average and how much money they had earned.

Next, participants were presented with three forecasting processes (human-only, model-only, and either adjust-by-10 or use-freely) and asked about their satisfaction with (1 = very dissatisfied; 5 = very satisfied) and confidence in (1 = none; 5 = a lot) these three forecasting processes. They were then asked to choose among these three forecasting processes for a second set of 10 forecasts with the same incentives as the first set. Participants in the use-freely condition and half of participants in the model-only condition rated and chose among the use-freely, model-only, and human-only forecasting processes. Participants in the adjust-by-10 condition and the other half of participants in the model-only condition rated and chose among the adjust-by-10, model-only, and human-only forecasting processes.

After completing the second set of 10 forecasts, participants estimated their own average error and the model’s average error, reported their confidence in the model’s forecasts and their own forecasts on 5-point scales (1 = none; 5 = a lot), and reported their thoughts and feelings about the statistical model. Finally, participants reported their age, gender, and highest level of education.

25 We do not report the results of these ratings in the paper as it is not clear if the differences among conditions are due to performance feedback or the context of the other options that participants were presented with. The results of these measures are presented in Figure S3 in the Online Supplementary Materials.
Results

Confidence in and satisfaction with forecasting process. After participants were randomly assigned to a forecasting process and made their first set of 10 incentivized forecasts, they rated their satisfaction with and confidence in their assigned forecasting process. Participants who were assigned to the model-only process in Round 1 were less satisfied with their forecasting process than participants who were assigned to the adjust-by-10, \( t(614) = -6.59, p < .001 \), and use-freely, \( t(620) = -6.17, p < .001 \), processes (see Figure 11). Also, participants who were assigned to the model-only process in Round 1 were directionally less confident in their forecasting process than participants who were assigned to the adjust-by-10 process, \( t(614) = -1.29, p = .196 \), and marginally less confident in their forecasting process than participants who were assigned to the use-freely process, \( t(620) = -1.68, p = .093 \). Thus, allowing participants to modify the model’s forecasts increased their satisfaction with their forecasting process.

Interestingly, participants’ satisfaction with and confidence in their forecasting process did not differ by whether they adjusted the algorithm’s forecasts in a constrained or unconstrained manner. Participants in the adjust-by-10 condition were about equally as satisfied with, \( t(410) = 0.56, p = .578 \), and confident in, \( t(410) = -0.29, p = .774 \), their assigned forecasting process as participants in the use-freely condition were, even though they had less freedom to modify the algorithm’s forecasts. This suggests that forecasters may not object to being constrained to an imperfect algorithm as long as they can modify its forecasts.
Figure 11. Study 3: Participants who could modify the model’s forecasts were more satisfied.

Note: Errors bars indicate ±1 standard error.

Choice of second forecasting process. When it came time to choose a forecasting process for the second round of incentivized forecasts, most participants chose to use the model’s imperfect forecasts (see Figure 12). In fact, participants in each condition chose to use the model in some manner (either exclusively or partially) over 80% of the time. Participants chose to combine their own judgment with the algorithm’s forecasts (51%-74%) most often, instead of using their own forecasts exclusively (11%-19%) or using the algorithm’s forecasts exclusively (10%-32%). This again suggests that adjusting an imperfect algorithm’s forecasts is a palatable forecasting process. Also, participants in the adjust-by-10 condition, who were constrained to the imperfect algorithm, chose to adjust the model’s forecasts by 10 at virtually the same rate (53%) as participants in the use-freely condition, who were unconstrained, chose to adjust the model’s forecasts freely (51%), $\chi^2(1, N = 410) = 0.15, p = .699$. This suggests that constraining forecasters to an
imperfect algorithm’s forecasts does not increase their likelihood of abandoning the algorithm or their forecasting process.

There is one substantial difference between the choices of participants who could adjust the model’s forecasts in Stage 1 and those who could not. Participants who could (restrictively or freely) modify the model’s forecasts in Stage 1 were much more likely to choose the “model-only” option (30%) than participants who could not modify the model’s forecasts in Stage 1 (12%), $\chi^2(1, N = 823) = 38.45, p < .001$. This suggests that participants who were able to modify the model’s forecasts may have held the model in higher regard (relative to themselves) compared to participants who were not able to modify the model’s forecasts.
Figure 12. Study 3: Participants in each condition chose to modify the model’s forecasts most often, instead of only using their own forecasts or only using the model’s forecasts.

Chosen Second Stage Forecasting Method

Note: Errors bars indicate ±1 standard error.

Perceptions of model and self. Participants’ confidence ratings and performance estimates suggest that allowing people to modify an imperfect algorithm’s forecasts may improve their perceptions of the algorithm relative to themselves (see Figure 13). Compared to participants who could not modify the model’s forecasts during the first set
of forecasts, participants who could modify the model’s forecasts had more confidence in
the model’s forecasts relative to their own, \( t(816) = -5.86, p < .001 \), and estimated that the
model’s average absolute error was better relative to their own, \( t(817) = 3.92, p < .001 \). These
results suggest that people may hold algorithms in higher regard relative to
themselves if they previously had the ability to modify the algorithm’s forecasts, and are
consistent with participants’ increased selection of the model-only forecasting process in
the adjust-by-10 and use-freely conditions.

Figure 13. Study 3: Participants who could modify the model’s forecasts were more
confident in the model’s forecasts and thought that the model performed better relative to
themselves.

Note. Error bars indicate \( \pm 1 \) standard error.

---

26 These analyses were conducted with OLS regressions that included participants’ confidence ratings or AAE
estimates as the dependent variable, with two observations per participant (confidence in model & confidence in
human, or model AAE estimate & human AAE estimate). The regressions included a dummy indicating whether each
observation described the model or the participant, a dummy indicating whether or not the participant could adjust the
model’s forecasts during the first round, and an interaction between both dummies. Standard errors were clustered by
participant. The t-tests reported correspond to the coefficient of the interaction term in each regression.
Figure 14. Study 3: Participants who had the option to adjust the model restrictively in the second stage of forecasts performed better and earned more money.

Forecasting Performance. As shown in Figure 14, participants in the adjust-by-10 condition deviated from the model’s forecasts far less than participants in the use-freely condition during the Stage 1 forecasts, $t(410) = -11.61, p < .001$. Although this only translated into directionally better average absolute errors in Stage 1 of this study, $t(410) = -1.22, p = .224$, we are confident that giving more weight to a model’s forecasts would increase forecasting performance in a larger sample of forecasts. Participants in the model-only condition, who could not adjust the model’s forecasts, had average absolute errors that were directionally better than participants in the adjust-by-10 condition, $t(616) = -0.63, p = .531$, and significantly better than participants in the use-freely condition, $t(622) = -2.32, p = .021$. As a result of these performance differences, participants in the use-freely condition earned significantly smaller bonuses ($M = $0.18) than participants in
the model-only condition \((M = 0.19)\), \(t(622) = 3.03, p = .003\), and directionally smaller bonuses than participants in the adjust-by-10 condition \((M = 0.19)\), \(t(410) = 1.27, p = .207\). Thus, the more participants were required to use the model in the first forecasting stage, the better they performed.

More important is how participants fared in Stage 2, when they could choose to either completely use the model, completely use their own forecasts, or to adjust the model’s forecasts. Participants in the adjust-by-10 condition deviated from the model’s forecasts less than participants in the use-freely condition, \(t(403) = -3.45, p < .001\). Also, participants in the model-only condition who had the option to use the adjust-by-10 process deviated from the model’s forecasts less than participants in the model-only condition who had the option to use the use-freely process, \(t(411) = -2.23, p = .027\). As described below, these differences in deviation from the model were driven by participants who chose to use the adjust-by-10 and use-freely forecasting processes.

Participants who chose to use the model-only process did not deviate at all from the model’s forecasts, and there were not significant between-condition differences in how far participants who chose the human-only process deviated from the model, \(F(3, 128) = 1.95, p = .125\). However, participants in the adjust-by-10 condition who chose to use the adjust-by-10 process deviated from the model substantially less \((M = 4.65)\) than participants in the use-freely condition who chose to use the use-freely process \((M = 6.84)\), \(t(208) = -4.21, p < .001\). Additionally, participants in the model-only condition who chose to use the adjust-by-10 process deviated from the model substantially less \((M = 5.31)\) than participants in the model-only condition who chose to use the use-freely
process \( (M = 8.26) \), \( t(300) = -7.23, p < .001 \). Thus, participants who were given the option to restrictively modify the model’s forecasts (i.e. adjust-by-10) deviated from the model much less than participants who were given the option to modify the model without restriction (i.e. use-freely). As shown in the next paragraph, this difference in deviation from the model did translate into significant performance differences.

In Stage 2, participants who used the adjust-by-10 process performed better than those who used the use-freely process. Participants who chose the adjust-by-10 process had lower average absolute errors \( (M = 17.90) \) than participants who chose the use-freely process \( (M = 20.25) \), \( t(510) = -6.28, p < .001 \), lower average absolute errors than participants who chose to use their own forecasts \( (M = 24.50) \), \( t(382) = 14.01, p < .001 \), and similar average absolute errors to participants who chose to use the model’s forecasts \( (M = 18.20) \), \( t(424) = 0.97, p = .331 \). Participants who used the adjust-by-10 process outperformed those who used the use-freely process specifically because they provided forecasts that were closer to the model’s \( (M = 5.04) \) compared to participants who used the use-freely process \( (M = 7.67) \), \( t(510) = 8.10, p < .001 \).27 As a result of these differences between the adjust-by-10 and use-freely forecasting processes, participants who had the option to use the adjust-by-10 process in Stage 2 (i.e. the adjust-by-10 condition, \( \frac{1}{2} \) of the model-only condition) had lower average absolute errors than, and

---

27 We conducted a mediation analysis, where the dependent variable was participants’ average absolute error, the mediator was participant’s average absolute deviation from the model’s forecasts, and the independent variable was whether participants used the adjust-by-10 process or the use-freely process. We included all participants who used the adjust-by-10 or use-freely processes for Stage 2. We then used Preacher and Hayes’s (2008) bootstrapping procedure to obtain unbiased 95% confidence intervals around the mediated effects. Average absolute deviation from the model’s forecasts significantly mediated the effect of having the adjust-by-10 versus the use-freely option on average absolute error (-2.19, -0.85).
earned more money than, participants who had the option to instead use the use-freely process (i.e. the use-freely condition, the other ½ of the model-only condition).  

**Discussion.** Taken together, these results inform multiple open questions related to people’s use of imperfect algorithms. First, they highlight the substantial (and, to us, surprising) benefits of letting people modify an imperfect algorithm’s forecasts. It increases their satisfaction with the process, their confidence in and perceptions of the model relative to themselves, and their use of the model on subsequent forecasts. Second, the results show that people who are able to modify an imperfect algorithm’s forecasts by a limited amount will not necessarily be less satisfied than if they can modify it by an unlimited amount, and that people will elect to modify an algorithm’s forecasts in a constrained manner even after using this process and seeing it err. Third, the results show that restricting people’s adjustments to the model, rather than allowing them to use it freely, prevents one from making forecasts that deviate greatly from the model and thus improves forecasting performance.

**General Discussion**

Our studies show that people will use imperfect algorithms to make incentivized forecasts so long as they can slightly modify them. Although people often fail to use imperfect algorithms exclusively, they will commit to use imperfect algorithms in a

---

28 Participants in the adjust-by-10 condition had lower average absolute errors than, and earned more money than, participants in the use-freely condition had, \( t(403) = -3.86, p < .001 \), and earned, \( t(403) = 3.59, p < .001 \), or participants in the model-only condition who had the use-freely option had, \( t(404) = 3.92, p < .001 \), and earned, \( t(404) = -3.68, p < .001 \). Participants in the model-only condition who had the adjust-by-10 option had lower average absolute errors than, and earned more money than, participants in the use-freely condition had, \( t(410) = -2.59, p = .010 \), and earned, \( t(410) = 2.99, p = .003 \), or participants in the model-only condition who had the use-freely option had, \( t(411) = -2.71, p = .007 \), and earned, \( t(411) = 3.10, p = .002 \).
constrained manner. Further, we found evidence that people are insensitive to the amount by which they can modify the imperfect algorithm’s forecasts when making this decision. We also found that allowing people to adjust an algorithm’s forecasts has additional benefits. Participants who were able to modify an imperfect algorithm’s forecasts reported higher satisfaction with their forecasting process, and thought that the algorithm performed better relative to themselves compared to participants who could not modify the algorithm’s forecasts. Additionally, we found that people are not less satisfied modifying an algorithm’s forecasts in a constrained manner versus an unconstrained manner. Finally, we found that restricting the amount by which people can modify an algorithm’s forecasts leads them to deviate from the algorithm less and thus to perform better.

These findings have many important implications for managers trying to increase employees’ and customers’ use of algorithms. First, framing the decision of whether or not to use an algorithm as an all-or-nothing decision is likely to be counterproductive. People are unlikely to commit to using an algorithm’s forecasts exclusively after getting performance feedback or learning that it is imperfect. Further, forcing employees into a regime where they have to use an imperfect algorithm’s forecasts exclusively may lead them to become dissatisfied or push for a change. However, asking people to commit to modifying an algorithm’s forecasts by a limited amount seems much more palatable. People will be much more likely to choose to use an imperfect algorithm if they can modify its forecasts, and employees will not necessarily be dissatisfied if they are partially constrained to an imperfect algorithm’s forecasts. Finally, the results of Study 3
suggest that constraining employees to an imperfect algorithm will not only be acceptable to them, but may also lead them to give the algorithm’s forecasts more weight and perform better than letting them use the algorithm’s forecasts in an unconstrained manner. If for some reason having employees make constrained adjustments to an algorithm’s forecasts is not possible, Study 1 shows that having employees make unconstrained adjustments to an algorithm’s forecasts can also substantially improve their forecasting performance.

It is worth noting that our techniques for increasing people’s choice of algorithms are likely to work for people’s choice of other decision aids as well. For example, people often give less weight to other people’s advice than they should when left to their own devices (see Bonaccio & Dalal, 2006); however, it is possible that people will commit to using another person’s advice in a constrained manner. In order to test this possibility, we ran a study with four between-subjects conditions that differed in whether (1) participants had the option to use the imperfect forecasts of another person (Human conditions) or a statistical algorithm (Algorithm conditions), and (2) participants were be unable to adjust the forecasts of that entity (Can’t-change conditions) or were able to adjust the forecasts of that entity by up to 5 percentiles (Adjust-by-5 conditions). This resulted in a 2 (Entity: Human vs. Algorithm) x 2 (Adjustment: Can’t-change vs. Adjust-by-5) design (see Study S5 in the supplement for a detailed write up). We found that giving participants the ability to modify the other entity’s forecasts increased participants’ choice of the human’s forecasts (21% to 41%) to a similar degree that it increased their

29 All participants were told that the algorithm’s or human’s forecasts were off by 17.5 percentiles on average.
choice of the algorithm’s forecasts (46% to 69%; \( z (N = 2,014) = -0.01, p = .996 \)). Thus, in general, people may be more willing to pre-commit to using information in a constrained manner when they have the ability to modify it.

**Limitations and Future Directions.**

The studies in this paper leave some questions unanswered. First, the results may turn out differently with different forecasting tasks. For example, when forecasters have important information that an algorithm does not have or when they are forecasting a time series, allowing them to make large adjustments to the algorithm’s forecasts may actually increase accuracy (e.g. Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009; Lawrence, Goodwin, O’Connor, & Önkal, 2006). Thus, in these cases, constraining forecasters tightly to an imperfect algorithm’s forecasts may not improve forecasting performance.

Second, there could be conditions under which the effects we found would be diminished or eliminated. For example, people may not be willing to use algorithms that are far more imperfect than the algorithms that we employed. Additionally, although we did find that participants were insensitive to the amount that they could adjust the model’s forecasts, we only gave participants the option to adjust the model by 2 to 10 percentiles. It is possible that more participants would have chosen to use the model if they could adjust it to a greater degree (e.g. 20 percentiles), or that fewer participants would have chosen to use the model if they could adjust it to a smaller degree (e.g. 1 percentile). Third, while we did use two different populations in our studies, it is possible
that the effects we found are dependent on some characteristics of those tasks or populations. Future work could investigate the effects shown in this paper with different populations of participants, different algorithms that are more or less accurate than those used in our studies, and in different forecasting domains. Research with a population of professional forecasters would be especially informative.

In conclusion, we found that letting people adjust an imperfect algorithm’s forecasts increases their likelihood of using it and their confidence in it. We also found that people are insensitive to the amount by which they can adjust the algorithm’s forecasts, and that restricting the amount that people can adjust an algorithm’s forecasts leads to better performance. Participants in our studies did often worsen the algorithm’s forecasts when given the ability to adjust them; however, we may have to accept this error in order to have people make less error overall.
APPENDICIES

Appendix A
Participants in the average absolute error condition of Study 2 (Chapter 1) were paid as follows:
$10 - within 4 percentiles of student’s actual percentile on average
$9 - within 8 percentiles of student’s actual percentile on average
$8 - within 12 percentiles of student’s actual percentile on average
$7 - within 16 percentiles of student’s actual percentile on average
$6 - within 20 percentiles of student’s actual percentile on average
$5 - within 24 percentiles of student’s actual percentile on average
$4 - within 28 percentiles of student’s actual percentile on average
$3 - within 32 percentiles of student’s actual percentile on average
$2 - within 36 percentiles of student’s actual percentile on average
$1 - within 40 percentiles of student’s actual percentile on average

Appendix B
Participants in Studies 3a and 3b (Chapter 1) were paid as follows:
$1.00 - perfectly predict state's actual rank
$0.85 - within 1 rank of state's actual rank
$0.70 - within 2 ranks of state's actual rank
$0.55 - within 3 ranks of state's actual rank
$0.40 - within 4 ranks of state's actual rank
$0.25 - within 5 ranks of state's actual rank
$0.10 - within 6 ranks of state's actual rank

Appendix C
Participants in Study 1 (Chapter 2) were paid as follows:
$5 - within 5 percentiles of student's actual percentiles on average
$4 - within 10 percentiles of student's actual percentiles on average
$3 - within 15 percentiles of student's actual percentiles on average
$2 - within 20 percentiles of student's actual percentiles on average
$1 - within 25 percentiles of student's actual percentiles on average

Appendix D
Participants in Studies 2 and 3 (Chapter 2) were paid as follows:
$0.50 - within 5 percentiles of student's actual percentiles on average
$0.40 - within 10 percentiles of student's actual percentiles on average
$0.30 - within 15 percentiles of student's actual percentiles on average
$0.20 - within 20 percentiles of student's actual percentiles on average
$0.10 - within 25 percentiles of student's actual percentiles on average
*Participants in study 3 were paid separately for each 2 rounds of 10 forecasts


