Journal of Experimental Psychology: General

0096-3445/14/$12.00 http://dx.doi.org/10.1037/xge0000033

Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err

Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey
University of Pennsylvania

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 5 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

Supplemental materials: http://dx.doi.org/10.1037/xge0000033.supp

Imagine that you are an admissions officer for a university and 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 book *Clinical Versus Statistical Prediction: A Theoretical Analysis and Review of the Evidence*, Paul Meehl (1954) 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, Zald, Lebow, Snitz, and Nelson (2000) meta-analyzed 136 studies investigating the prediction of human health and behavior. They found that algorithms 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, Yankelevich, & Highhouse, 2011; Eastwood, Snook, & Luther, 2012), more strongly weigh human input than algorithmic input (Önkal, Goodwin, Thomson, Gönül, & Pollock, 2009; Promberger & Baron, 2006), and more harshly judge professionals who seek out advice from an algorithm rather than from a human (Shaffer, Probst, Merkle, Arkes, & Medow, 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.

Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey, The Wharton School, University of Pennsylvania.

We thank Uri Simonsohn and members of The Wharton Decision Processes Lab for their helpful feedback. We thank the Wharton Behavioral Laboratory and the Wharton Risk Center Ackoff Doctoral Student Fellowship for financial support.

Correspondence concerning this article should be addressed to Berkeley J. Dietvorst, The Wharton School, University of Pennsylvania, 500 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104. E-mail: diet@wharton.upenn.edu

[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.

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; 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 coworker 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 in 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 versus 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 master's of business administration (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 online supplemental materials.

Method

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-min 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 2 weeks; in Study 2 we recruited as many as we could in 1 week; and in Study 4, each participant was yoked to a different participant from Study 1, and so we decided to recruit exactly as many participants as had fully completed every question in Study 1. In Studies 1, 2, and 4, 8, 4, and 0 participants, respectively, exited the survey before completing the study's key dependent measure, leaving us with final samples

[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).

Table 1
Studies 1, 2, and 4: Belief and Confidence Measures

	Study 1	Study 2	Study 4
How much bonus money do you think you would earn if your own estimates determined your bonus? (0–10)		✓b	
How much bonus money do you think you would earn if the model's estimates determined your bonus? (0–10)		✓b	
What percent of the time do you think the model's estimates are within 5 percentiles of a student's true score? (0–100)	✓a		✓a
What percent of the time do you think your estimates are within 5 percentiles of a student's true score? (0–100)	✓a		✓a
On average, how many percentiles do you think the model's estimates are away from students' actual percentiles? (0–100)		✓a	
On average, how many percentiles do you think your estimates are away from students' actual percentiles? (0–100)		✓a	
How much confidence do you have in the statistical model's estimates? (1 = *none*; 5 = *a lot*)	✓a	✓a	✓a
How much confidence do you have in your estimates? (1 = *none*; 5 = *a lot*)	✓a	✓a	✓a
How well did the statistical model perform in comparison to your expectations? (1 = *much worse*; 5 = *much better*)		✓a	
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)	✓a	✓a	✓a
What are your thoughts and feelings about the statistical model? (open-ended)	✓a	✓a	✓a

Note. ✓a = the measure was collected *after* participants completed the Stage 2 forecasts; ✓b = the measure was collected *before* participants completed the Stage 2 forecasts. All measures were collected after participants decided whether to tie their bonuses to the model or the human. 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."

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 the Amazon.com Mechanical Turk (MTurk) Web site. 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 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

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

	Study 3a	Study 3b
On average, how many ranks do you think the model's estimates are away from states' actual ranks? (0–50)	✓a	✓b
On average, how many ranks do you think your estimates are away from states' actual ranks? (0–50)	✓a	✓b
How much confidence do you have in the statistical model's estimates? (1 = *none*; 5 = *a lot*)	✓a	✓b
How much confidence do you have in your estimates? (1 = *none*; 5 = *a lot*)	✓a	✓b
How likely is it that the model will predict a state's rank almost perfectly? (1 = *certainly not true*; 9 = *certainly true*)*	✓a	✓b
How likely is it that you will predict a state's rank almost perfectly? (1 = *certainly not true*; 9 = *certainly true*)		✓b
How many of the 50 states do you think the model would estimate perfectly? (0–50)		✓b
How many of the 50 states do you think you would estimate perfectly? (0–50)		✓b
How likely is the model to make a really bad estimate? (1 = *extremely unlikely*; 9 = *extremely likely*)		✓b
How well did the statistical model perform in comparison to your expectations? (1 = *much worse*; 5 = *much better*)	✓a	
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)	✓a	✓a
What are your thoughts and feelings about the statistical model? (open-ended)	✓a	✓a

Note. ✓a = the measure was collected *after* participants completed the Stage 2 forecasts; ✓b = the measure was collected *before* participants completed the Stage 2 forecasts. All measures were collected after participants decided whether to tie their bonuses to the model or the human. Questions are listed in the order in which they were asked.

* In Study 3a, the wording of this question was slightly different: "How likely is it that the model will predict states' ranks almost perfectly?"

annual poll of MBA students around the Unites States), 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."[3] Participants were further told that the model was designed to predict each applicant's percentile among his or 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.

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 ($n = 91$) 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."

Undergraduate Degree	Business
GMAT - Verbal	41/60
GMAT - Quantitative	47/60
Essay Score	Good
Interview Score	Good
Work Experience (years)	5
Average Salary	$55,333
Average of Parents' Education	Undergraduate degree(s)

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

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 the question "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.[4] 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.

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 if 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 that of Study 1, 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 ≤ 4, and this bonus decreased by $1 for each four additional units of average error. This payment rule is reproduced in Appendix A.

[3] The statistical model was built using the same data provided to participants and is described in the supplemental materials.

[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.

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.

The Study 3a procedure was the same as Study 1 except for a few changes. Most notably, the forecasting task was different. The Study 3a 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 to 2010 and the same variables provided to participants; it is described in more detail in the supplemental materials.

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.

Number of Major Airports	1
Census Population Rank - 2010	9
Number of Counties Rank	2
Median Household Income Rank - 2008	23
Domestic Travel Expenditure Rank - 2009	9

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

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.

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 that of Study 1, 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

[5] As described in the supplemental materials, the replication attempt of Study 3b was motivated by having observed some weaker results in similar studies run prior to Study 3a. This study ensured that the Study 3a findings were not due to chance.

Table 3
Studies 1–4: Forecasting Performance of Model Versus Human

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

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.

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.[6] 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.[7] 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 abandon humans that make mistakes, even though, as is often the case, the humans' mistakes were larger.

Figure 3 reveals additional findings of interest. First, although all three of the Study 2 incentive conditions showed the hypothesized aversion to using the algorithm, the magnitude of this aversion did differ across conditions, $\chi^2(2, N = 206) = 8.50, p =$

[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 as follows: 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, $\chi^2(1, N = 235) = 5.24, p = .022$.

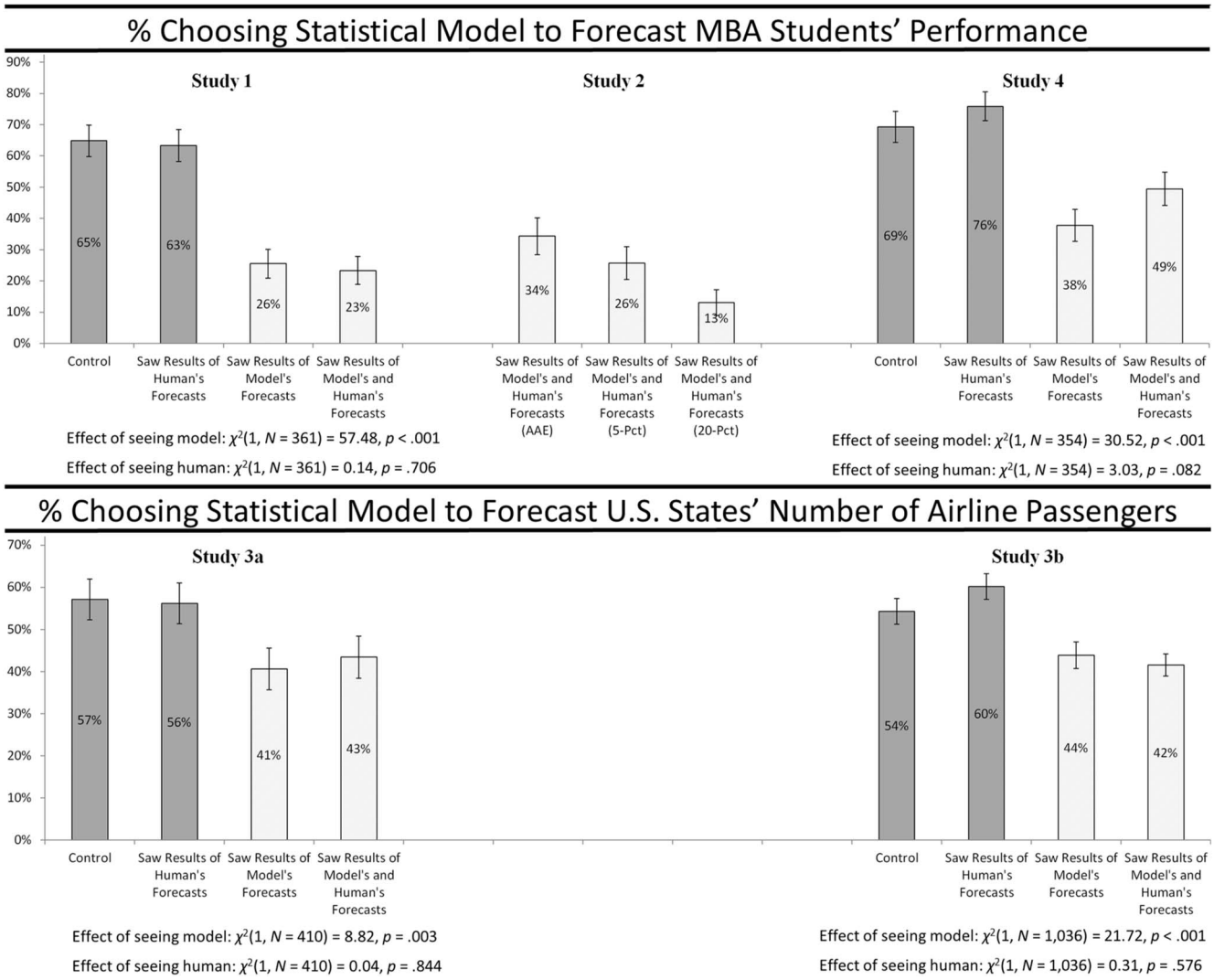

Figure 3. Studies 1–4: Participants who saw the statistical model's results were less likely to choose it. Errors bars indicate ±1 standard error. In Study 2, "AAE," "5-Pct," and "20-Pct" signify conditions in which participants were incentivized either for minimizing average absolute error, for getting within 5 percentiles of the correct answer, or for getting within 20 percentiles of the correct answer, respectively. AAE = average absolute error; Pct = percentile.

.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, $\chi^2(1, N = 136) = 8.56$, $p = .003$.[8] 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.[9]

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

[9] We conducted a binary mediation analysis, in which 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 (2008) 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, 95% CI [–.057, –.190], but confidence in the model's forecasts did not, 95% CI [–.036, .095].

Table 4
Confidence in Model's and Human's Forecasts: Means (and Standard Deviations)

	Control	Human	Model	Model-and-human
Confidence in model's forecasts				
Study 1	3.04_{a} (0.86)	3.17_{a} (0.82)	2.49_{b} (0.71)	2.63_{b} (0.68)
Study 2				2.77 (0.83)
Study 3a	3.40_{a} (0.83)	3.57_{a} (0.73)	3.34_{a} (0.79)	3.29_{a} (0.79)
Study 3b	3.75_{a} (0.75)	3.61_{a} (0.76)	3.34_{b} (0.74)	3.36_{b} (0.69)
Study 4	3.30_{a} (0.80)	3.28_{a} (0.75)	2.86_{b} (0.73)	2.87_{b} (0.86)
Confidence in human's forecasts				
Study 1	2.70_{a} (0.80)	2.47_{a} (0.69)	2.60_{a} (0.75)	2.66_{a} (0.75)
Study 2				2.85 (0.89)
Study 3a	2.85_{a} (0.83)	2.90_{a} (0.95)	3.07_{a} (1.01)	3.03_{a} (0.90)
Study 3b	2.92_{a} (0.85)	2.78_{a} (0.78)	2.83_{a} (0.81)	2.90_{a} (0.80)
Study 4	3.11_{a} (0.73)	2.79_{b} (0.69)	3.01_{ab} (0.73)	2.97_{ab} (0.83)

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

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). 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.[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 cases, confidence in the model's forecasts significantly mediated the effect, whereas confidence in the human did not.[11]

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 *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 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,

[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, $t(351) = 2.28$, $p = .023$.

[11] For confidence in the model's forecasts, the 95% confidence intervals were as follows: Study 1, 95% CI [–.165, –.070]; Study 3a, 95% CI [–.071, –.004]; Study 3b, 95% CI [–.112, –.060]; Study 4, 95% CI [–.174, –.068]. For confidence in the human's forecasts, the 95% confidence intervals were as follows: Study 1, 95% CI [–.029, .013]; Study 3a, 95% CI [–.073, .004]; Study 3b, 95% CI [–.033, .027]; Study 4, 95% CI [–.043, .026].

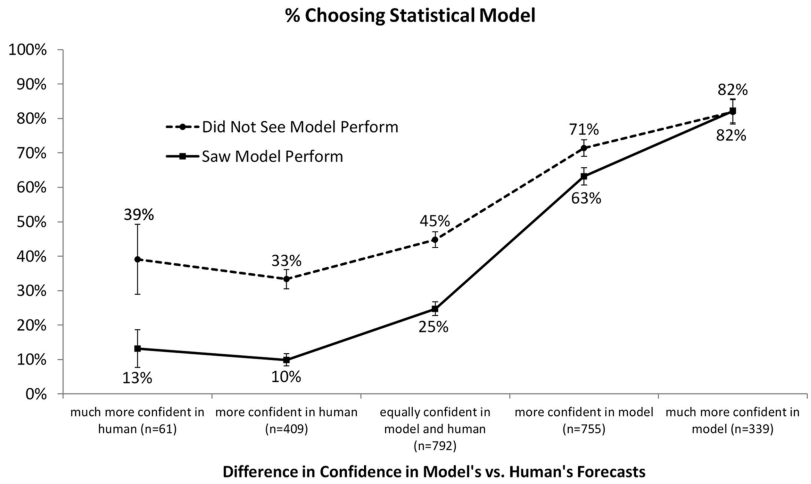

Figure 4. Most people do not choose the statistical model unless they are more confident in the model's forecasts than in the human's forecasts. Errors bars indicate ±1 standard error. The "Did Not See Model Perform" line represents results from participants in the control and human conditions. The "Saw 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.

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.

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 5 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

Table 5
Estimates of Model's and Human's Performance: Means (and Standard Deviations)

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

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)

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

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

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] 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.

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 Dawes (1979), Einhorn (1986), 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.

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; Grove et al., 2000; Meehl, 1954). The ubiquity of computers and the growth

[12] Sorry.

Table 7
Participants' Perceptions of the Model Versus Human Forecaster on Specific Attributes: Means (and Standard Deviations)

	Study 2	Study 3a	Study 3b	Study 4
Detecting exceptions	3.55_h (0.99)	3.02 (1.08)	2.98 (1.08)	3.91_h (0.97)
Finding underappreciated candidates	3.74_h (0.96)			4.04_h (0.98)
Avoiding obvious mistakes	2.68_m (1.10)	2.64_m (1.03)	2.62_m (1.02)	2.55_m (1.13)
Learning from mistakes	3.91_h (0.81)	3.74_h (0.92)	3.67_h (0.95)	3.81_h (0.99)
Appropriately weighing a candidate's qualities (state's attributes)	2.98 (1.09)	2.50_m (0.92)	2.34_m (0.93)	2.81_m (1.11)
Consistently weighing information	2.33_m (1.10)	2.49_m (1.00)	2.29_m (0.98)	2.05_m (1.02)
Treating each student (state) individually	3.60_h (1.02)	2.94 (1.02)	2.89_m (1.02)	3.48_h (1.25)
Getting better with practice	3.85_h (0.82)	3.66_h (0.96)	3.63_h (0.98)	3.77_h (1.08)

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, from 1 (*Model is much better*) to 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.

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.

References

Arkes, H. R., Dawes, R. M., & Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes, 37,* 93–110. http://dx.doi.org/10.1016/0749-5978(86)90046-4

Clark, D. (2012, November 7). How Nate Silver won the 2012 presidential election. *Harvard Business Review Blog*. Retrieved from http://blogs.hbr.org/cs/2012/11/how_nate_silver_won_the_2012_p.html

Davenport, T. H., & Harris, J. G. (2007). *Competing on analytics: The new science of winning*. Boston, MA: Harvard Business Press.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571–582. http://dx.doi.org/10.1037/0003-066X.34.7.571

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243,* 1668–1674. http://dx.doi.org/10.1126/science.2648573

Diab, D. L., Pui, S. Y., Yankelevich, M., & Highhouse, S. (2011). Lay perceptions of selection decision aids in U.S. and non-U.S. samples. *International Journal of Selection and Assessment, 19,* 209–216. http://dx.doi.org/10.1111/j.1468-2389.2011.00548.x

Eastwood, J., Snook, B., & Luther, K. (2012). What people want from their professionals: Attitudes toward decision-making strategies. *Journal of Behavioral Decision Making, 25,* 458–468. http://dx.doi.org/10.1002/bdm.741

Einhorn, H. J. (1986). Accepting error to make less error. *Journal of Personality Assessment, 50,* 387–395. http://dx.doi.org/10.1207/s15327752jpa5003_8

Gerson, M. (2012, November 5). Michael Gerson: The trouble with Obama's silver lining. *The Washington Post*. Retrieved from http://www.washingtonpost.com/opinions/michael-gerson-the-trouble-with-obamas-silver-lining/2012/11/05/6b1058fe-276d-11e2-b2a0-ae18d6159439_story.html

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2,* 293–323. http://dx.doi.org/10.1037/1076-8971.2.2.293

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19–30. http://dx.doi.org/10.1037/1040-3590.12.1.19

Heath, C., & Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty, 4,* 5–28. http://dx.doi.org/10.1007/BF00057884

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1,* 333–342. http://dx.doi.org/10.1111/j.1754-9434.2008.00058.x

Honan, D. (2012, November 7). The 2012 election: A big win for big data. *Big Think*. Retrieved from http://bigthink.com/think-tank/the-2012-election-a-big-win-for-big-data

Isidore, C. (2012, November 7). Nate Silver gets a big boost from the election. *CNN Money*. Retrieved from http://money.cnn.com/2012/11/07/news/companies/nate-silver-election/index.html

Kruger, J. (1999). Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology, 77,* 221–232. http://dx.doi.org/10.1037/0022-3514.77.2.221

McDermott, J. (2012, November 7). Nate Silver's election predictions a win for big data, the New York Times. *Ad Age*. Retrieved from http://adage.com/article/campaign-trail/nate-silver-s-election-predictions-a-win-big-data-york-times/238182/

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and review of the literature*. Minneapolis, MN: University of Minnesota Press.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115,* 502–517. http://dx.doi.org/10.1037/0033-295X.115.2.502

Noveck, J. (2012, November 9). Nate Silver, pop culture star: After 2012 election, statistician finds celebrity. *Huffington Post*. Retrieved from http://www.huffingtonpost.com/2012/11/09 nate-silver-celebrity_n_2103761.html

Önkal, D., Goodwin, P., Thomson, M., Gönül, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making, 22,* 390–409. http://dx.doi.org/10.1002/bdm.637

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40,* 879–891. http://dx.doi.org/10.3758/BRM.40.3.879

Promberger, M., & Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making, 19,* 455–468. http://dx.doi.org/10.1002/bdm.542

Shaffer, V. A., Probst, C. A., Merkle, E. C., Arkes, H. R., & Medow, M. A. (2013). Why do patients derogate physicians who use a computer-based diagnostic support system? *Medical Decision Making, 33,* 108–118. http://dx.doi.org/10.1177/0272989X12453501

Silver, N. (2012). *The signal and the noise: Why so many predictions fail–but some don't*. New York, NY: Penguin Press.

Taylor, C. (2012, November 7). Triumph of the nerds: Nate Silver wins in 50 states. *Mashable*. Retrieved from http://mashable.com/2012/11/07/nate-silver-wins/

Tiku, N. (2012, November 7). Nate Silver's sweep is a huge win for "Big Data". *Beta Beat*. Retrieved from http://betabeat.com/2012/11/nate-silver-predicton-sweep-presidential-election-huge-win-big-data/

(*Appendices follow*)

Appendix A

Payment Rule for Study 2

Participants in the average absolute error condition of Study 2 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

Payment Rule for Studies 3a and 3b

Participants in Studies 3a and 3b 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

Received July 6, 2014
Revision received September 23, 2014
Accepted September 25, 2014 ■