Organizational Behavior and Human Decision Processes 151 (2019) 90–103

Contents lists available at ScienceDirect

Organizational Behavior and Human Decision Processes

journal homepage: www.elsevier.com/locate/obhdp

Algorithm appreciation: People prefer algorithmic to human judgment

Jennifer M. Logg[a,*], Julia A. Minson[a], Don A. Moore[b]

[a] Harvard Kennedy School, Harvard University, United States
[b] Haas School of Business, University of California, Berkeley, United States

ARTICLE INFO

Keywords:
Algorithms
Accuracy
Advice-taking
Forecasting
Decision-making
Theory of machine

ABSTRACT

Even though computational algorithms often outperform human judgment, received wisdom suggests that people may be skeptical of relying on them (Dawes, 1979). Counter to this notion, results from six experiments show that lay people adhere *more* to advice when they think it comes from an algorithm than from a person. People showed this effect, what we call *algorithm appreciation*, when making numeric estimates about a visual stimulus (Experiment 1A) and forecasts about the popularity of songs and romantic attraction (Experiments 1B and 1C). Yet, researchers predicted the opposite result (Experiment 1D). Algorithm appreciation persisted when advice appeared jointly or separately (Experiment 2). However, algorithm appreciation waned when: people chose between an algorithm's estimate and their own (versus an external advisor's; Experiment 3) and they had expertise in forecasting (Experiment 4). Paradoxically, experienced professionals, who make forecasts on a regular basis, relied less on algorithmic advice than lay people did, which hurt their accuracy. These results shed light on the important question of when people rely on algorithmic advice over advice from people and have implications for the use of "big data" and algorithmic advice it generates.

1. Introduction

Although people often receive advice from other people, the rise of "big data" has increased both the availability and utility of a new source of advice: algorithms. The superior accuracy of algorithmic judgment relative to human judgment (Dawes, Faust, & Meehl, 1989) has led organizations to invest in the power of algorithms – scripts for mathematical calculations – to sift through data and produce insights. Companies including Johnson & Johnson and Jet Blue invest in complex algorithms to hire promising employees, track the satisfaction of current employees, and predict which employees are at risk for leaving the organization. Individuals in need of legal counsel can now use "Robo-advisors" such as DoNotPay to help them contest parking tickets or even file for political asylum. Indeed, decision-makers now rely on algorithms for personal use, in place of human secretaries, travel agents, headhunters, matchmakers, D.J.s, movie critics, cosmetologists, clothing stylists, sommeliers, and financial advisors (Siri, Google Search, and spell-check; Kayak, LinkedIn, OkCupid, Pandora, Netflix, Birchbox, Stitch Fix, Club W, and Betterment, respectively).

Such widespread reliance on algorithmic advice seems at odds with the judgment and decision-making literature which demonstrates human distrust of algorithmic output, sometimes referred to as "algorithm aversion" (Dietvorst, Simmons, & Massey, 2015).[1] This idea is so prevalent that it has been adopted by popular culture and the business press (Frick, 2015). Articles advise business leaders on how to overcome aversion to algorithms (Harrell, 2016). Companies in the business of selling algorithmic advice often go to great lengths to present it as purely human-generated. Stitch Fix, for example, uses a combination of algorithmic and human judgment to provide clothing recommendations to consumers. Yet, each shipment of clothes includes a personalized note from a stylist in order to focus consumers' attention on the human component.

In the present paper, we trace the history of research examining peoples' responses to algorithmic output and highlight boundary conditions for empirical evidence supporting algorithm aversion. We then present results showing that under conditions that apply to many decisions, and across a variety of estimation and forecasting tasks, people actually prefer advice from algorithms to advice from people. We call this effect "algorithm appreciation."

1.1. Prior research on algorithm aversion

The first scholarly reference to psychological distrust of algorithms may belong to Meehl (1954). Importantly, it appears in the discussion

* Corresponding author at: Harvard Kennedy School, Harvard University, 79 John F. Kennedy, Street, Cambridge, MA 02138, United States.
E-mail address: jennifer_logg@hks.harvard.edu (J.M. Logg).

[1] While this influential paper is about the effect that seeing an algorithm err has on people's likelihood of choosing it, it has been cited as being about how often people use algorithms in general.

https://doi.org/10.1016/j.obhdp.2018.12.005
Received 20 April 2018; Accepted 4 December 2018
Available online 05 February 2019
0749-5978/ © 2018 Elsevier Inc. All rights reserved.

section of his classic book outlining the predictive superiority of algorithms over human experts. Specifically, when researchers shared their findings with the experts in question, the conclusions were met with skepticism. It appears that the experts in the 1950s were hesitant to believe that a linear model could outperform their judgment. Similar anecdotes have circulated from other scholars (Dana & Thomas, 2006; Dawes, 1979; Hastie & Dawes, 2001). Over time, these anecdotal claims have become received wisdom in the field of judgment and decision making. In his best-selling book, *Thinking Fast and Slow*, Kahneman recounts Meehl's story: "From the very outset, clinical psychologists responded to Meehl's ideas [accuracy of algorithms] with hostility and disbelief" (2013, pg. 227).

Decades passed before the mistrust of algorithms was empirically tested. The results did not always support the received wisdom. On the one hand, in subjective domains governed by personal taste, people relied on friends over algorithmic recommender systems for book, movie, and joke recommendations (Sinha & Swearingen, 2001; Yeomans, Shah, Mullainathan, & Kleinberg, unpublished data). Participants who imagined themselves as medical patients more frequently followed a subjectively worded recommendation for an operation from a doctor than from a computer (Promberger & Baron, 2006). And an influential set of papers demonstrates that after seeing an algorithm err, people relied more on human judgment than an algorithm's (Dietvorst et al., 2015; Dzindolet, Pierce, Beck, & Dawe, 2002).

On the other hand, work in computer science shows that participants considering logic problems agreed more with the same argument when it came from an "expert system" than when it came from a "human" (Dijkstra, Liebrand, & Timminga, 1998). This preference for algorithmic output persisted, *even* after they saw the algorithm err (Dijkstra, 1999). In other studies, people outsourced their memory of information to algorithmic search engines (Sparrow, Liu, & Wegner, 2011; Wegner & Ward, 2013). In the meantime, companies continue to worry about aversion to their algorithms (Haak, 2017) from their own employees, other client organizations, and individual consumers, even while many people entrust their lives to algorithms (e.g., autopilot in aviation).

1.2. Who is relying on which advice and for what purpose?

Given the richness of the decision-making landscape, it is perhaps not surprising that different studies have arrived at different results. One important feature that distinguishes prior investigations is the comparison of people's reliance on algorithmic judgment to their reliance on their own, self-generated judgments. This approach has intuitive appeal, as managers and consumers often choose whether to follow their own conclusions or rely on the output of an algorithm. However, given the evidence that people routinely discount advice, such paradigms do not answer the question of how people react to algorithms, as compared to *other advisors*. Indeed, the robust result from literature on utilization of human advice is that individuals regularly (and inaccurately) discount the advice of others when making quantitative judgments (Yaniv & Kleinberger, 2000). Research has attributed such underweighting of other's input to egocentrism (Soll & Mannes, 2011). Furthermore, an extensive literature on overconfidence repeatedly demonstrates that individuals routinely report excessive confidence in their own judgment relative to that of their peers (Gino & Moore, 2007; Logg, Haran, & Moore, 2018; Moore & Healy, 2008; Moore, Tenney, & Haran, 2015). These literatures therefore raise the question of whether individuals insufficiently trust algorithms (relative to human advisors) or merely overly trust themselves. Thus, a direct comparison of advice utilization from an algorithm versus a human advisor would prove especially useful.

Another important concern is the quality of the advice in question. The now classic literature on clinical versus actuarial judgment has demonstrated the superior accuracy of even the simplest linear models relative to individual expert judgment (for a meta-analysis, see Grove, Zald, Lebow, Snitz, & Nelson, 2000). Although it is unlikely that research participants are aware of these research findings, they may believe that algorithmic judgment is only useful in some domains. This issue of advice quality particularly plagues studies that compare recommendations from an algorithm to recommendations from a human expert, such as a doctor (Promberger & Baron, 2006). As the normative standard for how a single expert should be weighted is unclear, it is also unclear what we should conclude from people choosing an algorithm's recommendation more or less frequently than a doctor's. Thus, the accuracy of algorithmic advice becomes important for interpreting results. In order to isolate the extent to which judgment utilization is specifically driven by the fact that the advisor happens to be an algorithm, the human and algorithmic advice should be of comparable quality.

Yet another important factor is the domain of judgment. It may be perfectly reasonable to rely on the advice of close friends rather than a "black box" algorithm when making a decision reflecting one's own personal taste (Yeomans, Shah, Mullainathan, & Kleinberg, 2018). Conversely, individuals may feel more comfortable with algorithmic advice in domains that feature a concrete, external standard of accuracy, such as investment decisions or sports predictions. Relatedly, the extent to which some domains may appear "algorithmically appropriate" may depend on the historical use of algorithms by large numbers of people. For example, most people have grown comfortable with weather forecasts from meteorological models rather than one's neighbors because meteorological models have enjoyed widespread use for decades. Conversely, the idea of fashion advice from algorithms is still relatively new and may face greater resistance.

When understanding algorithmic utilization, the algorithms' prior performance, when made available, is useful to consider. The work of Dietvorst et al. (2015) demonstrates that when choosing between their own (or another participant's) estimate and an algorithm's estimate, participants punished the algorithm after seeing it err, while showing greater tolerance for their own mistakes (or another participant's). Importantly, because the algorithm is, on average, more accurate than a single participant, choosing the human estimate decreases judgment accuracy. However, in the control conditions of the Dietvorst et al. studies, participants chose the algorithm's judgment more frequently than they chose their own (or another person's). And although the authors consistently and clearly state that "seeing algorithms err makes people less confident in them and less likely to choose them over an inferior human forecaster" (pg. 10), other research cites the paper as demonstrating generalized algorithm aversion.[2]

1.3. Measuring algorithm appreciation

The current research revisits the basic question of whether individuals distrust algorithmic advice more than human advice. In our experiments, we benchmark people's responses to advice from an algorithmic advisor against their responses to advice from a human advisor. Doing so allows us to take into account that people generally discount advice relative to their own judgment.

Importantly, we also control for the quality of the advice. Participants in different conditions receive identical numeric advice, merely labeled as produced by a human or an algorithmic source, which differentiates our work from past work. Thus, any differences we observe are not a function of advice accuracy, but merely the inferences participants might make about the source. Across experiments, we describe the human and algorithmic advice in several different ways.

[2] For instance, Petropoulos, Fildes, & Goodwin (2016) cite Dietvorst et al. (2015) as evidence of people "rejecting models in favor of their own (mis) judgments even when given evidence of the superior performance of models" (p. 851). And Prahl and Van Swol (2017) cite the paper as evidence that "humans are trusted more than computers" (p. 693).

Table 1
Manipulation wording for the source of advice across experiments.

Experiment Description and Sample Size	Algorithmic Advice Description	Human Advice Description
1A: Weight estimate N = 202	An algorithm ran calculations based on estimates of participants from a past study. The output that the algorithm computed as an estimate was: 163 pounds.	The average estimate of participants from a past experiment was: 163 pounds.
1B: Song Rank Forecasts N = 215	An algorithm estimated: X.	The predicted song rank based on an aggregation of 275 other participants is: X.
1C: Attraction Forecast N = 286	An algorithm estimated this (wo)man's attractiveness (humor/enjoyableness) from Mike's (Julia's) perspective. The algorithm's estimate was: X	In another study, 48 people estimated this (wo)man's attractiveness (humor/enjoyableness) from Mike's (Julia's) perspective. Their estimate was: X
1 D: Researcher Predictions N = 199	Same as 1C	Same as 1C
2: Joint vs. Separate Evaluation N = 154	The output that an algorithm computed as an estimate was: 163 pounds.	The estimate of another participant was: 163 pounds.
3: Self/Other Choice N = 403	Use only the statistical model's estimated rank to determine my bonus.	*Self condition*: Use only my estimated rank to determine my bonus. *Other condition:* Use only the other participant's estimated rank to determine my bonus.
4: National Security Experts N = 301; 70	The estimate from an algorithm is: X.	*Weight estimate:* The estimate of another person was: 163 pounds. *Forecasts:* The average estimate from forecasters in the forecasting tournament is: X.
General Discussion: Normative Standard Benchmark N = 671	… an algorithm, based on estimates of 314 participants who took a past study.	…a randomly chosen participant from a pool of 314 participants who took a past study.

Thus, we rule out alternative explanations based on participants' interpretations of our descriptions (Table 1). Finally, we use a variety of contexts to examine utilization of advice across different judgment domains.

We employ the Judge Advisor System (JAS) paradigm to measure the extent to which people assimilate advice from different sources (Sniezek & Buckley, 1995). The JAS paradigm requires participants to make a judgment under uncertainty, receive input ("advice"), and then make a second, potentially revised judgment. Where appropriate, we incentivized participants' final judgments. The dependent variable, Weight on Advice (WOA), is the difference between the initial and revised judgment divided by the difference between the initial judgment and the advice. WOA of 0% occurs when a participant ignores advice and WOA of 100% occurs when a participant abandons his or her prior judgment to match the advice.

Extensive prior research has documented that when seeking to maximize judgment accuracy, a person who receives advice from a single randomly-selected individual should generally average their own judgment with the advice, as reflected in a WOA of 50% (Dawes & Corrigan, 1974; Einhorn & Hogarth, 1975). Yet, people tend to only update 30–35% on average toward advice, incurring an accuracy penalty for doing so (Liberman, Minson, Bryan, & Ross, 2012; Minson, Liberman, & Ross, 2011; Soll & Larrick, 2009). Various attempts to increase advice utilization have demonstrated the effectiveness of paying for advice (Gino, 2008); highlighting advisor expertise (Harvey & Fischer, 1997; Sniezek, Schrah, & Dalal, 2004); and receiving advice generated by multiple individuals (Mannes, 2009; Minson & Mueller, 2012). Could providing advice from an algorithm also increase people's willingness to adhere to advice?

In addition to capturing advice utilization using a more precise, continuous measure, WOA also taps a somewhat different psychological phenomenon than does choosing between one's own judgment and that of an algorithm. Prior work demonstrates that people are quite attached to their intuitive judgments and prefer to follow them even while explicitly recognizing that some other judgment is more likely to be objectively correct (Woolley & Risen, 2018). This could lead people to endorse an advisor but ultimately fail to act on the advice. By contrast, WOA, especially under incentivized conditions, captures actual change in the participant's *own* judgment as a function of exposure to the advice.

1.4. The present research

In six experiments, participants make quantitative judgments under uncertainty and receive advice. In the majority of experiments, we manipulate the source of advice (human versus algorithm) and examine how much weight participants give to the advice. In Experiments 2 and 4, participants choose the source of advice they prefer. In one experiment, we ask active researchers in the field to predict how participants in one of our experiments responded to advice (Experiment 1D).

Across our experiments, we find that people consistently give more weight to equivalent advice when it is labeled as coming from an algorithmic versus human source. We call this effect *algorithm appreciation*. Yet, researchers predict the opposite; they predict algorithm aversion. Together, these results suggest that algorithm aversion is not as straightforward as prior literature suggests, nor as contemporary researchers predict.

In our experiments, we report how we determined sample sizes and all conditions. All sample sizes were determined a priori by running power analyses (most at 80% power). Where possible, we based effect size estimates on prior experiments. Where that was not possible, we conservatively assumed small effect sizes, which led us to employ larger sample sizes. Pre-registrations (including all exclusions), materials, and data are posted as a supplement online at the Open Science Framework: https://osf.io/b4mk5/. We pre-registered analyses (and exclusions) for Experiments 1B, 1C, 1D, 2, 3, and 4. We ran Experiment 1A before pre-registration became our standard practice.

2. Experiments 1A, 1B, and 1C

2.1. Benchmarking reliance on algorithmic advice against reliance on human advice

Experiments 1A, 1B, and 1C test the extent to which people are willing to adjust their estimates in response to identical advice, depending on whether it is labeled as coming from a human versus an algorithmic advisor. To generate advice, we use one of the simplest algorithms: averaging multiple independent judgments. Doing so allows us to provide high quality advice and truthfully describe it as coming from either people or from an algorithm. We test how people respond to algorithmic versus human advice in different domains, starting in an objective domain, a visual estimation task. In Experiment 1A, participants guess an individual's weight from a photograph. Next, we examine a more subjective domain. In Experiment 1B, participants forecast the popularity of songs on the upcoming week's Billboard Magazine Hot 100 Music Chart. Finally, we examine the most subjective domain we could think of: person-perception. In Experiment 1C, participants play matchmaker and predict how another person would judge a potential romantic partner.

2.2. Method: Experiment 1A (visual estimation task)

The sample included 202 participants (90 women; 112 men; *Mdn* age = 28). Participants estimated the weight of a person in a photograph (see Fig. A1 in the Appendix) at two points in time: before and after receiving advice, which either came from a person or an algorithm. Accurate final responses entered participants into a raffle for a $10 bonus. Immediately after each estimate, participants indicated their confidence in that estimate: "How likely is it that your estimate is within 10 lb of the person's actual weight?" on a scale from 0 (*no chance*) to 100 (*absolutely certain*).

Everyone received the same advice (163 lb), which was actually the average estimate of 415 participants in another experiment (Moore & Klein, 2008). Importantly, this estimate was nearly perfect (164 lb was the actual weight). The advice was either labeled as coming from other people or an algorithm. The manipulation wording for this and all other experiments is listed in Table 1. After reporting their second estimate, their confidence in it, and how difficult it was to determine the person's weight, participants answered an 11-item Numeracy Scale (Schwartz, Woloshin, Black, & Welch, 1997). Higher scores (0–11) reflect a greater comfort with numbers.

We calculated Weight on Advice (WOA) by dividing the difference between the final and initial estimate produced by each participant by the difference between the advice and the initial estimate. This produces a measure of advice utilization that, in most cases, ranges from 0% (full advice discounting) to 100% (full advice adherence). Following prior research, we Winsorized any WOA values greater than 1 or less than 0.

2.2.1. Results

Participants showed an appreciation of algorithms, relying more on the same advice when they thought it came from an algorithm ($M = 0.45$, $SD = 0.37$), than when they thought it came from other people ($M = 0.30$, $SD = 0.35$), F (1, 200) = 8.86, $p = .003$, $d = 0.42$; see Fig. 1. Results hold when controlling for gender, numeracy, and confidence in the initial estimate, F (1, 197) = 9.02, $p = .003$. There are no main effects of these variables ($Fs < 0.39$, $ps > 0.52$).

Similarly, confidence increased more from Time 1 to Time 2 in the algorithmic advice condition (Time 1: $M = 71.30$, $SD = 18.11$; Time 2: $M = 79.91$, $SD = 16.93$) than the human advice condition (Time 1: $M = 70.62$, $SD = 18.15$; Time 2: $M = 75.10$, $SD = 17.81$), as reflected in the main effect of time, F(1, 200) = 77.09, $p = .001$, and interaction between source and time, F(1, 200) = 5.62, $p = .019$. Furthermore, higher numeracy correlated with greater reliance on algorithmic advice, r(1 0 0) = 0.21, $p = .037$; this however, was not the case in human advice condition, r = -0.12, $p = .225$.

Next, some factors that one might expect to affect our results do not. Specifically, perceived difficulty of the task does not correlate with reliance on advice for either condition. $ps > .37$. But does distance from the advice matter? Perhaps people inferred the quality of advice based on its proximity to their original estimate (Minson, Liberman, Ross, 2011). No; neither Time 1 estimates ($p = .55$) nor Time 1 confidence ratings ($p = .37$) correlate with reliance on advice. Lastly, age seems especially relevant to any question related to perceptions of technology. Specifically, older people may feel less familiar with algorithms, which could in turn encourage their resistance to algorithmic advice. Yet, our more senior participants relied on algorithmic advice as much as our younger participants did, r(102) = -0.05, $p = .616$.

2.2.2. Discussion

Our results suggest that people display algorithm appreciation, even despite a minimal description of the algorithm; participants relied *more* on identical advice when they thought it came from an algorithm than when they thought it came from other people. Importantly, given that the advice was nearly perfect, greater reliance on it improved judgment accuracy. Although participants underweighted advice in both conditions (they adjusted less than half-way toward advice that had been generated by averaging the estimates of many people), they discounted algorithmic advice less than they discounted advice from other people. These results suggest that one way to increase adherence to advice is to provide advice from an algorithm.

Updating to algorithmic advice without access to its equations or processes reflects people's willingness to listen to algorithms even with uncertainty about its inner workings. Our operationalization of a "black box" algorithm parallels the widespread appearance of algorithms in daily life, including Netflix, weather forecasts, population estimates, and economic projections. Presenting a black box algorithm allowed participants the use of their own default interpretations of algorithmic judgment. This begs the question, what *are* people's default interpretations of an algorithm?

2.2.3. Defining "Algorithm"

An important consideration in examining people's reactions to algorithms is identifying how participants define the construct. If individuals are uncertain about what an algorithm is, then this uncertainty might account for some of the mistrust postulated by the algorithm aversion literature. Or perhaps the typical research participant is certain but incorrect. We conceptualize an algorithm as a series of mathematical calculations. Similarly, the field of mathematics defines an algorithm as "a procedure for computing a function" (Rogers, 1987).

In order to establish whether most participants share this definition, we asked participants in Experiment 2 ($N = 149$), as well as participants from an MBA sample at a large West Coast university ($N = 77$) to define what an algorithm is. A research assistant coded participants' open-ended responses using thematic coding (Pratt, 2009). Specifically, the research assistant was instructed to create as few categories as possible without making them too general (see Table 2).

Overall, the categories based on participants' responses provide evidence for (1) a high level of consensus among participants regarding what an algorithm is (42% of responses fell into the first category) as well as (2) agreement with the expert definition. These data suggest that our participants are mostly familiar with algorithms and conceptually understand what they do.

2.3. Method Experiment 1B (song forecasting task)

Next, we test whether algorithm appreciation holds in a more subjective domain: predicting the popularity of songs. We pre-registered collecting data from 200 participants. The final sample included 215

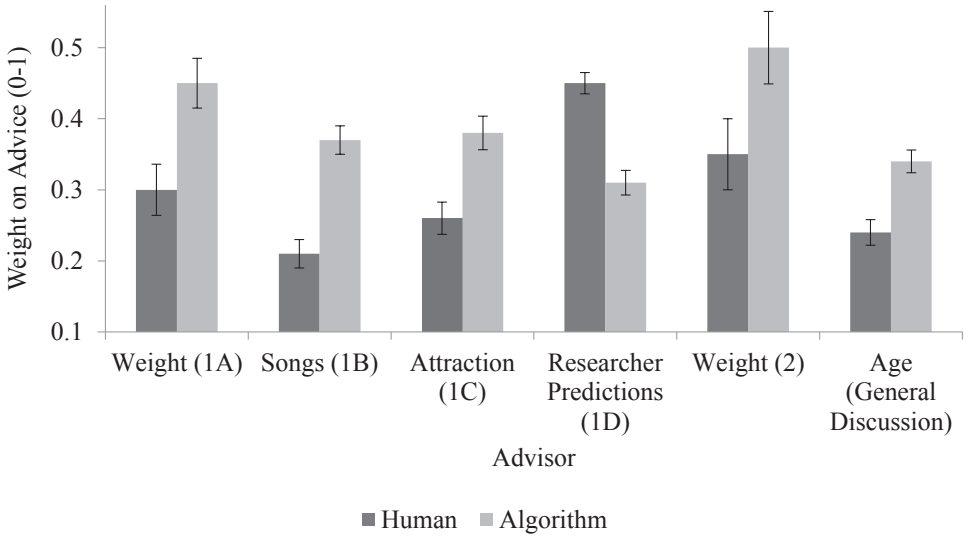


Fig. 1. Weight on Advice (WOA) as a function condition, Experiments 1A, 1B, 1C, the joint conditions of Experiment 2, and the normative experiment in the general discussion. The higher the WOA, the more participants revised their estimate toward the advice they received. For Experiment 1D, means represent researchers' predicted WOA for participants in Experiment 1C. Error bars indicate standard errors.

Table 2
Participant-generated definitions of an algorithm.

Category of Definition	Example Definition	% N = 226
Math/Equation/Calculation	"An algorithm is a set of equations to find an answer. It will spit out an answer."	42%
Step by Step Procedure	"An algorithm is a systematic way of solving problems. It looks at a problem and goes through a process to figure out the solution."	26%
Logic/Formula	"A formula that can be used to obtain a quantity or characteristic. An algorithm should be able to yield the same result for the same input exactly and consistently."	14%
Other	"A series of formulas that will generate an output with the correct inputs. Usually used on the computer."	18%

participants (114 women; 101 men; *Mdn* age = 31) as we overestimated the number we would exclude based on repeat I.P. addresses. We pre-screened participants who were 18–40 years old to ensure that they had some familiarity with current music. In this and all subsequent experiments, we pre-registered excluding participants associated with identical I.P. addresses (accepting the first instance and excluding all additional instances; see online Supplement on OSF for any other pre-registered exclusions specific to each experiment).

Participants predicted the rank of ten randomly selected songs on the Billboard Magazine's "Hot 100" which had placed on the chart in previous weeks (see Table A1 in the Appendix). For each song, participants saw a graph with that song's ranks from prior weeks and made a forecast by entering a number from 1 to 100. For example, one forecast asked:

> "What rank will 'Perfect' by Ed Sheeran place on the Billboard Magazine 'Hot 100' this week?"

They were asked to not search the internet for information. After making the first forecast, all participants received identical advice and a chance to make a second, incentivized forecast. Each exact correct final answer entered them into a raffle for $10. Participants were randomly assigned to receive advice described as either aggregated from 275 other participants' estimates or from an algorithm (see Table 1). For each song, advice was an average of forecasts from 275 past participants. We calculated WOA as we did in Experiment 1A.

2.3.1. Results

Consistent with our pre-registered analysis plan, we treated the ten forecasts as repeated measures. We regressed WOA on source of advice (−1: Algorithm; +1: Human) using an OLS regression and clustering observations at the participant level. Again, participants relied more on identical advice when they thought it came from an algorithm than from other people, $\beta = -0.34$, $t(2\,1\,4) = 5.39$, $p < .001$. Results hold analyzing an averaged WOA measure for each participant (Algorithm: $M = 0.37$, $SD = 0.02$; Other People: $M = 0.21$, $SD = 0.02$), $t(2\,1\,3) = 5.54$, $p < .001$; see Fig. 1.[3]

2.3.2. Discussion

Experiment 1B speaks to the generalizability of algorithm appreciation by testing the phenomenon using a forecasting task (rather than an estimation task) with a randomly selected set of stimuli. Participants relied more on identical advice when they thought it came from an algorithm than when they thought it came from other people, even in a highly familiar and taste-based domain, predicting the popularity of songs. We examine algorithm appreciation in Experiment 1C within an even more subjective domain: predicting romantic attraction.

2.4. Method: Experiment 1C (romantic attraction forecasting task)

The final sample included 286 participants (157 women; 129 men; *Mdn* age = 37). They were asked to play matchmaker. They were randomly assigned to read identical descriptions about a heterosexual woman, Julia, or man, Mike. These descriptions only differed in their use of pronouns (see Supplement). Then, participants saw one photograph of a different person, a man (woman), and predicted how Julia (Mike) would evaluate the target using scales from 1 (*not at all*) to 100 (*extremely*):

- How attractive do you think Julia (Mike) would find this (wo)man?

[3] For the figure, to create an appropriate comparison with Experiments 1A, 1C, and 1D, we use the means and standard errors obtained by averaging all observations produced by one participant into a single data point.

- How funny do you think Julia (Mike) would find this (wo)man's sense of humor?
- How much do you think Julia (Mike) would enjoy a dinner date with this (wo)man?

On the following page, participants received advice about each of the three forecasts and made their final forecasts. Again, participants received identical advice but were randomly assigned to read that the advice came from either 48 people from another study or an algorithm (see Table 1). All advice was an average of forecasts from 48 past participants. These participants evaluated the attractiveness, humor, and enjoyableness respectively for Julia's potential date as: 54, 59, and 58; and Mike's potential date as: 64, 57, 66.

2.4.1. Results

As pre-registered, we calculated WOA for attractiveness, humor, and enjoyableness forecasts and averaged them into a single index ($\alpha = 0.77$). Again, participants relied more on identical advice when they thought it came from an algorithm ($M = 0.38$, $SD = 0.28$) than when they thought it came from other people ($M = 0.26$, $SD = 0.27$), $t(2\,8\,4) = 3.50$, $p = .001$, $d = 0.44$; see Fig. 1. Algorithm appreciation is robust to whether people made predictions for Mike or Julia, as evidenced by a non-significant interaction between source of advice (algorithm vs. people) and target of evaluation (Julia vs. Mike's date), $F(1, 282) = 0.05$, $p = .818$. There was a main effect of source, $F(1, 282) = 13.48$, $p < .001$, but not of target, $F(1, 282) = 2.32$, $p = .129$.[4]

2.4.2. Discussion

Even when predicting interpersonal attraction, a domain where individual preferences (and emotions) rule, participants relied more on identical advice when they thought it came from an algorithm than when they thought it came from other people. As in Experiment 1B, Experiment 1C merely manipulated the label of the advice, ensuring that the algorithm was described as simply as was the advice from people ("algorithm" vs. "48 people estimated"). Next, we test whether these results are counterintuitive to researchers. Do active researchers successfully predict how our participants responded to algorithmic advice or do they expect our participants to display algorithm aversion?

2.5. Method: Experiment 1D (researchers' predictions)

In Experiment 1D, we recruited academic researchers to predict the results of Experiment 1C. Participants, whom we will call "researchers," followed a link circulated to the Society for Judgment and Decision Making email distribution list and via personal requests. As pre-registered, the final sample was collected over a period of two weeks. It included 119 participants (55 women; 64 men; *Mdn* age = 32). We excluded one participant from the 120 completed surveys who provided an answer outside of the range of 0 to 100%.[5] More accurate predictions increased participants' chances to win a $100 bonus.

The researchers predicted the mTurk participants' responses from Experiment 1C after viewing those survey materials. Researchers saw materials from *both* the advisor conditions for *either* Mike or Julia. They read that the descriptions of Julia and Mike were identical but for the names and pronouns. Researchers next read how the WOA measure is calculated, how to interpret the different values, and how participants generally respond to advice within the paradigm based on prior research (see Supplement). Then, they predicted how mTurk participants weighted the advice they received. These consisted of six incentivized predictions: WOA for each of the attraction, humor, and enjoyableness predictions for both the human and algorithm advisors. Specifically, participants were asked to enter a number between 0 to 100 for each prediction (to provide a prediction of 0% to 100%). Lastly, they answered questions about their own academic background.

2.5.1. Results

We averaged researchers' predictions of WOA across attractiveness, humor, and enjoyableness into a "predicted people WOA" index ($\alpha = 0.92$) and a "predicted algorithm WOA" index ($\alpha = 0.91$). Then we subtracted the *predicted algorithm WOA index* from the *predicted people WOA index*. Here, a positive value represents a prediction of algorithm aversion and a negative one represents a prediction of algorithm appreciation.

Even though mTurk participants in Experiment 1C displayed algorithm appreciation (-0.11), the researchers predicted algorithm aversion ($M = 0.14$, $SD = 0.20$), one-sample $t(1\,1\,8) = 14.03$, $p < .001$, $d = 1.25$. These results not only reflect a magnitude difference between the researchers' predictions and mTurk participants' actual responses but also a *directional* difference. Indeed, the algorithm appreciation exhibited by participants in experiment 1C was significantly different from equal utilization of human and algorithmic advice, $t(2\,8\,4) = 3.50$, $p = .001$, $d = 0.44$, as was the algorithm aversion predicted by researchers in experiment 1D, $t(1\,1\,8) = 7.92$, $p < .001$, $d = 0.80$; see Fig. 1.

2.5.2. Discussion

Although mTurk participants displayed algorithm appreciation when predicting romantic attraction, researchers expected them to display algorithm aversion. The idea that people are averse to algorithmic advice is evidently pervasive. Of the researchers in our sample, 34% identified themselves as graduate students. Researchers, both junior and senior, did not give people the credit they deserve in their willingness to rely on algorithmic advice (graduate students: $M = 0.16$, $SD = 0.17$; senior researchers: $M = 0.14$, $SD = 0.21$), $t(1\,1\,7) = 0.52$, $p = .607$.[6] But how robust is algorithm appreciation? Perhaps our experiments' presentation of advice (in a between-subjects design) led people to rely on algorithmic advice more than they might otherwise when forced to make a choice between one advisor directly compared with the other.

3. Experiment 2: Joint versus separate evaluation

Experiments 1A, 1B, and 1C demonstrated that in separate evaluation of advisors, our between-subjects manipulation of the source affected advice utilization. Yet, prior work finding algorithm aversion asked participants to choose between sources (joint evaluation). As attributes are easier to evaluate in a joint evaluation, due to increased information (Bazerman, Loewenstein, & White, 1992; Bazerman, Moore, Tenbrunsel, Wade-Benzoni, & Blount, 1999; Hsee, 1996; Hsee, Loewenstein, Blount, & Bazerman, 1999), perhaps providing participants with a counterfactual to the algorithmic advice decreases reliance on it. Experiment 2 examines whether joint versus separate presentation of advisors determines when people display algorithm appreciation or algorithm aversion.

3.1. Method

3.1.1. Participants

We set out to collect data from 150 participants to detect a medium sized effect (Cohen's $d = 0.4$) at 80% power. The final sample included

[4] When we enter gender of the participant as an additional factor, the effect of source remains significant, $F(1, 278) = 12.81$, $p < .001$, and there are no other significant main effects nor interactions, $ps > 0.093$.

[5] Not pre-registered.

[6] Not pre-registered. We had pre-registered that if we failed to find an effect, we would exclude graduate students to test for the effect again. However, after finding an effect, we wanted to compare the samples to make sure they did not differ.

154 participants (104 women; 50 men; *Mdn* age = 21) from a West Coast U. S. university's credit and paid subject pools.

3.1.2. Design

The experiment followed the design of Experiment 1A but additionally included a third condition where participants chose directly between an algorithmic and human advisor. Thus, it featured a 3-cell (person only vs. algorithm only vs. choice between a person and algorithm) between-subjects design that manipulated the presentation of the advisors.

3.1.3. Procedure and materials

After participants estimated the weight of the man in the same photograph as Experiment 1A, they learned that they were about to receive advice. In each of the separate conditions, participants read simple descriptions of their advisors, *either* another participant, or an algorithm, and then received the advice on the next page. In the joint condition, participants read the same descriptions of *both* advisors: "you will choose whether you will see advice regarding the same person's weight from another study participant or from an algorithm" and chose their advisor.

On the next page, participants received advice either based on their randomly assigned advisor (separate conditions) or on their choice (joint condition) (see Table 1). Again, we measured WOA. In the choice condition, we additionally measured whether participants chose to receive advice from the person or algorithm, which was our main variable of interest.

3.2. Results

As in our prior experiments, participants who evaluated advisors separately relied more on the advice when they thought it came from an algorithm ($M = 0.50$, $SD = 0.37$) than from another participant ($M = 0.35$, $SD = 0.36$), $t(1\,0\,0) = 2.10$, $p = .038$, $d = 0.44$; see Fig. 1. However, evaluating the two advisors jointly did not reverse the preference for algorithmic advice; 75% of participants in the joint condition also preferred the algorithm ($N = 39$) over the other participant ($N = 13$).[7]

3.3. Discussion

Reliance on algorithmic advice appears robust to different presentations of advisors. The results speak to the strength of algorithm appreciation. This durability might be impressive, given that many decisions are affected by joint-versus-separate evaluation: willingness to pay for consumer goods, willingness to pay for environmental issues, support for social issues, and voter preferences (Hsee, 1996, 1998; Irwin, Slovic, Lichtenstein, & McClelland, 1993; Nowlis & Simonson, 1997).

Thus far, our experiments have intentionally controlled for excessive certainty in one's own knowledge by providing advice from external advisors in both the human and algorithm conditions. Doing so ensures that participants compare their own judgment with the advice from both other people as well as advice from an algorithm. Yet, work that finds algorithm aversion tends to ask participants to choose between their own judgment and the algorithm's advice. Indeed, work on advice-taking shows that the role of the self influences willingness to use advice from people (Soll & Mannes, 2011). These differences might explain why we find different results from past work. In Experiment 3, we examine whether unwarranted confidence in one's own judgment moderates the use of algorithmic judgment.

[7] Not surprisingly, participants relied similarly on the advisor they chose, be it the algorithm ($M = 0.52$, $SD = 0.37$) or person ($M = 0.41$, $SD = 0.36$), $t(50) = 0.90$, $p = .371$.

4. Experiment 3: The role of the self

In the introduction, we alluded to the importance of distinguishing between the extent to which individuals might disregard *algorithmic* advice from the extent to which they might disregard advice *in general*, with a preference for their own judgment. Addressing this question requires a comparison of how advice is utilized both when it comes from human versus algorithmic sources, as well as when the human judgment is produced by the participant him or herself. Such a design necessitates that we abandon the WOA measure (because it cannot capture advice to oneself). Instead, we measure choice between human versus algorithmic judgment. Importantly, we manipulate whether the human judgment either comes from the self or another person.

We used the materials from Dietvorst et al. (2015), specifically Experiment 3A. Here, people chose between their *own estimate* and an algorithm's. The key comparison in the Dietvorst et al. paper was between people who were not exposed to performance information versus people who saw the (imperfect) performance of the human, the algorithm, or both. In our replication, we used the control condition from Dietvorst et al. (wherein people chose to use their own judgment versus an algorithm's, without information about prior performance). We added a new condition; participants chose between *another person's* answer and an algorithm's. This method allowed us to manipulate the source of human judgment: the participant's *own estimate* versus *another person's estimate*.

4.1. Method

We aimed to collect a sample of 400 participants, following a power analysis based on Dietvorst et al. (2015). The final sample included 403 participants (177 women; 226 men; *Mdn* age = 32). The more accurate participants' answers, the greater their bonus payment (from $0.10 for an answer within 6 ranks of the truth, increasing in 15 cent increments for each closer rank, to $1.00 for a correct answer). The experiment employed a 2-cell (self vs. other) between-subjects design. Participants either chose between their own estimate and an algorithm's estimate (self condition), or between another person's estimateand an algorithm's (other condition; see Table 3). They made this choice prior to making any estimate of their own and even prior to seeing the actual information that would inform their estimate.

Participants began the experiment by reading about the upcoming task: estimating the rank of one U.S. state from 1 to 50 in terms of the number of airline passengers who departed from the state in 2011. They learned that they would see the name of one U.S. state and then give it a rank. A rank of 1 indicated the most departing passengers, and a rank of 50 indicated the fewest departing passengers. Then, they read an overview of the information which all judges would receive in order to make the estimate. The information would include background information specific to that state, such as the number of major airports, median household income for 2008, etc. (see Supplement).

Participants either read that in addition to background information about the state, they would receive an estimate from an algorithm (self condition) or that they would receive estimates from an algorithm or *another participant* (other condition). All participants read a description of the algorithm:

> "…statistical model developed by experienced transportation analysts. …The model does not have any additional information that you will not receive. This is a sophisticated model, put together by thoughtful analysts."

Prior to making their own estimate, participants chose how they wanted to determine their bonus pay, which was based on accuracy. Specifically, participants in both conditions chose whether they wanted their bonus to be determined by an estimate produced by a human or by an estimate produced by an algorithm. As the conditions differed in whether the human estimate came from *the participant* him or herself or

Table 3
Design of Experiment 3: Participants chose to base their bonus on a human or algorithm estimate. The human was either the self or another participant.

Self Condition	Other Condition
Choose: Algorithm's Estimate vs. Own Estimate	Choose: Algorithm's Estimate vs. Other Participant's Estimate

another participant, this design allowed us to examine the influence of the self in adherence to algorithmic advice. Finally, participants reported their confidence in both human and algorithmic estimates prior to making their own estimate or seeing what the advisor(s) estimated.

4.2. Results

Overall, people preferred to base their bonus pay on algorithmic judgment rather than human judgment. The majority of participants chose to determine their bonus pay based on the algorithm's estimate rather than *another participant's* estimate (88%), χ^2 (1, N = 206) = 118.14, $p < .001$, $r = 0.76$, consistent with results from Experiments 1A, 1B, 1C, 2, and 3 in this paper. Similarly, participants even chose the algorithm's estimate over their *own* estimate (66%), $\chi 2$ (1, N = 197) = 20.15, $p < .001$, $r = 0.32$. Importantly, the preference for algorithmic judgment was attenuated by the introduction of the self as a direct comparison to the algorithmic estimate; participants chose the algorithm less frequently when they could choose their *own* estimate (rather than another participant's), $z = 6.62$, $p < .001$; see Fig. 2.

When we examined participants' confidence in the advisors' estimates, the pattern of results is consistent with both our predictions and the overconfidence literature. Although participants in both conditions were more confident in the accuracy of the algorithmic estimate ($M = 3.76$, $SD = 0.75$) than human estimate ($M = 2.69$, $SD = 0.83$; t (4 0 2) = 21.28, $p < .001$), participants were more confident in their *own* estimate ($M = 2.86$, $SD = 0.91$) than that of a *fellow participant* ($M = 2.53$, $SD = 0.71$), $t(369.93) = 4.03$, $p < .001$, correcting for unequal variances.[8] Individuals' confidence in the algorithmic advice remained consistent across conditions (confidence in algorithmic estimate when contrasted against *another participants'* judgment: $M = 3.80$, $SD = 0.72$; confidence in algorithm when contrasted against participants' *own* judgment: $M = 3.72$, $SD = 0.79$; $t(393.18) = 1.07$, $p = .286$, correcting for unequal variances).

4.3. Discussion

Experiment 3 demonstrates a reduction in algorithm appreciation when individuals are forced to choose between their own judgment and that of an algorithm. These results suggest that prior findings of algorithm aversion may have been boosted by people's excessive appreciation of their own opinions rather than comparing two types of advisors. Our results are also consistent with the robust literature on overconfidence, which has repeatedly demonstrated that individuals treat their judgment as superior to that of other people (Harvey, 1997). Consistent with the overconfidence literature, our participants were more confident in their own judgment than that of another participant. Yet, they appropriately judged the advice of a sophisticated algorithm as superior to both their own judgment and that of another person.

Our results may appear to contradict the results from Dietvorst et al. (2015), where participants' preference for algorithmic judgment decreased when they saw the algorithm err. However, a closer examination of the Dietvorst results reveals that *prior* to receiving performance data, participants in those studies either preferred algorithmic judgment to human judgment (as our participants do) or were indifferent between the two (as was the case for the experimental materials we used for our Experiment 3). Our current work focuses on people's perceptions of algorithmic advice *prior* to receiving any performance accuracy feedback, and is thus consistent with the prior results.

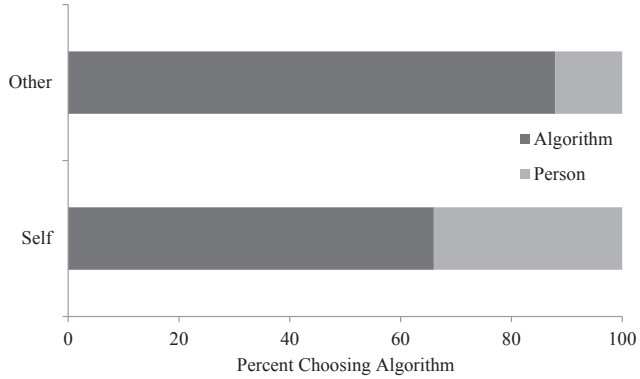


Fig. 2. Percent of participants choosing to base their bonus on the estimate of the algorithm by condition. More than 50% of participants chose the algorithm in both conditions, $ps < 0.001$.

The world outside of the laboratory is rife with situations in which we must decide whether to rely on algorithmic judgment in the absence of performance data. Forecasts in financial, geopolitical and environmental domains are often made years before accuracy data become available. And when it comes to selecting employees or dating partners, our choices rule out the possibility of observing the counterfactual of what would have happened had we made another choice. The richness of decision-making scenarios in the real world makes it doubly important to document use of advice across a variety of contexts.

5. Experiment 4: Decision-maker expertise

Our experiments thus far have focused on how elements of the judgment context's influence people's response to algorithmic versus human advice. Experiment 4 examines whether the expertise of the decision-maker influences algorithm appreciation. We recruited professionals whose work in the field of national security for the U.S. government made them experts in geopolitical forecasting and compared their advice utilization with the diverse online sample available through Amazon's Mechanical Turk. Although the two samples likely differ in many aspects beyond just expertise in forecasting, the mTurk participants' responses serves as a useful benchmark for the experts'. Both samples received the same forecasting tasks and identical advice. The advice came from forecasters in a follow-up tournament to the Good Judgment Project (GJP). The GJP was a government funded initiative aimed at improving forecasts of geopolitical events through statistical aggregation of individual judgments (Mellers et al., 2014).

5.1. Method

In selecting the sample size, we anticipated uncertainty in the number of experts who might take the survey, and thus aimed to collect a sample size of 200 mTurk workers and 75 experts, with the goal of collecting 100 experts if possible. We stopped data collection from experts when multiple days passed without new participants in order to ensure that all forecasts were made in a comparable time period. We recruited more mTurk participants than pre-registered because we overestimated the number of people we needed to exclude based on repeat I.P. addresses. The final sample included 301 mTurk participants (women = 154; men = 147; M age = 39) and 70 U.S. national security professionals (women = 3; men = 67; M age = 46) for a total of 371 participants. We recruited national security professionals with the help of a U.S. government employee who worked in national security and

[8] Not pre-registered.

shared the survey with email distribution lists dedicated to the topic of national security. Most respondents were U.S. government employees or contractors working for the government. We incentivized mTurk participants by entering them into a raffle for $10. The national security experts were entered into a raffle for an iPad. Participants who made more accurate forecasts received more entries into the raffle.

The experiment had a 2 (advisor: human vs. algorithm) × 2 (sample: lay vs. expert) design. This experiment used the same main dependent variable, WOA. for four tasks: a weight estimate, a business forecast, and two geopolitical forecasts. We included additional measures to better understand the experience of the expert sample.

We chose the tasks to amplify the difference in expertise between samples. We expected experts to feel greater expertise for the geopolitical forecasts than the lay sample. We expected both samples to feel similarly low expertise for the weight estimate and business forecasts. The weight estimate was identical to Experiment 1A. The forecasts were:

- "What is the probability that Tesla Motors will deliver more than 80,000 battery-powered electric vehicles (BEVs) to customers in the calendar year 2016?"
- "What is the probability that a North American country, the EU, or an EU member state will impose sanctions on another country in response to a cyber attack or cyber espionage before the end of 2016?"
- "What is the probability that the United Kingdom will invoke Article 50 of the Lisbon Treaty before July 1, 2017?"

As in Experiment 1A, participants made two judgments. Prior to their second estimate, all participants received human or algorithmic advice (see Table 1). For the forecasts, we provided participants with information about the forecasting tournament:

> The Good Judgment Open is a forecasting (prediction) tournament, hosted by academic researchers, where thousands of people around the world compete to make the most accurate forecasts (predictions) about global political events.

We provided high quality advice to all participants. The target in the photo weighed 164 lb and the advice was 163. Tesla did *not* sell more than 80,000 battery-powered electric vehicles (BEVs) in 2016 (it sold 76,230) and the advice was that the event was 20% likely to occur. The cyber and Brexit events were more difficult events to predict. The cyber event occurred *just* before the end of 2016, on December 29, 2016 but the advice was 12% likelihood of occurrence. The Brexit event *did* occur on March 29, 2017 but the advice was 54% likelihood of occurrence.

Following prior research (Mellers et al., 2014), we used Brier scores (Brier, 1950) to assess forecasting accuracy. A Brier score is a measure of error that ranges from 0 to 2, with higher numbers reflecting greater error. To calculate Brier scores, we squared the distance between the estimated probability and the actual result. For example, if a participant estimated a 0.8 probability of an event occurring (and thus a 0.2 probability of it not occurring), and the event occurred, we calculated: $(0.8-1)^2 + (0.2-0)^2 = 0.08$.

Participants also reported how frequently they made forecasts for their occupation and their familiarity with the word algorithm:

- "For your job, how often do you make forecasts (predictions)?" on a scale from 1 (*barely ever*) to 7 (*multiple times a day*).
- "How certain are you that you know what an algorithm is?" on a scale from 0 (*NA, I am certain that I do NOT know what it means*); 1 (*not at all certain*) to 7 (*extremely certain*).

5.2. Results

The expert sample (national security experts) reported making forecasts for their jobs more frequently ($M = 3.73$, $SD = 2.30$) than the lay sample ($M = 2.29$, $SD = 2.02$), $t(95.30) = 4.84$, $p < .001$, correcting for unequal variances. Participants overall said they were familiar with algorithms ($M = 5.04$, $SD = 1.72$), but this did not significantly differ by expertise (lay: $M = 4.96$, $SD = 1.73$, expert: $M = 5.40$, $SD = 1.63$), $t(3\,6\,9) = -1.96$, $p = .051$.

Consistent with our pre-registered analysis plan, we submitted WOA to a 2 (advisor: algorithm vs. human) × 2 (sample: lay vs. expert) repeated measures ANCOVA with the four tasks as repeated measures and familiarity with algorithms as a covariate. This analysis included those who answered all tasks (282 lay people and 61 experts). There is an effect of familiarity, $F(1, 338) = 8.86$, $p = .003$, $\eta^2 = 0.03$, $d = 0.29$, such that participants who claimed greater familiarity with algorithms took less advice, collapsing across advisors.

Controlling for familiarity, we observed a main effect of advisor, $F(1, 338) = 9.46$, $p = .002$, $\eta^2 = 0.02$, $d = 0.29$. As in earlier experiments, our participants placed more weight on algorithmic than human advice. Furthermore, experienced judges (the national security experts) took less advice than lay people, $F(1, 338) = 32.39$, $p < .001$, $\eta^2 = 0.08$, $d = 0.60$. Importantly, we also observed a significant interaction between judge expertise and the source of advice: whereas lay judges placed more weight on algorithmic than human advice, the experts heavily discounted all advice sources, $F(1, 338) = 5.05$, $p = .025$, $\eta^2 = 0.01$, $d = 0.23$; see Figs. 3 and 4.

5.2.1. Forecast accuracy

The geopolitical and business forecasts were difficult events to predict. The Tesla event did not occur, as Tesla sold merely *3770 cars less* than the number in the forecast question. Thus, the advice, a low probability of the event occurring (20%), was very accurate (Brier score: 0.08). The lower accuracy of the cyber forecast advice reflects its difficulty: the event occurred only *days* before the deadline in the forecast question, and because the advice was a low probability (12%) of the event occurring, the advice was not very accurate (Brier score: 1.55). The advice was ambiguous for the Brexit forecast (54%) but because the event actually *did* occur, the accuracy of the advice was fairly accurate (Brier: 0.42).

We submitted Brier scores to the same 2 × 2 repeated measures ANCOVA as WOA with three forecasts as repeated measures and familiarity with algorithms as a covariate. There is an effect of familiarity, $F(1, 366) = 5.71$, $p = .017$. Controlling for familiarity, there is no main effect of advisor, $F(1, 366) = 0.09$, $p = .764$, but there is an effect of sample expertise, $F(1, 366) = 9.53$, $p = .002$, and an interaction, $F(1, 366) = 4.16$, $p = .042$; see Fig. 5. As a point of reference, high performing forecasters in the GJP ("Superforecasters") reached an average Briere score of .25 in their first week of the tournament (see Table 1 in Mellers et al., 2014). Lay people and experts achieved similar accuracy when they received advice from a person. Importantly, lay people ironically achieved *greater* accuracy than experts when they received advice from an algorithm because experts heavily discounted algorithmic advice.

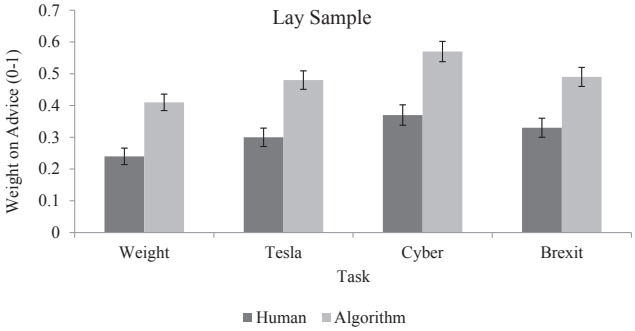


Fig. 3. Weight on Advice (WOA) as a function of advisor for the lay sample.

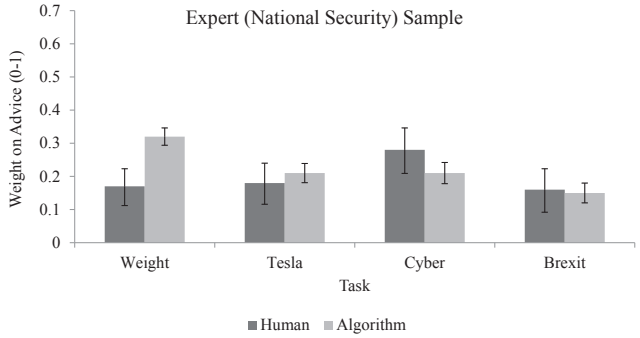


Fig. 4. Weight on Advice (WOA) as a function of advisor for the expert sample.

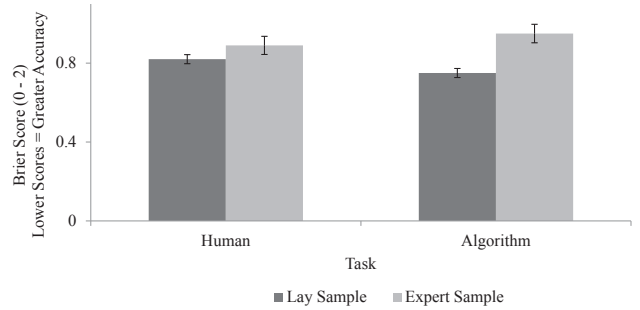


Fig. 5. Brier score as a function of advisor (another person/forecaster vs. algorithm) and sample (lay vs. expert), collapsed across forecasts, Experiment 4. The lower the Brier Score, the more accurate the forecast. A Brier Score of 0 means that the forecast was perfectly accurate. Note: $p < .001$, in the algorithm conditions between samples. Interaction: F (1, 366) = 4.16, $p = .042$.

5.3. Discussion

Unlike participants in our earlier experiments, the experts in Experiment 4 did not recognize the value of algorithmic advice. Their adherence to their prior judgments and their failure to utilize the information offered to them ultimately lowered their accuracy relative to the lay sample. Furthermore, the experts weighted advice less than the 30% WOA observed in advice-taking research (see Fig. 4). These results are also consistent with prior research demonstrating that expert attorneys are less likely than law students to give weight to advice in a verdict estimation task (Jacobson, Dobbs-Marsh, Liberman, & Minson, 2011). These results might help explain why pilots, doctors, and other experts are resistant to algorithmic advice (Meehl, 1954). Although providing advice from algorithms may increase adherence to advice for non-experts, it seems that algorithmic advice falls on deaf expert ears, with a cost to their accuracy.

6. General discussion

Counter to the widespread conclusion that people distrust algorithms, our results suggest that people readily rely on algorithmic advice. Our participants relied more on identical advice when they thought it came from an algorithm than from other people. They displayed this algorithm appreciation when making visual estimates and when predicting: geopolitical and business events, the popularity of songs, and romantic attraction. Additionally, they chose algorithmic judgment over human judgment when given the choice. They even showed a willingness to choose algorithmic advice over their own judgment.

Algorithm appreciation proved robust to a variety of elicitation methods. It also held across a variety of descriptions of algorithms (see Table 1). Additionally, people displayed algorithm appreciation regardless of their age. However, appreciation of algorithms was not uniform. It was lower among less numerate participants and, ironically, also among experts, who were simply less open to taking any advice. The accuracy of their judgments suffered as a result. Algorithm appreciation waned (but did not disappear) when algorithmic advice was pitted against the participants' own judgment. These results shed light on when people are most likely to improve their accuracy by listening to algorithms and have implications for the use of algorithms by decision makers within organizations.

6.1. Theoretical implications

Our results suggest that one simple way to increase adherence to advice is to provide advice from an algorithm. Many decisions are ripe for the use of algorithmic advising, considering the low cost and widespread availability of algorithmic advice relative to expert advice. For instance, a number of applications are now available to provide financial advice and investing guidance (Common Cents Lab, 2017). These sorts of "Robo-advisors" sometimes begin serving those who find it too costly to access human advisors. However, given the excellent performance record of algorithmic advice, there are many instances in which it is both cheaper and better.

Our results challenge the widespread assertion that people are averse to algorithms (Bazerman, 1985; Dawes, 1979; Dawes et al., 1989; Kleinmuntz, 1990; Kleinmuntz & Schkade, 1993; Meehl, 1954; Meehl, 1957). They suggest that the story of algorithm aversion is not as straightforward as the literature might otherwise lead us to believe. Importantly, our results are consistent with those of Dietvorst et al. (2015) as well as Prahl and Van Swol (2017). In those studies as well as in ours, participants were quite willing to rely on algorithmic advice before seeing the algorithm err. Although it is important to understand how people react to the performance of human and algorithmic advisors, many consequential decisions are made without the benefit of performance feedback. For instance, the accuracy of some forecasts is not available until years or even decades later; consider financial outcomes, outcomes of climate change, or political events such as conflicts between nations. It is therefore useful to examine responses to algorithmic advice prior to feedback, as we do in this paper.

Our experiments connect the research streams on perceptions of algorithmic judgment and advice taking in quantitative judgments by employing the Weight on Advice (WOA) measure. This continuous measure of advice utilization is more sensitive than the categorical measure of choice that is found in much of the prior research on human responses to algorithmic advice. In order to further compare our results with the literature on advice taking, we conducted another experiment where participants received information about how the human and algorithmic advice was produced.

Specifically, we told participants who were completing a visual estimation task that the human advice came from one person, randomly chosen from a pool of 314 people, whereas the algorithmic advice was based on estimates from 314 people (see Table 1). Presenting participants with the number of judgments that produced the advice provided the opportunity to measure how *effectively* people use algorithmic advice by comparing WOA to a normative benchmark. Doing so also allowed the accuracy of human and algorithmic advice to vary, as it often does in the real world. As in our prior experiments, participants relied more on advice from the algorithm ($M_{WOA} = 0.34$, $SD = 0.34$) than another person ($M_{WOA} = 0.24$, $SD = 0.27$), $F(1, 669) = 17.68$, $p < .001$, $d = 0.33$; see Fig. 1. But do people differentially underweight advice based on the advisor?

A sizable literature shows that to maximize accuracy, people should simply average their own guess with that of another person, yielding a WOA of 50% (Galton, 1907; Soll & Larrick, 2009; Suroweicki, 2004). However, WOA is usually too low; that is, people discount advice from others (Bonaccio & Dalal, 2006; Yaniv, 2004; Yaniv & Kleinberger, 2000). We replicate prior work, such that participants underweighted advice from the person ($M = 0.26$, $SD = 0.27$) relative to how much

they *should* have weighted advice from one other person (WOA = 0.5). By contrast, participants who received advice from an algorithm that averaged advice from 314 other people should have abandoned their own guess and weighted the algorithm's advice 100%. They failed to do so ($M = 0.66$, $SD = 0.34$), thereby underweighting algorithmic advice to an even greater extent than they underweighted human advice, $F(1, 669) = 275.08$, $p < .001$, $d = 1.30$. Even though our participants weighted algorithmic advice more heavily than advice from other people, they did not weigh it heavily enough. This result suggests that although people display algorithm appreciation, there is still room for them to improve their accuracy (for more details, see Supplement).

6.2. Practical implications

Understanding how people respond to algorithmic advice holds implications for any decision maker or organization with the potential to learn from lessons produced by "big data." As organizations invest in the collection, analysis, and exploitation of ever larger quantities of data, they use algorithms to sift through such information to produce advice. Advances in technology have improved the speed and efficiency of this process (Laney, 2012). Furthermore, many organizations have begun to make their data publically available (for example, Google (as used in Reips & Matzat, 2014) and Twitter (as used in Reips & Garaizar, 2011)). Algorithms rely on data to hone the accuracy of their advice, so continued collection of data increases the potential value of algorithmic advice in domains as diverse as movie recommendations and medical diagnosis.

6.3. Future research: Theory of machine

As humans interact more frequently with programmed agents in their workplaces, homes, and cars, we need to understand their *theory of machine*. Theory of machine, like theory of mind, requires people to consider the internal processes of another agent. Philosophical work on theory of mind considers how people infer other people's intentions and beliefs (Dennett, 1987; Saxe, 2007), In contrast, theory of machine considers people's lay perceptions of how algorithmic and human judgment differ in their *input*, *process*, and *output*. This theoretical framework can guide future research in examining the psychological mechanisms that shape how people expect human and algorithmic judgment, at their finest, to differ.

Social psychology has taught us much about how people think about others' minds. For instance, people differentiate others' behavior based on whether they perceive the behavior as intentional or unintentional (Malle & Knobe, 1997). People also use cues about experience for mind perception (Gray, Gray, & Wegner, 2007). The fundamental attribution error ascribes more credit to dispositional attributes (Lassiter, Geers, Munhall, Ploutz-Snyder, & Breitenbecher, 2002) and attitudes (Ajzen, Dalto, & Blyth, 1979; Jones & Harris, 1967) than the situation warrants. Anthropomorphizing machines, like the self-driving car, influences our trust in them (Waytz, Heafner, & Epley, 2014) and can backfire (Gray & Wegner, 2012).

Rather than testing how people impart human judgment on algorithms, this paper tests lay beliefs about how algorithmic and human advice differ. Doing so scratches the surface of *theory of machine* by testing people's responses to algorithmic output. Often, algorithmic advice is produced by "black box" algorithms, where the user is not privy to its inner workings. What matters most in these situations, and even when people *do* have some information about the algorithm, is people's lay theories about what kind of information the algorithm uses as inputs, how the information is processed, and the value of the output. Understanding how people expect algorithms and humans to differ might facilitate examination of how people respond to algorithmic compared with human advisors.

6.3.1. Algorithmic process

We find consistent evidence that participants who faced "black box" algorithms in our experiments were willing to rely on that advice despite its mysterious origins. We are comforted that the participant definitions of algorithms were similar to our conceptualization. Nevertheless, it is worth asking how appreciation of algorithmic advice is affected by *transparency* of the algorithm's operations as well as people's more sophisticated understanding of them.

Given the excellent track record of so many algorithms, it might stand to reason that more knowledge about an algorithm's process would increase algorithm appreciation. On the other hand, as algorithmic operations become more complex, they also become inscrutable. Especially for the innumerate, learning about linear models that minimize mean squared error without overfitting historical data might raise more concerns and doubts than it assuages. It is easy to imagine that some information could actually reduce users' appreciation of the algorithm.

6.3.2. Expertise process

The experiments we present compare participants' reliance on advice from algorithms with advice from other people who are similar to them (e.g., 275 other participants in Experiment 1B and 48 people in another study in Experiment 1C). That comparison allowed us to make the contest between algorithms and people a fair fight to control for a number of potential differences between different sources of advice. It also allowed us to avoid the use of deception by actually presenting advice derived from averaging estimates of past participants. However, it is possible that people prefer advice from human experts than from others like them (see Supplement), and that this preference is stronger than any appreciation of algorithms (Önkal, Goodwin, Thomson, Gönül, & Pollock, 2009; Promberger & Baron, 2006). Future work will have to investigate how people's beliefs about others' expertise affect reliance on their advice. Any such research program will need to address the question of what exactly people believe about expertise and how it improves the value of advice (a *theory of expertise*). The operations of expert minds will remain an inscrutable "black box" to those who receive advice from them. Thus, it may prove useful to compare a *theory of expertise* with a *theory of machine*.

6.3.3. Room for increased reliance on algorithms

Our results provide a more nuanced picture of individuals' willingness to take algorithmic advice. On the one hand, participants adjusted their judgments more in light of algorithmic versus human input. On the other hand, and consistent with prior work on advice taking, individuals adjusted their judgments too little, and seemed largely insensitive to the quality of advice that was offered to them. While an influential paper shows that people are more willing to use an *imperfect* algorithm if they are able to adjust its output (Dietvorst, Simmons, & Massey, 2016), future work could also examine how algorithmic advice can be presented in a way that maximizes its uptake prior to any performance information. It could also test if people expect algorithms and people to differ in how quickly they learn from mistakes.

7. Conclusion

Technological advances make it possible for us to collect and utilize vast amounts of data like never before. "Big data" is changing the way organizations function and communicate, both internally and externally. Organizations use algorithms to hire and fire people (Ritchel, 2013), manage employees' priorities (Copeland & Hope, 2016), and help employees choose health care plans (Silverman, 2015). Companies such as Stitch Fix use algorithms to provide clients with clothing recommendations (Ahuja, 2015; Hu, 2014).

Information technologies may increase the prevalence of "big data"

initiatives but how do these initiatives and the algorithmic advice gleaned from them change how people see the world? Organizations have an opportunity to learn from the ever-increasing amount of information they can access. Yet, if they only focus on collecting and analyzing data and overlook how people respond to the algorithmic advice produced by it, they will not fully maximize the benefits of algorithmic advice.

Without understanding how people incorporate information from algorithms into their decisions, organizations run the risk of misusing the opportunities presented by technological advances. On the one hand, our results suggest optimism about the potential for usage of algorithmic advice. Indeed, we are deeply hopeful about the potential for humans to use algorithmic advice to help them make wiser decisions. At the same time, our results raise questions about the willing reliance on technological guidance. In an age when we routinely allow technological systems to direct our attention and our spending, when we rarely bother to understand those systems, let alone read the terms and conditions, we all ought to be concerned about our vulnerability to manipulation. We hope our work stimulates future research to explore the consequences of the human willingness to rely on algorithms and their broader theory of machine. Doing so can help decision makers increase their awareness of being potentially manipulated and also help them better extract the knowledge that is available in order to thrive in the world of "big data."

Acknowledgements

The Authors wish to thank Leif D. Nelson, Cameron Anderson, Michael A. Ranney for their helpful comments and insights. Thanks also to Clayton Critcher, Linda Babcock, and Jennifer Lerner as well as the seminars in the Negotiation, Organization & Markets unit at Harvard Business School, the Operations, Information & Decisions department at the Wharton School of the University of Pennsylvania, the Social & Decision Sciences group at Carnegie Mellon University, the Public Policy department at London School of Economics, the behavioral seminar at the University of California, Los Angeles, Anderson School of Management, the Affective Brain Lab at University College London, the Scalable Cooperation Group at the Massachusetts Institute of Technology's Media Lab, the Program on Negotiation seminar at Harvard University, the Management department at the University of Miami, Miami Business School, and the Economics and Strategic Management department at the University of California, Rady School of Business for their thoughtful feedback and discussions.

Thank you to Jeff Hannon for his generosity in recruiting participants who work in national security for the U.S. Government and to Terry Murray and Nick Rohrbaugh for sharing forecasting questions from the Good Judgment Open. Thanks to Berkeley Dietvorst for generously sharing experimental materials. Thanks to the Intelligence Advanced Research Projects Activity (IARPA), UC Berkeley Haas Dissertation Fellowship, and the Behavioral Lab at Haas for their generous financial support. Thanks to Isaac Weinberg, Julia Prims, Marija Jevtic, Noor Ul Ann, and Sue Wang for their assistance in coding.

Appendix A

.
.



Fig. A1. The photograph viewed by participants in Experiment 1A.

Table A1
Participants forecasted the rank of 10 songs on the "Hot 100" in Experiment 1B.

Song	Artist	Advice for Song Rank
"Love Lies"	Khalid & Normani	56
"Never Be The Same"	Camila Cabello	16
"I Fall Apart"	Post Malone	25
"Most People Are Good"	Luke Bryan	48
"Bad At Love"	Halsey	30
"Outside Today"	YoungBoy Never Broke Again	37
"Broken Halos"	Chris Stapleton	46
"Heaven"	Kane Brown	35
"Perfect"	Ed Sheeran	5
"Dura"	Daddy Yankee	52

Appendix B. Supplementary material

Supplementary data, materials, and pre-registrations to this article can be found online at the Open Science Framework here: https://osf.io/b4mk5/

References

Ahuja, S. (2015). What Stitch Fix figured out about mass customization. *Harvard Business Review*. https://hbr.org/.

Ajzen, I., Dalto, C. A., & Blyth, D. P. (1979). Consistency and bias in the attribution of attitudes. *Journal of Personality and Social Psychology, 37*(10), 1871.

Bazerman, M. H. (1985). Norms of distributive justice in interest arbitration. *Industrial and Labor Relations Review, 38*, 558–570. https://doi.org/10.2307/2523991.

Bazerman, M. H., Loewenstein, G. F., & White, S. B. (1992). Reversals of preference in allocation decisions: Judging an alternative versus choosing among alternatives. *Administrative Science Quarterly,* 220–240.

Bazerman, M. H., Moore, D. A., Tenbrunsel, A. E., Wade-Benzoni, K. A., & Blount, S. (1999). Explaining how preferences change across joint versus separate evaluation. *Journal of Economic Behavior and Organization, 39*, 41–58.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes, 101*(2), 127–151.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*, 1–3.

Common Cents Lab (2017, August 18). Retrieved from http://advanced-hindsight.com/commoncents-lab/.

Copeland, R., & Hope, B. (2016). The World's Largest hedge fund is building an algorithmic model from its employees' brains. *The Wall Street Journal*. https://www.wsj.com/.

Dana, J., & Thomas, R. (2006). In defense of clinical judgment...and mechanical prediction. *Journal of Behavioral Decision Making, 19*(5), 413–428.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*(7), 571. https://doi.org/10.1037/0003-066x.34.7.571.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*(2), 95.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573.

Dennett, D. (1987). *The intentional stance.* Cambridge: MIT Press.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2016). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*.

Dijkstra, J. J. (1999). User agreement with incorrect expert system advice. *Behaviour & Information Technology, 18*(6), 399–411. https://doi.org/10.1080/014492999118832.

Dijkstra, J. J., Liebrand, W. B., & Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour & Information Technology, 17*(3), 155–163. https://doi.org/10.1080/014492998119526.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 44*(1), 79–94. https://doi.org/10.1518/0018720024494856.

Einhorn, & Hogarth (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance, 13*(2), 171–192.

Frick, W. (2015). Here's Why People trust human judgment over algorithms. *Harvard Business Review*. https://hbr.org/.

Galton, F. (1907). Vox populi. *Nature, 75*, 450–451.

Gino, F. (2008). Do we listen to advice just because we paid for it? The impact of advice cost on its use. *Organizational Behavior and Human Decision Processes, 107*(2), 234–245.

Gino, F., & Moore, D. A. (2007). Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making, 20*(1), 21–35. https://doi.org/10.1002/bdm.539.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812) 619-619.

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, 125*(1), 125–130.

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*(1), 19.

Haak, T. (2017). Algorithm aversion (HR trends, 2017). Retrieved March 22, 2017, from https://hrtrendinstitute.com/2017/02/13/algorithm-aversion-hr-trends-2017-5/.

Harrell, E. (2016). Managers shouldn't fear algorithm-based decision making. *Harvard Business Review*.

Harvey, N. (1997). Confidence in judgment. *Trends in Cognitive Sciences, 1*(2), 78–82.

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes, 70*, 117–133.

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making.* Thousand Oaks, Calif.: Sage.

Hu, E. (2014). Try This On For Size: Personal Styling That Comes In The Mail [Audio file]. *National Public Radio: All Tech Considered.*.

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes, 67*(3).

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin, 125*(5), 576.

Irwin, J. R., Slovic, P., Lichtenstein, S., & McClelland, G. H. (1993). Preference reversals and the measurement of environmental values. *Journal of Risk and Uncertainty, 6*(1), 5–18.

Jacobson, J., Dobbs-Marsh, J., Liberman, V., & Minson, J. A. (2011). Predicting civil jury verdicts: How attorneys use (and misuse) a second opinion. *Journal of Empirical Legal Studies, 8*(s1), 99–119.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology, 3*(1), 1–24. https://doi.org/10.1016/0022-1031(67)90034-0.

Kahneman, D. (2013). *Thinking, fast and slow* (1st pbk. ed.). New York: Farrar, Straus and Giroux.

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. *Psychological Bulletin, 107*(3), 296–310. https://doi.org/10.1037/0033-2909.107.3.296.

Kleinmuntz, D. N., & Schkade, D. A. (1993). Information displays and decision processes. *Psychological Science, 4*(4), 221–227. https://doi.org/10.1111/j.1467-9280.1993.tb00265.x.

Laney, D. (2012). The importance of 'Big Data': A definition. *Gartner*.

Lassiter, G. D., Geers, A. L., Munhall, P. J., Ploutz-Snyder, R. J., & Breitenbecher, D. L. (2002). Illusory causation: Why it occurs. *Psychological Science, 13*(4), 299–305.

Liberman, V., Minson, J. A., Bryan, C. J., & Ross, L. (2012). Naïve realism and capturing the "wisdom of dyads". *Journal of Experimental Social Psychology, 48*(2), 507–512.

Logg, J. M., Haran, U., & Moore, D. A. (2018). Is Overconfidence a Motivated Bias? Experimental Evidence. *Journal of Experimental Psychology: General, 147*(10), 1445–1465.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology, 33*(2), 101–121.

Mannes, A. E. (2009). Are we wise about the wisdom of crowds? The use of group judgments in belief revision. *Management Science, 55*(8), 1267–1279.

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

Meehl, P. E. (1957). When shall we use our heads instead of the formula? *Journal of Counseling Psychology, 4*(4), 268–273. https://doi.org/10.1037/h0047554.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., ... Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological science, 25*(5), 1106–1115. https://doi.org/10.1177/0956797614524255.

Minson, J. A., Liberman, V., & Ross, L. (2011). Two to tango: Effects of collaboration and disagreement on dyadic judgment. *Personality and Social Psychology Bulletin, 37*(10), 1325–1338.

Minson, J. A., & Mueller, J. S. (2012). The cost of collaboration: Why joint decision making exacerbates rejection of outside information. *Psychological Science, 23*(3), 219–224.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*(2), 502.

Moore, D. A., & Klein, W. M. (2008). Use of absolute and comparative performance feedback in absolute and comparative judgments and decisions. *Organizational Behavior and Human Decision Processes, 107*(1), 60–74.

Moore, D. A., Tenney, E. R., & Haran, U. (2015). Overprecision in judgment. *The Wiley Blackwell Handbook of Judgment and Decision Making,* 182–209.

Nowlis, S. M., & Simonson, I. (1997). Attribute-task compatibility as a determinant of consumer preference reversals. *Journal of Marketing Research,* 205–218.

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*(4), 390–409.

Petropoulos, F., Fildes, R., & Goodwin, P. (2016). Do "big losses" in judgmental adjustments to statistical forecasts affect experts' behaviour? *European Journal of Operational Research, 249*(3), 842–852.

Prahl, A., & Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting, 36*(6), 691–702.

Pratt, M. G. (2009). From the editors: For the lack of a boilerplate: Tips on writing up (and reviewing) qualitative research. *Academy of Management Journal, 52*(5), 856–862.

Promberger, M., & Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making, 19*, 455–468.

Reips, U. D., & Garaizar, P. (2011). Mining twitter: A source for psychological wisdom of the crowds. *Behavior Research Methods, 43*(3), 635–642. https://doi.org/10.3758/s13428-011-0116-6.

Reips, U. D., & Matzat, U. (2014). Mining "Big Data" using big data services. *International Journal of Internet Science, 9*(1), 1–8. http://www.ijis.net/.

Ritchel, M. (2013). How big data is playing recruiter for specialized workers. *New York Times*. http://www.nytimes.com.

Rogers, H. (1987). *Theory of recursive functions and effective computability, Vol. 5*. New York: McGraw-Hill.

Saxe, R. (2007). Theory of mind. *The Oxford Handbook of Cognitive Neuroscience*.

Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine, 127*(11), 966–972. https://doi.org/10.7326/0003-4819-127-11-199712010-0000.

Silverman, R. (2015). Picking a health plan? An Algorithm could help. *New York Times*. http://www.nytimes.com/.

Sinha, R. R., & Swearingen, K. (2001). Comparing Recommendations Made by Online Systems and Friends. In DELOS workshop: Personalisation and recommender systems in digital libraries (p. 106).

Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision making. *Organizational Behavior and Human Decision Processes, 62*(2), 159–174. https://doi.org/10.1006/obhd.1995.1040.

Sniezek, J. A., Schrah, G. E., & Dalal, R. S. (2004). Improving judgement with prepaid expert advice. *Journal of Behavioral Decision Making, 17*(3), 173–190.

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(3), 780–805. https://doi.org/10.1037/a0015145.

Soll, J. B., & Mannes, A. E. (2011). Judgmental aggregation strategies depend on whether the self is involved. *International Journal of Forecasting, 27*(1), 81–102.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science, 333*(6043), 776–778.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations.* New York: Doubleday.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52*, 113–117.

Wegner, D. M., & Ward, A. F. (2013). How Google is changing your brain. *Scientific American, 309*(6), 58–61.

Woolley, K., & Risen, J. L. (2018). Closing your eyes to follow your heart: Avoiding information to protect a strong intuitive preference. *Journal of personality and social psychology, 114*(2), 230.

Yaniv, I. (2004). The benefit of additional opinions. *Current Directions in Psychological Science, 13*(2), 75–78.

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes, 83*(2), 260–281.

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. Conditional acceptance. Sense of recommendations. *Journal of Behavioral Decision Making* (unpublished data).