Theory of Machine: When Do People Rely on Algorithms?

HARVARD | BUSINESS | SCHOOL

Theory of Machine: When Do People Rely on Algorithms?

Jennifer M. Logg

Working Paper 17-086

Theory of Machine: When Do People Rely on Algorithms?

Jennifer M. Logg
Harvard Business School

Working Paper 17-086

Theory of Machine: When do people rely on algorithms?

Jennifer M. Logg
Harvard University

3.13.2017

Under Review: Please do not circulate without author's permission.

Author Note

Jennifer M. Logg, Harvard Business School, Harvard University

The Author wishes to thank Don A. Moore, Leif D. Nelson, Cameron Anderson, and Michael A. Ranney for their helpful comments and insights. Thanks also to Clayton Critcher, Linda Babcock, and Rick Larrick 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, and the behavioral seminar at the University of California, Los Angeles, Anderson School of Management for their thoughtful feedback and discussions. Thank you to Jeff Hannon for his generosity in recruiting participants who work in National Security. 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 and Julia Prims for their research assistance.

Correspondence concerning this article should be addressed to Jennifer M. Logg, Harvard Business School, Harvard University, Baker Library, Bloomberg Center 433, Harvard Business School, Boston, MA 02163. E-mail: jlogg@hbs.edu

Abstract

Algorithms--scripts for mathematical calculations--are powerful. Even though algorithms often outperform human judgment, people resist allowing a numerical formula to make decisions for them (Dawes, 1979). Nevertheless, people increasingly depend on algorithms to inform their decisions. Eight experiments examined trust in algorithms. Experiments 1A and 1B found that advice influenced participants *more* when they thought it came from an algorithm than when they thought it came from other people. This effect was robust to presenting the advisor jointly or separately (Experiment 2). Experiment 3 tested a moderator; excessive confidence in one's own knowledge attenuated reliance on algorithms. These tests are important because participants can improve their accuracy by relying more on algorithms (Experiment 4). Experiments 5 and 6 tested a mechanism for reliance: subjectivity of the decision. For objective decisions, participants preferred algorithmic advice and for subjective decisions, participants preferred advice from people. Experiment 6 tested the interaction of subjectivity and the availability of expert advice. Participants preferred an expert to an algorithm, regardless of the domain (Experiment 6). Experiment 7 examined how decision makers' own expertise influenced reliance on algorithms. Experts in national security, who regularly make forecasts, relied less on algorithmic advice than lay people did. 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 technological algorithms.

Keywords: algorithms, accuracy, decision-making, advice-taking, forecasting

Theory of Machine: When do people rely on algorithms?

Algorithms--scripts for sequences of mathematical calculations or procedural steps--are powerful. They can complement human judgment and are increasingly used to inform it. Professionals routinely extract information from the Internet using Google Search, perhaps while listening to music recommendations from Pandora. Algorithms help people make decisions from the mundane to the consequential. They help us travel from point A to point B with Google Maps while using Global Positioning System (GPS). More organizations are using start-ups like Gild to recruit programmers (Richtel, 2013; Miller, 2015). Gild's algorithms predict programmers' work performance based on clues scraped from the Internet including their reputation amongst and communication with peers and whether others use their code. Organizations are also incorporating algorithms into daily routines: police departments increasingly use algorithms to predict areas where crime is likely in order to determine where to patrol (Lynch, 2016).

Although organizations have traditionally relied on humans to forecast future events and outcomes (Hartford, 2015), a long literature documents the superiority of algorithms over unaided human intuition. When provided the same information, algorithms outperformed expert forecasts of: survival of cancer patients (Einhorn, 1972), severity of pathologies (Goldman et al., 1977), heart attacks (Hedén, Öhlin, Rittner, & Edenbrandt, 1997), recidivism of parolees (Carroll, Wiener, Coates, Galegher, & Alibrio, 1982), magnitude of operational risk (Tazelaar & Snijders, 2013), and answers to trivia questions (Tesauro, Gondek, Lenchner, Fan, & Prager, 2013). In these instances, algorithms more appropriately weighted the same informational cues (Dawes, 1979). In fact, a newer algorithm predicted the severity of breast cancer better than

pathologists, partly because it identified *new* cues currently overlooked by pathologists (Beck, Sangoi, Leung, Marinelli, Nielsen, van de Vijver, & Koller, 2011).

Widespread evidence demonstrates the potential of algorithms to improve the accuracy of human judgment across domains. But are people willing to listen to them? This paper examines the apparent distrust in algorithms and simultaneous widespread dependence on them. It tests whether people are ever willing to leverage the power of algorithm and if so, when they are most likely to do so.

Theory of Machine

As humans interact more frequently with programmed agents in their cars, homes, and workplaces, we need to understand their "theory of machine."[1] By *theory of machine*, I refer to lay theories about how algorithmic judgment works. Philosophical work on *theory of mind* considers how people infer intentionality and beliefs in other people and even in other non-humans, such as anthropomorphization of inanimate objects (Dennett, 1987).

Social psychology has taught us much about how people think about other's minds. To pick one example, the fundamental attribution error ascribes more credit to human personality and intentionality than the situation warrants (Ajzen, Dalto, & Blyth, 1979; Jones & Harris, 1967). More recent work tested how anthropomorphizing machines, like the self-driving car, influences our trust in them (Waytz, Heafner, & Epley, 2014). Rather than testing how people impart human judgment on algorithms, this paper tests lay beliefs about how algorithmic and human judgment differ. Theory of mind examines perceptions of algorithms, a flavor of theory of mind.

Apparent Distrust of Algorithms

Despite the helpfulness of algorithms, people appear resistant to allowing a numerical formula to make decisions for them (Bazerman, 1985; Dawes, 1979). The consequences of distrust can prove substantial. To pick one tragic example, in 2004, the captain of Flash Airlines Flight 604, who experienced a condition known as spatial disorientation, trusted his own flawed judgment over the aircraft's instruments. He crashed the plane into the Red Sea, resulting in the largest death toll in Egypt's aviation history (Sparaco, 2006). Reliance on human judgment over a machine's calculations resulted in disaster. Another life or death decision domain is medicine. Cardiologists in one study ignored recommendations from algorithms based on medical board recommendations more than half the time for decisions such as prescribing medicines (Keeffe et al., 2005). Even in industries where algorithms have won widespread acceptance, such as Silver's PECOTA (Player Empirical Comparison and Optimization Test Algorithm) in baseball, acceptance can take time and face substantial resistance (Silver, 2012). Experts appear to prefer their own judgment to algorithmic advice, especially when algorithmic advice threatens their expertise and job security.

Positioning algorithms and computers as competitors, rather than complements, to human judgment may exacerbate distrust. Computers have served as competitors to human experts in games from chess to "Jeopardy!" to Go (IBM's Deep Blue, IBM's Watson, and Google's DeepMind AlphaGo respectively). Popular books explain algorithms in terms of a threat to human agency (Steiner, 2012) and researchers have speculated about which jobs computers might eliminate (Frey & Osborne, 2013). This competitor framing encourages an out-group bias (Brewer, 1979; Harvey, White, Hood, & Sherif, 1961; Sumner, 1906) against algorithms and algorithm-driven computers in favor of human judgment.

An out-group presentation of algorithms along with a baseline of disregarding advice from others may further fuel distrust in algorithmic advice. People show a general tendency to favor their own judgment over advice from other sources (Bonaccio & Dalal, 2006; Yaniv & Kleinberger, 2000; Yaniv, 2004). Specifically, participants fail to adjust their original estimates to incorporate advice from peers as much as they should.

Widespread Dependence on Algorithms

Yet, looking around, people readily rely on algorithms. We use algorithms in place of human secretaries, travel agents, headhunters, matchmakers, D.J.s, movie critics, beauty specialists, clothing stylists, sommeliers, and financial advisors (Siri and Google Search to Kayak, LinkedIn, OkCupid, Pandora, Netflix, Birchbox, Stitch Fix, Club W, and Betterment respectively). Improved accessibility to and affordability of algorithmic advice are two reasons for our growing dependence.

Originally, mainframe computers played an important role in improving decisions. Yet, they were only accessible to large organizations, like governments, with deep pockets and enough square footage to house them. In World War II, the British government relied on Turing's machines to crack German coded communications (Hodges, 2014). With the spread of the smartphone, which moves with us, algorithms seamlessly integrate into our everyday decisions. The affordability of algorithms further increases dependence on them. The rise of algorithms is no more apparent than in robo-advising on financial decisions, where algorithms are replacing human advisors ("Ask the Algorithm," 2015). Expert advice often requires time and money. Algorithms can potentially replace experts by providing instant, inexpensive advice.

Empirical Evidence for Algorithm Aversion

Do people use algorithmic advice optimally? This paper explores the tension between the apparent distrust in, and the obviously widespread dependence on, algorithms. I designed Experiments 1A and 1B to test two competing predictions:

1. People rely on advice from algorithms *less* than advice from people.
2. People rely on advice from algorithms *more* than advice from people.

Anecdotal evidence on the accuracy of algorithms would have us believe that people rely less on algorithmic advice than human advice (Dawes, 1979; Dawes, Faust, & Meehl, 1989; Kleinmuntz, 1990; Meehl, 1954; Meehl, 1957; see also Kleinmuntz & Schkade, 1993). There is surprisingly little experimental evidence testing how people perceive algorithmic advice. Moreover, existing evidence is contradictory.

A handful of empirical experiments from the psychological and computer science literatures have examined people's perceptions of algorithms. In subjective domains governed by personal taste, participants relied on friends over recommender systems for book, movie, and joke recommendations (Sinha & Swearingen, 2001; Yeomans, Shah, Mullainathan, & Kleinberg, 2017). In another experiment, participants who imagined themselves as medical patients preferred to receive a subjectively worded diagnosis from a doctor rather than from a computer (Promberger & Baron, 2006). After seeing an algorithm err, people relied more on themselves than the algorithm, (Dietvorst, Simmons, & Massey, 2015; Dzindolet, Pierce, Beck, & Dawe, 2002) a result the authors called algorithm aversion.

Empirical Evidence for Reliance on Algorithms

The second prediction, that people are willing to rely more on advice from an algorithm than another person, is not without evidentiary support. The computer science literature shows that when solving a logic problem, participants relied on advice from an algorithm more than

advice from other people (Dijkstra, Liebrand, & Timminga, 1998). In one experiment, participants relied more on an algorithm than themselves, even after they saw the algorithm err (Dijkstra, 1999). This result directly contradicts the evidence from Dietvorst et al. (2015) and Dzindolet et al. (2002). Psychological work shows that people depend on algorithmic search engines to remember information they already know (Sparrow, Liu, & Wegner, 2011; Wegner & Ward, 2013), although it does not examine reliance on algorithms for prospective events.

Limitations of Existing Evidence

This section outlines the limitations of existing evidence and how this paper addresses them. First, I address potential confounds of past work: confounding human judgment with the self and confounding both advisors with expertise. Second, I address limitations regarding the subjectivity of decision domains and repeated interactions in past work. The existing empirical work leaves open key questions regarding when people accept or reject algorithmic advice.

Human Judgment Confounded with Self. Extant literature suggests that people are resistant to allowing algorithms to make decisions for them (Dzindolet, et al., 2002; Keeffe et al., 2005; Dietvorst et al., 2015). This comparison confounds human judgment with one's own judgment. The literature on advice-taking shows a robust effect of discounting advice from peers because people have access to their own reasoning and not to others' (Bonaccio & Dalal, 2006; Yaniv & Kleinberger, 2000; Yaniv, 2004). Thus, people often disregard advice when comparing it to their own knowledge. Indeed, work on overconfidence suggests that people rely on their own over others' judgments. People often show excessive confidence in the quality and accuracy of their own knowledge and beliefs, a phenomenon termed overprecision (Moore & Healy, 2008; Moore, Tenney, & Haran, 2016).

Anecdotal observations of algorithm aversion similarly conflate human judgment and the self, where people, especially experts, reject algorithms. The previously discussed doctors (Keeffe et al., 2005) and talent scouts (Silver, 2012), rejected algorithms that can perform their jobs. It seems unsurprising that talent scouts, doctors, and other professionals find a threat to their job aversive. Asking people if they prefer an algorithm to do their job instead of doing themselves is a different question from asking if they prefer to rely on advice from an algorithm or another person. The first question conflates human judgment with one's own knowledge.

Although people may distrust algorithms relative to their own knowledge, they may rely more on algorithmic advice than advice from other people. For example, a driver may deny that he is lost on back roads near his house and rely on his intuition before consulting Google Maps. And people may prefer to rely on Google Maps before asking a stranger for directions. This paper compares reliance on advice between two advisors, controlling for overconfidence in one's own knowledge in Experiments 1A, 1B, 2, 4, and 7. Experiment 3 tests overconfidence as a moderator to reliance on algorithmic advice.

Advisor Confounded with Expertise. Past work conflates expertise with the advisors. Work suggesting algorithm aversion confounds human judgment with expertise. Where participants relied on a doctor more than a computer for a medical decision (Promberger & Baron, 2006), the study potentially confounded expertise and human judgment of the doctor. The same conflation occurs where results show reliance on algorithms. Where participants relied more on an "expert system" than another person (Dijkstra et al., 1998; Dijkstra, 1999), the experiment left open the possibility that participants trusted the computer because of the explicit expert label. Experiments 1 through 4 control for expertise by presenting the advice from another person or other people. In order to examine whether people infer expertise from

algorithmic advice, Experiment 6 orthogonally manipulates expertise of the person and algorithmic versus human advisor. Experiment 7 tests whether expertise of the decision maker moderates reliance on algorithms.

Subjective Domain. Participants relied on other people for recommendations of products (Sinha, & Swearingen, 2001; Yeomans, et al., 2017). Individual preferences are an interesting domain to consider, but one limitation is that it does not allow for an objective standard of accuracy across an entire sample. Expert literary and art critics even differ in their stylistic preferences. Furthermore, a long literature has shown the malleability of individual preferences. Preferences are easily influenced by the contexts in which they are elicited (Lichtenstein & Slovic, 2006). In order to understand when people improve their accuracy by listening to algorithms, this paper first focuses on more objective domains (estimates and forecasts of future events) where accuracy is assessed by numeric measurement or the occurrence of an event. Experiments 5 and 6 then test whether subjectivity of decisions moderates responses to algorithmic advice.

Repeated Interactions. Prior findings of algorithm aversion examined situations when participants saw an algorithm make errors (Dietvorst et al., 2015; Dzindolet et al., 2002). There are many situations in which people have limited experience with a particular algorithm, making it important to examine perceptions prior to repeated feedback. For instance, forecasts with a long time horizon require forecasters to decide if they want to rely on an algorithm prior to any feedback. People may not find out the accuracy of their forecast until years or even decades later; consider outcomes of climate change, political events such as conflicts between nations, economic events, relationship success, etc. The experiments in this paper focus on willingness to rely on algorithms prior to performance feedback.

Overview of Experiments

This paper tests the psychological processes behind people's response to algorithmic and human advice. The experiments test:

1. *If* people are willing to rely on algorithmic advice (while seeking to avoid problems with prior empirical evidence),
2. *When* people are willing to rely on algorithmic advice (by reconciling the results of this paper, reliance on algorithmic advice, with work that finds aversion), and
3. If people rely on algorithmic advice as much as they *should,* in order to maximize their accuracy.

Experiments 1A and 1B test whether people rely on algorithmic advice over human advice. In the interest of making human and algorithmic advice as comparable as possible, this paper focuses on one of the simplest algorithms: aggregation of human judgments. Merely changing the label indicating whether the advice came from other people or an algorithm further enhances comparability between conditions in these studies. Counter to the story of algorithm aversion, the results in Experiment 1A and 1B reveal that participants relied more on the same advice when they thought it came from an algorithm than when they thought it came from other people.

Experiment 2 and 3 test potential moderators to reliance on algorithmic advice: joint versus separate presentation of advisors (Bazerman, Loewenstein, & White, 1992) and confidence in participants' own knowledge. Experiment 4 measures whether people relied on algorithmic advice as much as they *should,* provided normative information, in order to maximize their accuracy. Normatively, people should rely on aggregated judgments (the wisdom of the crowd) more than one advisor (Mannes, 2009; Soll & Larrick, 2009).

Experiment 5 and 6 examine subjectivity of a decision as a moderator. Experiment 6 examines the interaction between decision's subjectivity and expertise of the human advisor. Experiment 7 tests the expertise of the decision maker as a moderator to reliance on human advice, comparing responses from experts who work in national security to those from lay people. The results from these eight experiments help explain why and when people are willing to rely on algorithmic advice, suggesting that algorithm aversion is not as straightforward as prior literature may lead one to expect. They also reconcile the mixed and contradictory prior results.

Maximizing power. In Experiment 1A and all other experiments, I report how I determined sample sizes, pre-registered data exclusions, and all conditions. All sample sizes were determined a priori by running power analyses on G*power (most at 80% power). Where possible, I based effect sizes on prior experiments in the paper. Where that was not possible, I estimated smaller effects for a conservative test (which produces larger sample sizes). Final sample sizes include the number of participants after removing survey responses based on pre-registered exclusionary criteria. I report all exclusions in the Appendix, for ease of reading. The links to pre-registrations, materials, and data are posted online (Simmons, Nelson, & Simonsohn, 2012). I pre-registered Experiments 1B, 2, 3, 4, 5, 6, and 7 (including exclusions) at the Open Science Framework (I had not yet discovered the joys of pre-registration at the time of this experiment).

Diversity and inclusiveness. In order to increasing the generalizability of the results, I collected data from a variety of populations. Convergent evidence from the studies suggest that the results are generalizable to online participants who vary in age, MBA students at a West Coast university, and undergraduates at a West Coast university. In order to understand the

limits to these results' generalizability, I also ran exploratory analyses to understand how some demographic factors, such as age and gender related to reliance on algorithms. Age especially seems relevant to any question related to perceptions of technology, such that older people may feel less familiar with it, in turn influencing their responses. Surprisingly, my results suggest that neither age nor gender have a relationship with reliance on algorithmic advice. Experiment 7 specifically focused on a special sample of participants, people with occupations related to national security in order to test an important moderator to the findings in this paper: expertise with forecasting.

Experiment 1A: Do people rely more on advice from an algorithm or other people?

Experiment 1A tested how much people relied on advice when it came from an algorithm or other people, for a judgment that could be judged on accuracy. This experiment aimed to create a simple, clean test and avoid some potential confounds in prior work: confounding human judgment with the self and confounding expertise with either human or algorithmic advice. All participants received the same advice but the experimental manipulation varied the label of the advisor. In fact, I created the advice by averaging estimates of past study participants.

An average is one of the simplest forms of an algorithm. Using it has the advantage that I can describe it, truthfully, as coming from people or from an algorithm, as it is a mathematical calculation based on human judgment. Aggregation of individual judgements outperforms judgments made by individuals (Larrick & Soll, 2006) because it cancels out errors from individuals (Galton, 1907; Hastie & Kameda 2005; Soll & Larrick 2009; Surowiecki, 2004). Providing good quality advice in terms of accuracy allows me to test if people are willing to listen to advice that can improve their judgment.

The paradigm, adapted from the advice-taking literature's judge-advisor system (JAS) by Sniezek & Buckley (1995), allowed participants to decide how much they wanted to weigh the advice relative to their initial estimate, in both the human and algorithm conditions. This paradigm avoids confounding human judgment with the self by allowing participants in both conditions to weigh their own knowledge relative to an external advisor's advice. The paradigm also controls for expertise and provides advice from peers, rather than experts.

Another important feature of Experiment 1A is the presentation of algorithmic advice as a "black box"—that is, without explanation of the algorithm's inner mechanics. Without access to the algorithm's mechanics, the experiment presents algorithmic advice in a manner similar to how people routinely interact with algorithms on a day-to-day basis. This operationalization parallels the wide-spread appearance of algorithms in daily life, including weather forecasts, population estimates, and economic projections.

Method

Participants

The final sample included 202 participants (90 women; 112 men; *Mdn* age = 28). I determined the sample size a priori with the goal of including 100 participants in each condition. I opened the survey to 200 participants online via Amazon's Mechanical Turk for $0.25.[2] Instructions noted that accurate responses increased participants' chances of winning a $10 bonus. All experiments in this paper incentivized final answers where participants provided estimates or forecasts.

Design

The experiment had a 2-cell (advisor: people vs. algorithm) between-subjects design that manipulated the source of advice participants received. Participants estimated the weight of a

person in a photograph twice: before and after receiving advice. The experimental manipulation varied whether they believed the advice came from another person or an algorithm. I measured how much people relied on the advice. Consistent with the advice-taking literature, I refer to this as Weighting of Advice (WOA).

Procedure and Materials

Overview. I adapted the judge-advisor system (JAS) for the procedure. Participants viewed a photograph of a person (See Figure 1), made an initial weight estimate, received advice about the person's weight, and made a final incentivized estimate. In all conditions, participants read advice that the person weighed 163 pounds.

Advisor manipulation. The instructions described the advice as an estimate from either another person or an algorithm.

In the human condition, participants read, "The average estimate of participants from a past experiment was: 163 pounds."

In the algorithm condition, participants read, "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 advice was the average estimate from 415 participants in another experiment (Moore & Klein, 2008). It was actually quite accurate and only one pound off from the person's actual weight (164 pounds).

Main dependent measure: Weight of Advice (WOA). I measured how much participants relied on the advice. WOA measures the degree to which participants move their estimates toward the advice. The continuous measure captures more information than a binary

choice can capture. In addition to fully discounting or fully updating to the advice, participants could rely on the advice as little or as much as they like.

I measured how much participants relied on the advice by dividing the difference between the final and initial estimate by the difference between the advice and the initial estimate. The higher the WOA, the greater the reliance on the information. Greater reliance also means greater accuracy here because the advice is accurate (one pound away from the actual weight). A WOA of 1 means that the participant matched his or her final estimate to the advice. A WOA of .5 means that the participant averaged the advice with his or her initial estimate. A WOA of 0 means the participant did not change their final estimate from the initial estimate and connotes 100% discounting of the advice.

I included other measures in order to test potential mechanisms for reliance on advice.

Confidence. After each estimate (both initial and final), participants indicated their confidence in it, "How likely is it that your estimate is within 10 pounds of the person's actual weight?" on a scale from 0 = *no chance* to 100 = *absolutely certain*.

Difficulty. After providing their final estimates, participants reported how difficult they found the task, "How easy was it to determine the person's weight from the photograph?" on a scale from 1 = *not at all easy* to 6 = *extremely easy*.

Numeracy. At the end of the survey, participants answered an 11-item Numeracy Scale, consisting of math questions (Schwartz, Woloshin, Black, & Welch, 1997). The higher a participant's numeracy score, the greater their mathematical literacy and comfort with numbers (0 to 11).

Results

Did participants rely more on advice when it came from an algorithm or people? Participants relied more on the same advice when they thought it came from an algorithm (M = .45, SD = .37) than when they thought it came from other people (M = .30, SD = .35), $F(1, 200) = 8.86$, p = .003, d = .42. Results hold when controlling for gender, numeracy, and Time 1 confidence, $F(1, 197) = 9.02$, p = .003. There are no main effects of these three variables (Fs < .39, ps > .52). See Figure 2. The pattern of WOA is mirrored in participant's change in confidence. Participants in the algorithm showed a greater increase in confidence at Time 2 (M = 7.79, SD = 10.90) than participants in the human condition (M = 4.48, SD = 8.83), $t(200) = 2.37$, p = .019, d = .33.

Did numeracy correlate with reliance on advice (WOA)? Interestingly, for participants in the algorithm condition, higher numeracy correlates with WOA, $r(100) = .21$, p = .037. As expected for the human condition, there is no correlation, r = -.12, p = .225. Neither age, ps > .61, nor perceived difficulty of task, ps > .37, correlate with WOA in either condition. Did distance from the advice matter? Perhaps people inferred the quality of advice based on its proximity to their original estimate. In a non-preregistered analysis, neither Time 1 estimates (p = .55) nor Time 1 confidence ratings (p = .37) correlate with WOA.

Discussion

Participants relied more on advice when they thought it came from an algorithm than when they thought it came from another person. This result contrasts the conclusion from the literature that people are averse to algorithms. It suggests that one way to increase adherence to advice is to provide algorithmic advice. An important feature of this experiment was the objective nature of the task. It provided an objectively correct answer that allows for a measure

of accuracy. As the advice used many estimates, it was fairly accurate in this experiment (as is also the case in Experiments 1B, 2, 5, and 6). Importantly, greater reliance on advice means greater accuracy. People were willing to listen to algorithmic advice, which allowed them to improve their accuracy.

Despite uncertainty about the black box algorithm's process, participants showed willingness to rely on it. A black box operationalization of algorithmic advice creates a conservative test of reliance on algorithms for two reasons. First, without seeing an equation, the algorithmic advice is more equivalent to the human advice. Second, distrust of algorithms may arise for many reasons, including the opacity of its process. Uncertainty in the algorithm's process increases when its calculations are not available. Regardless of opacity, people should have a better understanding of how others arrive at judgments, based on their own experience making the judgment. Familiarity with the *kinds* of calculations commonly driving algorithms could increase reliance on a black box algorithm, as suggested by the result that mathematical knowledge enhances reliance on algorithms. Opacity of an algorithm's process (Finkel, Eastwick, Karney, Reis, & Sprecher, 2012) may become more relevant to reliance on algorithmic advice as organizations begin to share data with the public.

Another important feature of this experiment is operationalization of human judgment. Participants in the human advice condition read about an average estimate, rather than a single individual. One possible explanation for prior results that found reliance on algorithms is that participants chose between advice from an "expert system" (algorithm) and a *single* individual (Dijkstra, 1999; Dijkstra, Liebrand, & Timminga, 1998). Participants may have assumed that it could draw from multiple inputs. It is rational to rely on the wisdom of the crowd over a single judgment (Mannes, 2009; Soll & Larrick, 2009). Experiment 1A avoids the asymmetry of an

algorithm which may access multiple inputs and a single human judgment and still finds reliance on algorithms.

At least for this incentivized, objective task, participants relied more on algorithmic advice than human advice. These results suggest that the assumption of algorithm aversion is not as straightforward as past work suggests. Experiment 1B rules out a few alternative explanations for Experiment 1A's results.

Experiment 1B: Forecasts by MBA Students

Experiment 1A found that participants relied on algorithmic advice more than advice from other people. Experiment 1B sought to replicate these results while increasing their generalizability and ruling out potential alternative explanations. It also tested whether familiarity with the domain affected reliance on algorithmic advice. The experiment tested the hypothesis that people would rely on algorithmic advice, but less so for a task where they have experience making judgments. To increase generalizability, it used a different sample of participants, MBA students.

Experiment 1B operationalized both human and algorithmic advice differently than 1A. To ensure that the conditions did not differ in number of words and to allow participants the use of their own default interpretations of human judgment and algorithmic advice, Experiment 1B, used simpler advisor labels ('person' and 'algorithm'). To better understand people's default interpretations of algorithms, I asked them to define the term at the end of the survey.

Experiment 1A created a conservative test by basing both conditions on estimates from multiple people. Such a comparison gives the human advisor a fighting chance (as opposed an algorithm that scans the photo or uses complex equations). Although unlikely in a between-subjects design, a potential alternative explanation for the results is that participants viewed

"averaging" in the human condition as a second-rate algorithm. It is also a possibility that people found the word "average" aversive, thinking about its connotation of mediocrity. There is evidence that people equate the accuracy of averaged estimates with that of an "average individual" (Larrick & Soll, 2006). Experiment 1B attempts to rule out these explanations by operationalizing human advice as a single person instead of an average estimate.

A second potential explanation for Experiment 1A's result is that participants viewed the algorithm as a proxy for the researcher who may have known the answer to the estimation because she created the study. Experiment 1B rules out this explanation by including a forecast task. Participants cannot assume that the researcher knows the future. A different task also helps increase the generalizability of the results. I predicted that people rely on advice from algorithms more than advice from people, but that prior experience with the task (rather than experience with the algorithm) decreases reliance.

Method

Participants

The final sample included 77 participants (27 women; 50 men; *Mdn* age = 28). I determined the sample size based on the number of available students across two sections of an MBA class from a West Coast university.

Design

Experiment 1B had the same design and main dependent variables as Experiment 1A: a 2-cell (advisor: person vs. algorithm) between-subjects design that manipulated the source of advice participants received. I measured reliance on advice.

Procedure and Materials

Overview. Although similar to Experiment 1A, Experiment 1B (and later studies) did not include the difficulty measure and numeracy scale. Participants forecasted a movie's gross profit on its opening weekend and two geopolitical events, in addition to the same weight task.

The experiment held the order of tasks constant so that participants answered the weight-guessing task first. This order provided a direct replication of Experiment 1A. The participants had previously researched and forecasted the gross opening weekend for different movies as a weekly class assignment. This domain allowed me to test if people rely on algorithmic advice when they have experience with the task. Advice for the movie forecast came from a website that made forecasts for each movie opening the upcoming weekend. Participants saw two geopolitical forecasts, a topic participants had not researched previously for class. See Table 1 for forecast wording. I took the advice from a forecasting tournament (including thousands of participants) hosted by the federally funded Good Judgment Project (GJP), on which I collaborated. Prior to forecasting the two world events, participants received information about the global forecasting tournament run by the GJP, "The Good Judgment Project, a group of federally funded researchers, hosts a forecasting tournament where thousands of lay people

around the world compete to make the most accurate forecasts about global political events." I measured how much participants relied on the advice, Weighting of Advice (WOA).

Advisor manipulation. Prior to their second estimate, all participants received the same advice (163 pounds for the weight task, 14.40 million for the movie forecast, 25% for the first geopolitical forecast, and 8% for the second geopolitical forecast).

In the person condition for the weight task and movie forecast, participants read, "The estimate of another person was: X." For the two geopolitical forecasts, participants in the human condition read, "The estimate of another forecaster was: X%." The materials referenced another forecaster, rather than another person, in order to make it clear that the advice came from an individual in the forecasting tournament previously described.

In the algorithm condition for all tasks, participants read, "The estimate of an algorithm was: X."

Main dependent measure: Weight of advice (WOA). As in Experiment 1A, I measured how much participants relied on the advice.

Definition of algorithm. After Experiments 1B and 2, participants defined the term "algorithm."

Results

Did participants rely more on advice when it came from an algorithm or from other people? I submitted the weight estimation, movie forecast, and the two geopolitical forecasts (averaged together) to a 2-cell (advisor: person vs. algorithm) MANOVA. Seventy-three participants had useable WOA measures for all tasks. Again, participants relied more on the advice when it came from an algorithm than from another person, as evidenced by the main effect of advisor, $F(1, 71) = 13.97$, $p < .001$. See Figure 3. Participants relied more on

algorithmic advice than advice from another person for the weight estimate, $F(1, 71) = 13.97$, $p < .001$, and for the geopolitical forecasts, $F(1, 71) = 4.87$, $p = .031$.[3] However, for the movie forecast, people discounted the algorithmic advice as much as they discounted the advice from the person, $F(1, 71) = .72$, $p = .398$.

Participants in Experiment 1B and Experiment 2 merely read the label "algorithm" and thus needed to rely on their own default interpretations of the term. What *are* people's default interpretations? 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. Although people provided fairly general definitions, they alligned with the construct used in this paper: a series of mathematical calculations. See Table 2.

Discussion

Experiment 1B replicated reliance on algorithmic advice. As 1B did not include the word "average" in the human condition, the results cast doubt on the alternative explanation for 1A that participants relied on the algorithm because they viewed the human advisor as a second-rate algorithm. Reliance on the algorithm did not depend on conceptualizing human judgment as an average or as an estimate from a specific person. The forecasting tasks rule out the possibility that people saw the algorithm as a proxy for the researcher who had access to the correct estimate.

Participants relied on algorithmic advice with one exception. Interestingly, participants equally discounted all advice for the movie forecast; possibly due to experience with the task over the course of the class. Experiment 7 further explores the role of expertise as a moderator to reliance on algorithmic advice. The results of Experiments 1A and 1B contradict the idea of

algorithm aversion. Experiment 2 attempts to reconcile these results with past work, by experimentally manipulating a difference between Experiments 1A and 1B and work that finds aversion.

Experiment 2: The effect of joint versus separate evaluation on algorithm reliance

Experiment 2 tested a possible moderator to reliance on algorithmic advice: joint vs. separate evaluation. It also attempted to increase generalizability, by using yet a different sample of participants, undergraduate students. A major difference between Experiments 1A and 1B and virtually all work that finds aversion is the way the advisors are presented to participants. Preference reversals occur between joint and separate presentation of choices. Attributes which are otherwise difficult to evaluate without a comparison are easier to evaluate in a joint evaluation due to a greater amount of information (Bazerman, Loewenstein, & White, 1992; Bazerman, Moore, Tenbrunsel, Wade-Benzoni, & Blount, 1999; Hsee, 1996; Hsee, Loewenstein, Blount, & Bazerman, 1999).

Experiments 1A and 1B asked participants to evaluate either advice from a person *or* an algorithm (separate evaluation) while prior work asked participants to choose between a person (usually themselves) *and* an algorithm (joint evaluation). I predicted that people prefer an algorithm when evaluating the advisors separately but prefer a person when evaluating them jointly.

Method

Participants

The final sample included 154 participants (104 women; 50 men; *Mdn* age = 21) from a West Coast university's credit and paid subject pools. I determined the sample size a priori with the goal of including 150 participants, for 50 participants per cell. I based my power analysis on

an ANOVA to provide a more conservative sample size than a t-test and adjusted the sample size from a four-cell design to the experiment's three-cell design. I needed one hundred and ninety-nine participants, 50 per cell, to detect a medium effect size (Cohen's d = .4) for an interaction in 2 x 2 ANOVA at 80% power. This study had 3 cells and needed 150 participants.

Design

The experiment had a design similar to that of Experiment 1A, with the addition of a third condition: joint presentation of advisors. It had a 3-cell (person vs. algorithm vs. choice between a person and algorithm) between-subjects design that manipulated the presentation of the advisors, either separately or jointly in the choice condition.

Procedure and Materials

Overview. The materials differed from Experiment 1A in two ways: the addition of a third condition. Second, participants all learned about the advice that they were about to receive before they actually received it. They were randomly assigned to 1. receive advice from another participant, 2. receive advice from an algorithm, or 3. choose between another participant or algorithm as an advisor. In the choice condition, the materials counterbalanced the order of advisors. Then, participants read the advice.

Advisor manipulation.

This experiment uses a more specific operationalization of the human advisor than in Experiment 1A ("average of past participants") and Experiment 1B ("another person"): "another participant."

In the human condition, participants read, "The estimate of another participant was: 163 pounds."

In the algorithm condition, participants read, “The output that an algorithm computed as an estimate was: 163 pounds.” In the choice condition, participants read about the advisor they chose.

Weight of advice (WOA). I measured how much participants relied on the advice.

Choice. Participants chose to receive advice from the person or algorithm.

Results

As in Experiments 1A and 1B, participants who evaluated advisors separately relied more on the advice of the algorithm (M = .50, SD = .37) than the other participant (M = .35, SD = .36), $t(100) = 2.10$, $p = .038$, $d = .44$. See Figure 4. Evaluating the two advisors jointly did not reverse the preference for algorithmic advice. When choosing between the two advisors, 75% of participants chose to see advice from the algorithm (algorithm: N = 39; person: N = 13). Not surprisingly, participants relied similarly on the advisor they chose, be it the algorithm (M = .52, SD = .37) or other participant (M = .41, SD = .36), $t(50) = .90$, $p = .371$.

Discussion

Reliance on algorithmic advice appears robust to different presentations of advisors. Participants relied more on the algorithmic than human advice, regardless of whether they evaluated the advisors separately or jointly. The results speak to the strength of reliance on algorithmic advice, especially considering how 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). The experiment replicated reliance on algorithmic advice with undergraduate students and used different operationalizations of both advisors.

Experiments 1A, 1B, and 2 controlled for overconfidence by allowing participants to consider their own knowledge relative to the advice in both conditions, not just the algorithm condition. Such a design is in-line with work on advice-seeking but contrasts work showing algorithm aversion. Experiment 3 sought to reconcile the current work with past results by manipulating whether participants chose advice from an algorithm and either their own estimate or another person's advice.

Experiment 3: The effect of overconfidence on algorithm reliance

Experiment 3 tested a moderator of reliance on algorithmic advice: overconfidence. Experiments 1A, 1B, 2 and 3 purposefully controlled for overconfidence, excessive certainty in one's own knowledge. Prior work suggests that people are resistant to allowing algorithms to make decisions for them (Dzindolet, et al., 2002; Keeffe et al., 2005). Indeed, work on overconfidence and advice-taking shows that people rely more on their own judgment than others' judgments (Gino & Moore, 2007). Similarly, robust results on advice-taking show that people underweight advice from others (Gardner & Berry, 1995; Harvey & Fischer, 1997; Yaniv & Kleinberger, 2000; Soll & Larrick, 2009).

I used the materials of Dietvorst et al.'s Experiment 3A, where people chose between their knowledge and an algorithm's estimate. Then, I added a new condition where participants chose between another person's answer and an algorithm's, as in Experiments 1A, 1B, and 2. Thus, I manipulated the source of human judgment: the self or another person.

Work that finds algorithm aversion finds that people display it *after* people see an algorithm err (Dietvorst et al., 2015). Performance feedback is key within Dietvorst et al.'s insightful experiments. In fact, *prior* to seeing feedback, participants happily relied more on algorithms than themselves, which is consistent with the result of this paper. I chose to replicate

Experiment 3A in Dietvorst et al. because here, participants came closest to displaying algorithm aversion *prior* to feedback; they were indifferent between their own estimate and an algorithm's. This small difference in participants' choice provides the opportunity to test a potential moderator.

Experiment 3 tests whether the option to choose one's own knowledge over algorithmic advice moderates reliance on algorithms. I predicted that people rely less on algorithms when they compare an algorithm's advice to their own estimate but rely more on algorithmic advice relative to another person's advice.

Method

Participants

The final sample included 403 participants (177 women; 226 men; *Mdn* age = 32). Participants on Amazon's Mechanical Turk received $0.80 to take the survey. I used the effect size from the control condition of Dietvorst et al.'s Study 4 (not 3A, the materials I used that had the self versus algorithm comparison) because here, participants relied *more* on algorithmic advice than advice from another participant, r = .39. In order to detect an attenuated interaction, I doubled the participants per cell (Simonsohn, 2014), which produced 88 X 2 = 176 people per cell. This current experiment had 2 cells and needed 352 participants. I aimed to collect a sample of 400 participants to ensure a well-powered experiment.

Design

The experiment had a 2-cell (self vs. other) between-subjects design. The manipulation varied whether their own or another participant's estimate served as the human estimate. All participants had to choose whether to rely on an algorithm or a human. This choice determined

participants' bonus payment. Participants received $1 bonus if the estimate they selected proved accurate. I measured the percentage of people who chose the algorithm to determine their pay.

Procedure and Materials

Overview. I replicated the materials of the control condition in Dietvorst et al.'s (2015) Experiment 3A. All participants read about the estimation task:

> …rank (1 to 50) of individual U.S. states in terms of the number of airline passengers that departed from that state in 2011. A rank of 1 indicates that the state had the most departing airline passengers, and a rank of 50 indicates that it had the least departing airline passengers.

Then they read about the information provided to all judges (the algorithm and either themselves or another participant). See Figure 5. Prior to making their own estimate, participants chose how to determine their bonus pay, either based on the accuracy of the algorithm's estimate or on the accuracy of a person's estimate (either their own, as in Dietvorst et al., or another participant, consistent with Experiments 1A, 1B, and 2 in this paper). All participants read that "experienced transportation analysts" developed the algorithm and that they would receive the same information it used.

I determined the bonus as follows:

$1.00 - perfectly predict state's actual rank
$0.85 - within 1 rank of state's actual rank
$0.70 - within 2 ranks of state's actual rank
$0.55 - within 3 ranks of state's actual rank
$0.40 - within 4 ranks of state's actual rank
$0.25 - within 5 ranks of state's actual rank
$0.10 - within 6 ranks of state's actual rank

Participants chose their advisor before seeing the advice similar to Experiment 2. They also read that they would make an estimate, regardless of their choice. Then, participants made their own estimate, regardless of how they determined their final pay.

Self / Other manipulation. Participants either chose between their *own* estimate and an algorithm's estimate (as in the original materials) or between *another participant's* estimate and an algorithm's estimate (the single change to the original materials).

Participants in the self condition read that they could choose the algorithm's estimate or "your estimate" to determine their bonus pay.

Participants in the other condition read that they could choose the algorithm's estimate or "another participant's estimate" to determine their bonus pay.

Choice. Before making their own estimate, participants chose which estimate determined their pay, "Would you like your (the other participant's) estimated rank or the model's estimated rank to determine your bonus?"

Results

The results reveal a preference for algorithms. More participants chose an algorithm over a person to determine their bonus pay, χ^2 (1, N = 206) = 118.14, $p < .001$, $r = 0.76$, consistent with results from Experiment 1A, 1B, and 2. They also chose the algorithm's estimate over their own, $\chi 2$ (1, N = 197) = 20.15, $p < .001$, $r = 0.32$. The option to choose one's own estimate appears to attenuate reliance on algorithmic advice: the effect size for preferring the algorithm is greater in the other condition than the self condition, $z = 6.62$, $p < .001$. See Figure 6. The results are consistent with work on overconfidence in judgment, attesting to the excessive faith people have in the quality of their own judgment.

The odds of choosing algorithmic advice is 3.73 times higher when participants choose between an algorithm and another person's advice than when participants choose between an algorithm and their own estimate, $\chi 2$ (1, N = 403) = 27.35, $p < .001$, $r = 0.26$. Results hold when choice is regressed on self vs. other in a logistic regression, $\chi 2$(1, N = 403) = 28.10, $p < .001$.

Discussion

Introducing overconfidence to the judgment process decreased reliance on algorithms but did not produce aversion. When provided the option to choose their own judgment, more participants still chose the algorithm. Experiment 3 replicated reliance on algorithms using a different estimation task and shows that overconfidence attenuates the effect of reliance on algorithms. Replicating materials from past work decreased the number of factors that varied between Experiment 3 and past work. Experiment 3 may not have found a complete reversal of reliance to aversion because the materials produced indifference between the algorithm's estimate and participant's own estimate in the original paper. It seems that prior to feedback, people are open to algorithmic judgment.

These results suggest that prior work may have found algorithm aversion partly because participants chose between their own estimate and an algorithm's, as a robust literature on overconfidence would predict (Harvey, 1997). Overconfidence in the accuracy of one's judgment (also known as overprecision) is difficult to overcome (Soll, Milkman, & Payne, 2016; Soll & Klayman, 2004); but the results from Experiment 3 suggest that advice from an algorithm may help. Experiment 4 sought to better understand how effectively people utilize algorithmic advice.

Experiment 4: Do people rely on algorithmic advice as much as they should?

Experiments 1A, 1B, 2, and 3 controlled for accuracy of the advice by providing the same advice across conditions. This helped ensure that participants did not rely on algorithms simply because they provided superior advice. Experiment 4 allowed the accuracy of human and algorithmic advice to vary, as it often does in the real-world. It varied the amount of information (as operationalized by the number of individual estimates) used to produce advice. Crowds are

wiser than individuals as the robust findings on the wisdom of crowds demonstrates (Galton, 1907; Surowiecki, 2004), especially as the crowd size increases (Einhorn, Hogarth, & Klempner, 1977).

The literature on wisdom of crowds and advice-taking suggests a normative benchmark for optimal reliance on algorithmic advice based on aggregated individual judgments. In fact, participants should rely more on larger crowds (Mannes, 2009; Soll & Larrick, 2009). Thus, even with limited information about advice, people should optimally average between their own estimate and the estimate from one other person, a WOA of .5. When advice comes from more than 300 people (as in the case of the algorithmic advice in this paper), people should essentially weight it 100% (WOA of 1). This experiment provides participants with normative information by providing the number of individual estimates used to produce the advice. Thus, participants' responses can be compared to a normatively "correct" benchmark; how much they ought to weight advice, given the number of estimates taken into account. I predicted that people rely on algorithmic advice less than they should.

Method

Participants

The final sample included 671 participants (357 women; 314 men; *Mdn* age = 33). I needed a sample of 620 participants to detect an interaction with an effect size $d = .25$ ($f = .125$) at 80% power. Experiments run to-date had produced an effect for reliance on algorithms of $d = .35$. I powered the analysis to detect a smaller effect in order to appropriately power a covariate (experience with a failed algorithm). [4]

Design

The experiment had a 2-cell (advisor: person vs. algorithm) between-subjects design. As in Experiments 1A, 1B, and 2, the experimental manipulation varied whether participants saw advice labeled as having come from an algorithm or from human advisors. I measured WOA (weighting of advice).

Procedure and Materials

Overview. I used a similar procedure to Experiment 1A with two main changes. Instead of guessing weight, participants guessed the age of a different person in a photograph. See Figure 7. Participants made their first estimate and prior to reading the advice, read details about the advice. Then they read the advice and gave their final answer.

Advisor manipulation. Participants received normative information, the number of estimates used to produce the advice.

In the human condition, participants read that they would see advice from past pre-test participants, "a randomly chosen participant from a pool of 314 participants who took a past study." These participants received randomly chosen advice. Because another randomly selected participant is likely (on average) to have an estimate as accurate as one's own, it is normatively correct to average one's own estimate with this advice (WOA = .5).

In the algorithm condition, participants read that they would see advice from, "an algorithm, based on estimates of 314 participants who took a past study." These participants all received the average guess from the 314 past participants: 66. An algorithm based on such a large number of useful inputs is likely to be quite accurate, so it is normatively correct to update fully and match this advice (WOA = 1). The algorithmic advice was fairly accurate as the person in the photograph was actually 63 years old.

Main dependent measure: Weight of advice (WOA). I measured how much participants relied on the advice.

Results

Weighting of advice (WOA). Replicating prior results, participants relied more on advice from the algorithm (M = .34, SD = .34) than another person (M = .24, SD = .27), $F(1, 669) = 17.68$, $p < .001$, d = .33. But did participants rely on the algorithmic advice as much as they *should* have, given the normative information they received? Compared to the normative benchmark of how much people should have weighted the advice from the person (.5) and algorithm (1), participants underweighted advice from the algorithm more than (M = .66, SD = .34) they did from the person (M = .26, SD = .27), $F(1, 669) = 275.08$, $p < .001$, d = 1.30. See Figure 8. They should have relied on the algorithmic advice much more than they actually did.

As a (pre-registered) exploratory analysis, I compared the mean absolute error between the conditions (computed by taking the absolute difference between the actual age and each participants' final estimate). Participants achieved less mean absolute error when then received advice from the algorithm (M = 4.59, SD = 3.76) than another person (M = 5.67, SD = 3.57), $F(1, 669) = 19.63$, $p < .001$, d = -.29. Are younger people more willing to listen to algorithmic advice? As non-preregistered, exploratory analysis, age was not a significant predictor of WOA when I included it in the model, $F(1, 668) = 1.91$, $p = .167$.

Discussion

Again, participants relied more on an algorithm than a person. But participants could have used the algorithmic advice more than they actually did. Participants *should* have updated 100% to the algorithmic advice because it used more individual judgments. They greatly underweighted it, showing insensitivity to the number of estimates. Participants underweighted

the advice from other people, but because the normative benchmark is to update only 50%, participants underweighted advice from the algorithm more than they underweighted advice from a person. Although not averse to algorithmic advice, participants could rely on algorithms even more, which would boost their accuracy. Experiment 5 moves beyond purely objective domains to test a mechanism for why people rely on algorithms in the first place: subjectivity of the domain.

Experiment 5: The effect of subjectivity on algorithm reliance

Experiment 5 tested whether people are more likely to rely on algorithmic advice for more objective decisions. The results so far suggest that people are willing to rely on algorithmic advice more than other people, and even more than themselves. These results are a contrast to the algorithm aversion found in more subjective domains such as book, movie, and joke recommendations (Sinha, & Swearingen, 2001; Yeomans, et al., 2017). Could subjectivity of the domain moderate reliance on algorithms?

My results also contrast a preference for receiving a hypothetical medical diagnosis from a doctor rather than from an algorithm (Promberger & Baron, 2006). At first glance, this decision seems directly at odds with the current results. On closer inspection, participants may have viewed the decision as more subjective than objective. Prior to choosing an advisor, participants read test results such as, “LDL blood cholesterol levels can have the values low, normal, high, very high. Yours is very high.” Qualitative descriptions may frame the decision as more subjective. Doctors in the real world often provide numeric information and sometimes follow up with qualitative descriptions, which may have increased the salience of this qualitative description. For instance, patients may often ask a doctor to translate numerical results, asking how their result compare to a “normal” result.

I predicted that in a more objective domain like a financial decision, where people prefer more rational decision making (Hsee, Zhang, Yu, & Xi, 2003), people prefer algorithmic advice. In a more subjective domain like dating, where people think intuition plays a role and prefer more affect-based decision making (Weber & Lindemann, 2008), they prefer advice from people.

Method

Participants

The final sample included 276 decisions from 51 participants (27 women; 24 men; *Mdn* age = 28). I pre-determined a sample size of 50, as the power comes from the number of decisions produced by the participants. Participants provided three decisions for which they expected others to rely most heavily on advice from an algorithm and three for those which they expected others to rely most heavily on advice from other people.

Design

The experiment had a 2-cell (advisor: person vs. algorithm) within-subjects design. Participants listed decisions for which they thought people would rely most heavily on advice from an algorithm and decisions for which they thought people would most heavily rely on advice from another person. Two research assistants, blind to the conditions, coded the decisions as subjective or objective. I measured the number of decisions in each condition coded as subjective or objective.

Procedure and Materials

Participant-generated decisions. Participants provided decisions in response to these counter-balanced, open-ended questions:

"For what kinds of decisions do you expect other people to rely the most on advice from an algorithm (rather than on advice from a person)?"

"For what kinds of decisions do you expect other people to rely the most on advice from a person (rather than on advice from an algorithm)?"

Coding. Two research assistants coded the participant-generated decisions, randomly presented through the Qualtrics survey program, as 1 = *very subjective,* 2 = *neither subjective nor objective* or 3 = *very objective*.

Dependent variable: Number of subjective and objective decisions. I measured the number of subjective and objective decisions participants listed.

Results

In order to keep the sample size as large as possible, I made two decisions prior to data analysis that were not pre-registered: I broke 9 ties myself (blind to condition), when the decision was coded as objective by one coder and subjective by the other. Second, I removed the 15 decisions rated as "uncertain" by both coders. The coders showed high inter-rater reliability (α = .85) prior to tie-breaking and exclusions.

In order to test whether subjectivity moderated reliance on algorithmic advice, I submitted ratings to a 2 (subjective, objective) X 2 (person, algorithm) chi-square test. When the participant wrote about an algorithmic advisor, the odds listing an objective decision were 16.67 times higher. When the participant wrote a human advisor, the odds of a listing a subjective decision were 16.56 times higher, χ^2 (1, N = 276) = 92.66, $p < .001$, $r = 0.58$. See Figure 9. These results suggest that participants expected a preference for algorithmic advice for more objective decisions but a preference for human advice for more subjective decisions.

Participants expected greater reliance on algorithmic advice for a greater number of objective decisions, χ^2 (1, N = 144) = 87.11, $p < .001$, $r = 0.78$. They expected greater reliance on human advice for a greater number of subjective decisions, χ^2 (1, N = 132) = 16.03, $p < .001$, $r = 0.35$. People listed a greater variety of decisions for human than algorithmic advisors, $z = 5.78$, $p < .001$, suggesting that subjectivity influences preferences for algorithmic advice more than it influences preferences for human advice.

Discussion

In Experiment 5, subjectivity moderated the preference for algorithms: people listed an algorithm for objective decisions and a person for subjective decisions. Decisions for algorithmic advice showed a starker contrast in subjectivity, suggesting that subjectivity of the decision matters more when people consider algorithmic than human advice. These results help explain why prior work finds algorithm aversion to recommendations about books, jokes, and movies and why my experiments find reliance on algorithmic advice for estimates and forecasts. Experiment 6 extends the findings of subjectivity as a moderator to test it as a mediator. Additionally, it attempts to reconcile this paper's findings with past work suggesting aversion by examining the interaction of subjectivity and expertise.

Experiment 6: The interaction of subjectivity and expertise on algorithm reliance

The results from the participant-generated responses in Experiment 5 suggested that subjectivity moderates reliance on algorithmic advice. Experiment 6 used the most frequent responses from Experiment 5 to directly manipulate subjectivity. It tested subjectivity as a mediator and tested the interaction between subjectivity and the expertise of the advisor.

Results showing that people preferred to receive a medical diagnosis from a doctor (an expert) than from a computer (Promberger & Baron, 2006), do not clearly distinguish the

separable influences of human vs. expert guidance. Likewise, results that show people view "expert systems" as more objective and "rational" than humans (Dijkstra, Liebrand, & Timminga, 1998) are susceptible to the alternative explanation that participants read advice from an "expert system" and treated it as an expert. Experts often produce superior advice to non-experts. Experiment 6 sought to explain why people rely on algorithmic advice by distinguishing between the separable influences of algorithmic vs. expert guidance. I predicted that subjectivity moderates reliance on algorithmic advice; but that participants prefer expert advice when it is available, regardless of subjectivity.

Method

Participants

The final sample included 550 online participants (women = 222; men = 328; *Mdn* age = 31). In order to detect an attenuated interaction, I first estimated an effect of subjectivity as d = .35. I based the effect size on the average effect size from all studies with WOA run to-date. In order to detect that effect size at 80% power, I needed a total of 260 participants for a 2-cell design. To account for an attenuated interaction, I followed Simonsohn (2014) and doubled the number per cell, which produced 130 X 2 = 260 participants per cell. This current study had 2-cells and needed 520 participants.

Design

The experiment had a 2 (subjectivity: subjective vs. objective) X 2 (expertise of person: expert vs. non-expert) mixed design with subjectivity varying within-subjects and expertise varying between-subjects. I measured how much participants relied on advice from an algorithm relative to another person (or expert).

Procedure and Materials

Overview. Participants read twelve decision problems and were asked to, "imagine you are about to make each decision." See Table 3. They read six subjective and six objective decisions, with order of subjectivity counterbalanced.

I based the decisions on Experiment 5, prior work, and other real-world decisions where people could use algorithms. For instance, materials included a medical decision in the objective condition and a decision about books and movies in the subjective condition. The materials also included decisions that could be informed by algorithms in the real-world: investment, dating, and clothing decisions. While Betterment touts its financial advice as algorithmic, OkCupid does not explicitly label the date recommendations as algorithmic, and Stitch Fix focuses on the "personal stylist" component of their service.

Subjectivity manipulation. Participants read both subjective and objective decision problems (previously pre-tested with another sample of mTurkers).

Expertise manipulation. Participants read about either non-experts (another person) or an expert. Instead of using a general label of "expert," participants read about specific professional roles that signaled expertise (where the advisor made the decision or gave advice about that decision for a living).

Main dependent measure: Reliance. For each decision, participants reported, "When you make your decision, you have the opportunity to receive advice from an algorithm or another person. Do you rely more on advice from an algorithm or another person (or specific expert – e.g., doctor)?" on a scale from 1 = *rely most heavily on another person* to 7 = *rely most heavily on an algorithm*. This measure forced participants to directly compare the two advisors, which complements the WOA measure in the other experiments.

Perceived subjectivity. I tested perceived subjectivity of each decision as a mediator. As a clarification, participants read, "By subjective, we mean a decision that you make based on emotion or intuition. By objective, we mean a decision that you make based on logic or reason." Then they answered, "How subjective / objective do YOU think the following decisions are?" on a scale from *1 = completely subjective* to 7 = *completely objective*.

Importance. I statistically controlled for a potential confound to the subjectivity conditions: importance of the decision. Participants rated, "How important do YOU think the following decisions are?" on a scale from 1 = *not at all important* to 7 = *very important.*

Results

Consistent with my pre-registered analysis, I excluded one decision where perceived subjectivity did not differ from the mid-point of the scale: moving. Participants rated all subjective decisions lower than the mid-point of the scale (more subjective) and all of the other objective decisions higher than the mid-point of the scale (more objective). Participants rated the objective decisions as more important (M = 5.23, SD = 1.16) than the subjective decisions (M = 3.02, SD = 1.37), $t(549) = 28.69$, $p < .001$,[5] which suggests that subjectivity and importance are not independent. Thus, the analyses control for importance.

Reliance. I averaged responses across decisions and submitted them to a 2 (subjectivity: subjective vs. objective) X 2 (expertise of person: expert vs. non-expert) mixed ANCOVA with expertise as a between-subjects factor and subjectivity as a within-subjects factor. I averaged importance across the decisions for each subjectivity condition and included it as a within-subjects covariate. There is an effect of importance, $F(1, 1093.64) = 7.80$, $p = .005$, and controlling for it, there is both a main effect of subjectivity, $F(1, 831.05) = 94.60$, $p < .001$, and interaction between subjectivity and expertise, $F(1, 825.56) = 94.79$, $p < .001$. See Figure 10. A

main effect of expertise is not relevant to my predictions, $F(1, 545.97) = 31.49$, $p < .001$. These results are also significant without controlling for importance (ps $< .001$).[6]

In the non-expert condition, participants relied more on algorithmic advice in the objective (Adjusted $M = 4.32$, $SE = 0.08$) than subjective condition (Adjusted $M = 3.00$, $SE = 0.08$), $F(1, 713.55) = 208.85$, $p < .001$. The interaction suggests that responses in the expert condition showed a different pattern. In fact, participants preferred the expert advice, regardless of the decision's subjectivity (Objective: Adjusted $M = 3.20$, $SE = 0.08$; Subjective: Adjusted $M = 3.10$, $SE = 0.08$), $F(1, 714.50) = 1.05$, $p = .306$.

Within-subjects mediation analysis. I tested whether perceived subjectivity mediated the relationship between the subjectivity and reliance on algorithms, when an expert was not available. Because I wanted to control for importance of the decision within-subjects for each subjectivity condition, I followed the within-subjects mediation analysis in Critcher and Dunning (2009) based on Judd, Kenny, & McClelland (2001).

First, I tested that the subjectivity manipulation predicts perceived subjectivity, controlling for importance. I averaged perceived subjectivity ratings across decisions and submitted them to the 2-cell (subjectivity: subjective vs. objective) ANCOVA with importance as a within-subject covariate. Participants perceived the decisions as more objective in the objective condition (Adjusted $M = 5.08$, $SE = 0.08$) than in the subjective condition (Adjusted $M = 3.21$, $SE = 0.08$), $F(1, 401.40) = 245.07$, $p < .001$, controlling for importance (which was itself significant, $F(1, 444.77) = 40.09$, $p < .001$).

Next, I tested that perceived subjectivity correlated with reliance on algorithmic advice (relative to advice from another person) for each subjectivity condition. In both conditions, perceived subjectivity correlates with relative reliance on algorithmic advice, controlling for

importance, objective: $r(272) = .169$, $p = .005$, subjective: $r(272) = .484$, $p < .001$. The more participants perceive a decision as objective, the more they relied on algorithmic advice over a non-expert.

I then examined importance, relative reliance on the algorithm, and perceived subjectivity. I compared these measures within the subjective decisions relative to the objective decisions (objective minus subjective ratings). If the effect of the decision's subjectivity is mediated by *perceived* subjectivity, then the difference score of perceived subjectivity should correlate with the difference score of relative reliance on algorithmic advice, controlling for the difference score of importance. Based on these difference scores, perceived subjectivity correlates with relative reliance on algorithmic advice (relative to advice from another person), controlling for importance, $r(272) = .367$, $p < .001$.

Between-subjects mediation analysis. The nature of this experimental design allowed for a follow-up mediation analysis by setting aside the second condition participants saw and running a between-subjects analysis on the first condition participants saw (Critcher & Dunning, 2009).[7] As a more conservative test of mediation, I bootstrapped the indirect effect. See Figure 11. The subjectivity of the decision is a significant predictor of participant's perceived subjectivity of the decision ($\beta = -1.84$, $p < .001$). Perceived subjectivity is a significant predictor of algorithmic reliance ($\beta = .20$, $p < .001$). The effect of the decision's subjectivity on algorithmic reliance is significantly reduced (from $\beta = -.51$, $p < .001$ to $\beta = -.14$, $p = .330$) when participants' perceived subjectivity of the decision was included in the model. Including perceived subjectivity in the model, the amount of variance explained changed from $R^2 = 0.24$ to $R^2 = 0.31$. There is an indirect effect of the subjectivity condition on reliance on algorithmic advice through perceived subjectivity of the decision, ($\beta = .-37$), CI[-.53, -.21]. As the

confidence interval does not include 0, the results suggest a significant indirect effect (MacKinnon et al., 2007).

Discussion

Experiment 6 provides evidence that perceived subjectivity mediates reliance on algorithms. Regardless of subjectivity, participants relied more on an expert than an algorithm. These results help explain why participants in past work preferred a medical diagnosis from a doctor than an algorithm (Promberger & Baron, 2006). Taken together, the results so far suggest that when experts are not available, people are willing to rely on algorithms in objective domains. This suggests that where experts are costly or difficult to access, providing algorithmic advice is useful. One question that remains is whether experts themselves are willing to listen to algorithms in objective domains or if they discount all advice equally.

Experiment 7: Do experts rely on algorithmic advice?

Experiment 7 examined national security experts, people who make forecasts on a daily basis, to understand whether experts, themselves, are willing to rely on algorithmic advice. Experiment 7 tested whether experts are willing to rely on algorithmic advice. Experiments 1A, 1B, 2, 3, and 4 carefully isolated a robust effect of reliance on algorithmic advice and Experiments 5 and 6 examined perceptions of algorithmic advice using a wider range of decisions. The Good Judgment Project (GJP), on which I was a collaborator, aimed to develop better forecasts of geopolitical events through aggregation of individual judgments (Mellers et al., 2014). The question remains as to whether decision makers are willing to listen to advice from algorithms, especially when doing so could improve predictive accuracy.

Participants in Experiment 1B relied less on algorithmic advice for a task they had previously researched, movie profits. Experiment 7 directly tests expertise as a moderator of

reliance on algorithmic advice by comparing lay responses with responses from U.S. Government employees (or contractors) who worked in National Security. I predicted that expertise suppresses reliance on algorithmic advice.

Method

Participants

The final sample included 301 mTurkers (women = 154; men = 147; *M* age = 39) and 70 expert participants (women = 3; men = 67; *M* age = 46) for a total of 371. The results address the gender makeup of the expert sample. I recruited experts from a series of email list-serves dedicated to the topic of national security. Most of these respondents were government employees or consulted with the government. I found a large effect for the averaged geopolitical forecasts from Experiment 1B (d = .60). I estimated a smaller effect (d = .20) for an interaction with repeated measures and thus needed a sample of 200 overall.

To compensate for the uncertainty in the number of experts who might take the survey, I aimed to collect a sample size of 200 mTurkers and at least 75 experts if possible, with the aim of collecting 100. I collected almost as many experts as expected but stopped collecting data when multiple days passed without new participants. I needed to balance the goal of recruiting more participants with collecting their forecasts around the same time. Participants forecasted the probability of events occurring before the end of 2016, and to avoid any unfair advantage of forecasting closer to the date in question, I attempted to collect data within a reasonable time constraint and before the presidential election.

Design

The experiment had a 2 (advisor: human forecasters vs. algorithm) X 2 (sample: lay vs. expert) design. As in prior experiments, the experimental manipulation varied whether

participants saw advice labeled from an algorithm or from human advisors (forecasters where relevant). This experiment used a similar paradigm to Experiment 1B and used the same main dependent variable, WOA (weighting of advice), for four tasks: a weight estimate, business forecast, and two geopolitical forecasts.

Procedure and Materials

Overview. I used a similar procedure to Experiment 1B but included new forecasts and additional measures. See Table 4. I chose the tasks to amplify the difference in expertise between samples. I expected the experts to feel greater expertise for the geopolitical forecasts than the lay people and for both samples to experience similarly low expertise for the weight estimate (the same photograph as Experiment 1A) and business forecast. The business forecast provided a domain where experts and lay people both have little expertise, while controlling for the type of task (a forecast about a probability that is not currently knowable rather than an estimate about a fact that is currently knowable).

I took forecasts from Good Judgment Open tournament (the tournament created after funding ended for the GJP). The experiment held the order of tasks constant in order to replicate past experimental materials, where participants estimated the person's weight first.

Advisor manipulation. Prior to their second estimate, all participants received the same advice (163 pounds for the weight estimate, 20% for the business forecast (Tesla), 12% for the first geopolitical forecast (cyber), and 45% for the second geopolitical forecast (Brexit) described as an estimate from either another person or an algorithm.

In the human condition for the weight estimate, participants read, "The estimate of another person was: 163 pounds."

For the three forecasts, I purposely provided information about the tournament, so that lay people and experts had the same human reference. I wanted to avoid lay people and experts responding to advice from their different peer groups. Participants in the human condition read, "The average estimate from forecasters in the forecasting tournament is: X."

In the algorithm condition for all tasks, participants read, "The estimate from an algorithm is: X."

Main Dependent Measure

Weight of advice (WOA). I measured how much participants relied on the advice.

Additional Measure: Control Variable

Familiarity. I controlled for a possible difference between samples in familiarity with algorithms. Participants reported their familiarity with the word algorithm, "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*.

Exclusion Measures

In addition to the exclusion measures utilized in the past experiments, I excluded participants with prior exposure to the tournament (where they may have seen the same forecast) and those who failed to follow directions by reporting that they sought out information on the Internet. These important exclusions helped control for the amount of information participants accessed to make the forecasts (through links to information provided with the forecasts).

Prior participation in tournament. "Prior to this survey, have you participated as a forecaster in the Good Judgment Project or Good Judgment Open?" (*yes/no/unsure*)

External information search. "Did you search the internet or use other sources to find more information for the forecasts (beyond clicking the links provided)? Clicking the links provided does not count as searching the internet." (*yes/no*)

Additional Measures: Mechanism

Forecasting experience. Participants reported how frequently they made forecasts for their occupation, "For your job, how often do you make forecasts (predictions)?" on a scale from *1 = Barely ever to 7 = Multiple times a day*.

Felt expertise. Participants reported their expertise on each topic, "How much do you know about each topic in the survey?" on a scale from 1 = *very little* to 7 = *a lot*.

Exploratory measure

To better understand the types of decisions people tend to make and how they normally inform those decisions, especially experts, I included a few exploratory measures.

Types of decisions. Participants described, "What kinds of decisions do you make at work that depend on forecasts?" in an open-ended form.

Default advisor. Participants reported their default advisor, "At work, which source normally provides you with forecasts? (*Another person, a group of people, an algorithm or statistical model, I normally provide the forecasts, other*)

Reliance on default source. Participants compared an algorithmic advisor to their default advisor, "If you were given the option, how much would you rely on advice from an algorithm or your usual source, in order to make the types decisions you listed?" on a scale from 1 = *completely rely on algorithm* to 7 = *completely rely on usual source*.

Results

Forecasting experience. Did lay people and experts differ in their forecasting experience? The expert sample reported that they made 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.

Felt expertise. To understand how much expertise participants thought they brought to each task, I submitted reported expertise on each task to a 2 (expert vs. non-expert) X 4 (weight estimate vs. tesla forecast vs. cyber forecast vs. Brexit forecast) repeated measures ANOVA with task as a repeated measure.[8] There is a main effect of expertise, $F(1, 369) = 11.88$, $p = .001$, a main effect of task, $F(3, 1107) = 21.78$, $p < .001$, and an interaction, $F(3, 1107) = 34.35$, $p < .001$. Figure 12 shows that across tasks, experts reported greater expertise with each topic than lay people. As predicted, experts reported greater expertise both for the cyber, $F(1, 369) = 9.48$, $p < .001$, and Brexit forecast, $F(1, 369) = 30.02$, $p < .001$. The samples reported similar expertise for the Telsa forecast, $F(1, 369) = .39$, $p = .535$. Although I predicted that lay people and experts would report similar expertise for the weight estimate, lay people reported greater expertise (M = 3.67) than the experts (M = 3.01), $F(1, 369) = 9.48$, $p = .002$.

Weighting of Advice (WOA). I submitted WOA to a 2 (advisor: algorithm vs. human) X 2 (expertise: expert vs. non-expert) repeated measures ANCOVA with the four tasks as repeated measures and familiarity with the term algorithm as a covariate. This analysis included those who answered all tasks (282 lay people and 61 experts). There is an effect of familiarity, $F(1, 338) = 8.86$, $p = .003$, $\eta^2 = 0.02$, $d = 0.29$. Controlling for familiarity, there are main effects of advisor, $F(1, 338) = 9.46$, $p = .002$, $\eta^2 = 0.02$, $d = 0.29$, and expertise, $F(1, 338) = 32.39$, $p < .001$, $\eta^2 = 0.08$, $d = 0.60$, as well as an interaction, $F(1, 338) = 5.05$, $p = .025$, $\eta^2 = 0.01$, $d = 0.23$.[9] Figure 13 shows that for each task, lay people relied more on advice from an algorithm

than from other people, while Figure 14 shows that experts heavily discount all advice. I found virtually no evidence for mediation by reported felt expertise or forecasting experience.[10]

Familiarity. Although there was variance in people's familiarity with the term algorithm, I wanted to test how certain that they knew what it meant, on average. Across samples, reported familiarity differed from the mid-point of the scale (4), (M = 5.04, SD = 1.72), $t(370)$ = 17.30, $p < .001$. This held by expertise (Lay: M = 4.96, SD = 1.73), $t(300) = 9.61$, $p < .001$, (Experts: M = 5.40, SD = 1.63), $t(69) = 7.20$, $p < .001$.

Usual advisors. To take full advantage of the unique expert sample, I examined the types sources people used to inform their decisions. Both samples tended to rely on a group of people. See Table 5. Experts may show hesitation to rely advice because more than a third of them reported normally providing forecasts at work.[11]

Interestingly, when participants directly compared their usual source of information at work to an algorithm, participants preferred their usual advisor to an algorithm to inform their work decisions (M = 4.55, SD = 1.33, relative to the mid-point of the scale), $t(339) = 7.57$, $p <$ .001. This pattern held for both lay people, (M = 4.56, SD = 1.40), $t(273)$ 6.79, $p < .001$, and experts, (M = 4.47, SD = 1.13), $t(65) = 3.39$, $p = .001$. Surprisingly, forecasting expertise does not correlate with greater reliance on the default source, $r(340) = .01$, $p = .910$. A higher percentage of experts report providing forecasts that inform decision, which could affect their unwillingness to listen to advice, especially from a new advisor. These results suggest that for decisions where people have a default advisor for their jobs, they may show greater hesitation to rely on a new source, such as an algorithm.

As an exploratory, non-registered analysis, reporting one's self as the default advisor at work only affected reliance on advice for the Brexit forecast, $t(363) = -2.11$, $p = .036$, and not for

any other tasks (*t*s < -1.45, *p*s > .13). People who reported that they were the advisor relied less on the advice (*M* = .30, *SD* = .37) than people who did not (*M* = .39, *SD* = .38).

Age and Gender. The sample included a diversity of ages (ranging from 18 to 75 in the non-expert sample and from 25 to 88 in the expert sample). Although age did not correlate with reliance on algorithmic advice in Experiment 1A, the expert sample in Experiment 7 was recruited from a different participant pool, so I ran an additional, non-registered analyses and correlated reliance on algorithmic advice with age. In the expert sample, age does not correlate with reliance on algorithmic advice for any tasks (*p*s > .21). Combining the samples, the same was true (*p*s > .09) except for the business forecast (*r* = -.16, *p* = .030), where younger participants relied more on algorithmic advice. These results suggest that age has little to no relationship with reliance on algorithmic advice. I also checked (non-registered) whether gender correlated with reliance on either human or algorithmic advice because the expert sample was mostly men. Gender neither correlates with reliance on algorithmic (*p*s > .07) nor human advice (*p*s > .08).

Discussion

Experiment 7 shows that experts do not respond to algorithms in the same way that lay people do. In fact, experts heavily discounted advice, regardless of the advisor. Figure 14 shows heavier discounting than what is usually found in advice taking work (a 30% discounting rate on average). It is worth noting that the 30% average discount rate is driven by a trimodal distribution (0, 50 and 100%), rather than people consistently displaying heavy discounting (Soll & Larrick, 2009). The results from Experiment 7 help explain why pilots and doctors tend to ignore statistical advice while making decisions. Although providing advice from algorithms

may increase adherence to advice for non-experts, it seems that algorithmic advice falls on deaf expert ears.

General Discussion

Counter to the widespread conclusion that people are averse to algorithms, results from eight experiments suggest that people are willing to rely on algorithmic advice under circumstances that apply to many decisions. More importantly, they highlight contexts in which people are most likely to improve their accuracy by using advice generated by algorithms. When making estimates and forecasts, participants relied more on advice when they thought it came from an algorithm than from a person (Experiments 1A and 1B). One mechanism for this effect is numeracy: people with higher mathematical capabilities relied more on algorithms (Experiment 1A).

Reliance on algorithmic advice appears robust; it is resilient to how the algorithmic and human advisor are presented, either jointly or separately (Experiment 2). Algorithmic advice seems to partly help people overcome overconfidence; people are willing to choose an algorithmic estimate over their own. Although overconfidence reduces reliance on algorithms, it does not suppress or reverse it (Experiment 3). As much as people are willing to listen to algorithms, they still have room to improve their accuracy. Participants underweighted algorithmic advice more than they should have when they received normative information about the advice (Experiment 4).

Convergent evidence of moderation and mediation suggest a second mechanism: subjectivity of the decision context (Experiments 5 and 6). For objective decisions (made based on logic or reason), participants preferred algorithmic advice. For subjective decisions (made based on emotion or intuition), they preferred advice from people. Expertise of the advisor as

well as the expertise of the decision maker both appear to moderate the effect. Regardless of subjectivity, people preferred the expert advice to algorithmic advice (Experiment 6). Decision makers' own expertise also moderated reliance on algorithms. National security experts relied less on algorithmic advice than lay people did, heavily discounting any advice they received.

In sum, people *are* willing to rely on advice more from algorithms than other people in certain contexts. People are more to listen to algorithmic advice 1) the more comfortable they are with mathematics and 2) when they are making more objective decisions. This robust effect is moderated by overconfidence, the availability of an expert advisor, and the decision-maker's own expertise. These results shed light on when people are most likely to improve their accuracy by listening to algorithms and has implications for the use of algorithms by individual decision makers and decision makers within organizations.

Theoretical Implications

These results introduce an interesting contrast to the widespread assumption that people are averse to algorithms (Bazerman, 1985; Dawes, 1979; Dawes, Faust, & Meehl, 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, these results help reconcile the contradictory work that finds aversion under some conditions and reliance under others. See Table 6.

The results from this paper suggest that people will rely on algorithms more than another person in an objective domain. When there is a fair comparison, algorithms win. When the boundaries of those conditions change, so does reliance on algorithmic advice. When decision makers directly compare their knowledge with an algorithm's (allowing for overconfidence), rather than comparing between human and algorithmic advisor, reliance weakens. In fact, the

effect reverses with experts who prefer their own judgment, regardless of the advisor. People also prefer advice from other people for subjective decisions and when an expert advisor is available. Work testing accuracy declared algorithms the winner over experts (Beck et al., 2011; Carroll et al., 1982, Einhorn, 1972; Goldman, et al., 1977; Hedén et al., 1997), which suggests that reliance on expert advice may encourage suboptimal judgments.

Although the results of this paper may appear to most directly contradict the work from Dietvorst et al. (2015), a closer examination of their results suggest that they are compatible with the results from this paper. Although not the focus of the 2015 paper, prior to learning about the inaccuracy of an algorithm, participants relied on the algorithm. In one instance, they showed indifference between their answer and the algorithm's (the materials used in Experiment 3 of this paper). Yet, none of their studies found algorithm aversion prior to participants receiving evidence of a potentially broken algorithm. As discussed in the introduction, consequential decisions are often made based on forecasts with long time horizons. Thus, work focusing on perceptions of algorithms prior to repeated interactions is useful.

These results likewise hold implications for the advice-taking literature. One robust finding is that people discount advice from others (Bonaccio & Dalal, 2006; Yaniv & Kleinberger, 2000; Yaniv, 2004). The results presented in this paper suggest that one simple way to increase adherence to advice in situations where it is heavily discounted is to provide advice from an algorithm. An algorithmic advisor helped participants in Experiment 3 overcome their overconfidence; they relied more on the algorithm than themselves. This should prove especially useful for objective decisions when expert advice is unavailable. Many decisions are ripe for the use of algorithmic advising considering the low cost and widespread availability of algorithmic advice relative to expert advice.

Practical Implications

Understanding how people respond to algorithmic advice holds implications for any decision maker or organization that has the potential to learn from lessons produced by "big data." As organizations invest in the collection, analysis and exploitation of big data, they use algorithms to sift through the information and produce algorithmic advice. Advances in technology allow organizations to better access and analyze big data, which has lead them to collect greater amounts and variety of data, and all at a faster rate (Laney, 2012). Access to big data is opening up as well. Google (as used in Reips & Matzat, 2014) and Twitter (as used in Reips & Garaizar, 2011) already share enormous quantities of data publicly. Business organizations are not the only players cued in to big data. The Obama administration launched the Big Data and Research Initiative to improve aspects of government including national security (Kalil, 2012).

To improve the communication of information within any organization, one cannot focus solely on the accuracy of the information delivered but on how the receiver perceives it (Chan, 1979). For instance, the benefits of algorithms to more accurately forecast future political events (Baron et al., 2014) should have the U.S. Intelligence Community (IC) and policy makers alike jumping to take advantage of algorithms for improved policy decisions. However, the results from Experiment 7 suggest that even if algorithms produce more accurate information than experts, algorithmic advice falls on deaf ears. The results from this paper provide a more hopeful picture for lay people. Future work should examine how to increase experts' reliance on algorithms. The results in this paper shed light on how to best make use of efforts to present algorithmic advice from big data by testing when people are most likely to listen.

Limitations and Future Directions

Expert Advice. This work has only scratched the surface of what research on "Theory of Machine" can examine. It presents the most basic test of how people respond to algorithmic advice. I used simple tests designed to avoid confounds and to identify moderators in order to reconcile prior conflicting research. In order to avoid confounding expertise and human advice, Experiments 1 through 5 compared perceptions of algorithmic advice and advice from a peer. Experiment 7 provided advice from a forecaster in the forecasting tournament so that both the lay and expert sample had the same human comparison (rather than from a peer of the national security specialist and a peer on mTurk). However, a potential limitation to these studies is that people may sometimes find themselves in situations where they compare expert and algorithmic advice. In fact, Experiment 5 showed that when available, participants relied more on experts than algorithmic advice.

Although future work should compare expert to algorithmic advice where the comparison occurs in the real-world, experts are not the only source of human advice. Often, people seek out advice from friends and family members, who are clearly not experts. For instance, when things are broken, from hearts to home appliances, people call friends, partners, or parents for potential solutions. As discussed in the introduction, expert advice is not always a realistic comparison to algorithmic advice. We do not always call a therapist or a plumber at the first sign of a problem, simply because access to these experts is painfully costly and often delayed. Algorithmic advice is often cheaper and faster. In fact, the limited accessibility of experts has opened the door for a chatbot that provides free legal aid to refugees seeking asylum in the U.S. and Canada (Cresci, 2017). The chatbot asks the user questions in order to fill out the appropriate asylum application, as they are often difficult to understand.

Another constraint to comparing expert advice to algorithmic advice is the normative usage of expert advice. It is unclear how much participants *should* weight advice from an expert. A normative standard for a random individual or judgment based on multiple individual judges is clear. A normative standard for a single expert is not. Future work should consider expert advice but remain aware of its boundaries.

Theory of Machine: Input, Process, Output. There is much more work needed to examine how people expect algorithmic and human judgment to differ beyond their advice, or output. Future work on Theory of Machine should examine lay perceptions of how algorithms and humans differ in their *input*, *process*, and *output*. Understanding how people expect algorithms and humans to differ will allow examination of how algorithmic advisors, at their finest, compare to human advisors.

Do people assume that algorithmic and human judgment utilize different *input* when providing advice? One reason why people may rely on algorithmic advice is that they assume an algorithm often has access to more information than a person. In fact, a criterion that determines how much people rely on advice is the advisor's access to information (Birnbaum & Stegner, 1979; Birnbaum, Wong, & Wong, 1976; Budescu, Rantilla, Yu, & Karelitz, 2003; Sniezek & Buckley, 1995). Experiment 4 suggests that people are less sensitive to the number of judgments used to construct advice than they should be. Yet, people may still assume that an algorithm and person have access to vastly different magnitudes of data, beyond the number of judgments included in data based on aggregation. Likewise, perceptions about the *source* of the input to the advice may also differ. People may expect an algorithm to have access to less personalized information (data about other people) and other people to have access to more personalized information (data about them specifically). Instances in which algorithms do have personalized

information may provoke backlash due to privacy concerns about the personal data collected to inform the algorithm. These assumptions should influence when people prefer algorithmic to human advice.

Do people assume that algorithmic and human judgment *process* information differently? People may think that algorithms are more constrained in their processing of information. They may expect other people to process the same information as algorithms do in a more holistic way. These assumptions lead to interesting predictions about how people will respond to advice from either source. So for a medical diagnosis, if both an algorithm and person have access to only five cues related to a patient's cancer biopsy, participants may rely more on an algorithm because they expect it to process that piece of data more effectively or consistently. However, when the patient's entire medical history is available, participants may rely more on the person because they expect the physician will take a more holistic view. Participants may prefer a human advisor when advisors have access to lots of information. Ironically, the reason why algorithms provide more accurate forecasts than people is because they are not distracted by unhelpful cues and thus weigh cues more appropriately (Hedén et al., 1997). People may see algorithmic processing as efficient within environments of informational scarcity but less effective than a holistic approach when informational resources grow.

There are many more questions to ask about how people perceive algorithmic and human *output*, or advice. People may view an algorithm as a black box, either because they do not have access to or do not understand the mathematics or procedures of the algorithm. Future work could manipulate the opacity of an algorithm's process and whether people can understand it by manipulating the complexity of algorithmic calculations through the notation (see Supplementary Materials for a "file drawer" experiment on the topic of transparency). If people can

comprehend the calculations or procedures of an algorithm, they may rely on it more because they can understand it. Work on acceptance of climate change suggests that this is the case. It shows that acceptance of climate change increases when participants are provided mechanistic information about how global warming works (Ranney & Clark, 2016). The opposite could also occur; people may rely on an algorithm less when understanding the calculations means that they are better able to consider its potential flaws. Another possible outcome is that when people access an algorithm's calculations, they rely on the algorithm, even if they cannot comprehend the calculations. They may feel less knowledgeable when they view a complex calculation or procedural rules and infer that an algorithm uses a superior process to a person.

Conclusion

Big data are changing the way organizations function and communicate, both internally and with their clients. Organizations use algorithms to hire people (Ritchel, 2013), fire people, manage employees' priorities (Copeland & Hope, 2016), and help employees choose health care plans (Silverman, 2015). Organizations like Stitch Fix provide clients with clothing recommendations based on algorithms (Ahuja, 2015; Hu, 2014). Information technologies may increase exposure to big data but how do big data, 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 big data and ignore how people respond to the algorithmic advice produced by it, they will not fully capture that opportunity to learn.

Without understanding how people incorporate information from algorithmic advice into their decisions, organizations are spending precious resources to produce output that is not fully utilized. Understanding where people focus their attention within the deluge of information that

floods email inboxes, Internet searches, and discussions is useful for any organization or decision maker. The results of eight experiments shed light on how to best leverage algorithmic advice. Uncovering the mechanisms that encourage reliance on algorithmic over human advice, through research on "Theory of Machine," will help decision makers transition more effectively to the world of big data.

REFERENCES

Ahuja, S. (2015, May 26). 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.

Ask the Algorithm. (2015, May 9). *The Economist*. Retrieved from http://www.economist.com/news/special-report/21650292-human-wealth-advisers-are-going-out-fashion-ask-algorithm

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*. *11*(2), 133-145. doi:10.1287/deca.2014.0293

Bazerman, M. H. (1985). Norms of distributive justice in interest arbitration. *Industrial and Labor Relations Review, 38*, 558-570. doi: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.

Beck, A. H., Sangoi, A. R., Leung, S., Marinelli, R. J., Nielsen, T. O., van de Vijver, M. J., & Koller, D. (2011). Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science Translational Medicine, 3*(108), 108-113. doi:10.1126/scitranslmed.3002564

Birnbaum, M. H., Stegner, S. E. (1979). Source credibility in social judgment: Bias, expertise, and the judge's point of view. *Journal of Personality and Social Psychology, 37*(1),48-74.

Birnbaum, M. H., Wong, R., & Wong, L. K. (1976). Combining information from sources that vary in credibility. *Memory & Cognition, 4*(3), 330-336.

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.

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin, 86*(2), 307-324. doi:10.1037/0033-2909.86.2.307

Budescu, D. V., Rantilla, A. K., Yu, H. T., & Karelitz, T. M. (2003). The effects of asymmetry among advisors on the aggregation of their opinions. *Organizational Behavior and Human Decision Processes*, *90*(1), 178-194.

Carroll, J. S., Wiener, R. L., Coates, D., Galegher, J., & Alibrio, J. J. (1982). Evaluation, diagnosis, and prediction in parole decision making. *Law and Society Review*, 199-228. Chicago.

Chabris, C. (2015) Edge: What do you think about machines that think? [Web log post]. Retrieved from https://www.edge.org/response-detail/26224

Chan, S. (1979). The intelligence of stupidity: understanding failures in strategic warning. *The American Political Science Review*, *73*(1), 171-180. doi:10.2307/1954739

Copeland, R., Hope, B. (2016, December 22) The World's Largest Hedge Fund Is Building an Algorithmic Model From its Employees' Brains. The Wall Street Journal, Retrieved from https://www.wsj.com/

Cresci, E. (2017, March 6) Chatbot that overturned 160,000 parking fines now helping refugees claim asylum. The Guardian, Retrieved from https://www.theguardian.com

Critcher, C. R., & Dunning, D. (2009). Egocentric pattern projection: how implicit personality theories recapitulate the geography of the self. *Journal of personality and social psychology*, *97*(1), 1.

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

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668-1674. doi: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.

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

Dijkstra, J. J., Liebrand, W. B., & Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour & Information Technology*, *17*(3), 155-163. doi: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. doi:10.1518/0018720024494856

Einhorn, H. J. (1972) Expert measurement and mechanical combination. *Organizational Behavior and Human Performance*, *7*(1), 86-106. doi:10.1016/0030-5073(72)90009-8

Einhorn, H. J., Hogarth, R. M. Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin*, *84*(1), 158–172.

Finkel, E. J., Eastwick, P. W., Karney, B. R., Reis, H. T., & Sprecher, S. (2012). Online dating: A critical analysis from the perspective of psychological science. *Psychological Science in the Public Interest*, *13*(1), 3-66.

Frey, C. B., & Osborne, M. A. (2013). The future of employment: how susceptible are jobs to computerization. Retrieved April 2014.

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

Gardner, P. H. and Berry, D. C. (1995). The effect of different forms of advice on the control of a simulated complex system. *Applied Cognitive Psychology*, *9*, S55–S79. doi: 10.1002/acp.2350090706

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

Goldman, L., Caldera, D. L., Nussbaum, S. R., Southwick, F. S., Krogstad, D., Murray, B. & Slater, E. E. (1977). Multifactorial index of cardiac risk in noncardiac surgical procedures. *New England Journal of Medicine*, *297*(16), 845-850. doi:10.1056/nejm197710202971601

Hartford, T. (2015, September 14). How to see into the future. *Financial Times.* Retrieved from http://www.ft.com/

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.

Harvey, O. J., White, B. J., Hood, W. R., & Sherif, C. W. (1961). *Intergroup conflict and cooperation: The Robbers Cave experiment*, *10*, Norman, OK: University Book Exchange.

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review*, *112*(2), 494-508. doi:10.1037/0033-295x.112.2.494

Hedén, B., Öhlin, H., Rittner, R., & Edenbrandt, L. (1997). Acute myocardial infarction detected in the 12-lead ECG by artificial neural networks. *Circulation*, *96*(6), 1798-1802. doi:10.1161/01.cir.96.6.1798

Hodges, A. (2014) *Alan Turing: The Enigma*. Princeton, NJ: Princeton University Press.

Hu, E. (2014, September 8). Try This On For Size: Personal Styling That Comes In The Mail [Audio file]. *National Public Radio: All Tech Considered*. http://www.npr.org/

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.

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. (1998). Less is better: When low-value options are valued more highly than high-value options. *Journal of Behavioral Decision Making*, *11*.

Hsee, C. K., Zhang, J., Yu, F., & Xi, Y. (2003). Lay rationalism and inconsistency between predicted experience and decision. *Journal of Behavioral Decision Making*, *16*(4), 257-272. doi:10.1002/bdm.445

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

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

Judd, C. M., Kenny, D. A., & McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subject designs. *Psychological Methods*, *6*, 115–134.

Kalil, T. (2012, March 29) Big data is a big deal [Web log post]. Retrieved from http://www.whitehouse.gov/

Keeffe, B., Subramanian, U., Tierney, W. M., Udris, E., Willems, J., McDonell, M., & Fihn, S. D. (2005). Provider response to computer-based care suggestions for chronic heart failure. *Medical Care*, *43*(5), 461-465. doi:10.1097/01.mlr.0000160378.53326.f3

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

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

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

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*(1), 111-127. doi:10.1287/mnsc.1050.0459

Lichtenstein, S., & Slovic, P. (2006). *The Construction of Preference*. Cambridge, MA: Cambridge University Press.

Lynch, J. (2016, April 24). Is Predictive Policing the Law-Enforcement Tactic of the Future? *The Wall Street Journal*. Retrieved from http://www.wsj.com/

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. doi:10.1037/h0047554

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., & Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological science*, *25*(5), 1106-1115. DOI: 10.1177/0956797614524255

Miller, C. (2015, June 25). Can an algorithm hire better than a human? *The New York Times*. Retrieved from http://www.nytimes.com/

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502-517. doi: 10.1037/0033-295X.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. (2016). Overprecision in judgment. In G. Wu and G. Keren (Eds.), *Handbook of Judgment and Decision Making*. New York: Wiley.

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.

Ranney, M.A. & Clark, D. (2016). Climate change conceptual change: Scientific information can transform attitudes. *Topics in Cognitie Science*. *8*, 49-75. DOI: 10.1111/tops.12187

Reips, U. D., & Garaizar, P. (2011). Mining twitter: A source for psychological wisdom of the crowds. *Behavior research methods*, *43*(3), 635-642. doi: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. Retrieved from http://www.ijis.net/

Ritchel, M. (2013, April 27) How big data is playing recruiter for specialized workers. *New York Times*. Retrieved from: http://www.nytimes.com/

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. doi:10.7326/0003-4819-127-11-199712010-00003

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

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

Silverman. R. (2015, November 10). Picking a health plan? An Algorithm could help. *New York Times.* Retrieved from: http://www.nytimes.com/

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. Available at SSRN 2160588.

Simonsohn, U. (2014, March 12) ns [Web log post]. Retrieved from http://datacolada.org/

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. doi:10.1006/obhd.1995.1040

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 299.

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. doi:10.1037/a0015145

Soll, J. B., Milkman, K. L., & Payne, J. W. (2016). A user's guide to debiasing. In G. Wu & G. Keren (Eds.), *Handbook of Judgment and Decision Making*. New York: Wiley.

Sparaco, P. (2006, April 10). Safety First, Always. *Aviation Week & Space Technology*. Retrieved from http://en.wikipedia.org

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.

Steiner, C. (2012). *Automate This: How Algorithms Came to Rule Our World*. United Kindgom: Penguin Group.

Sumner, W. J. Graham. (1906). *Folkways: A Experiment of the social importance of usages, manners, customs, mores, and morals*. Boston, MA: Ginn.

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.

Tazelaar, F., & Snijders, C. (2013). Operational risk assessments by supply chain professionals: Process and performance. *Journal of Operations Management, 31*(1), 37-51.

Tesauro, G., Gondek, D., Lenchner, J., Fan, J., & Prager, J. M. (2013). Analysis of watson's strategies for playing Jeopardy!. *Journal of Artificial Intelligence Research, 47*, 205-251. doi:10.1613/jair.3834

Ungar, L., Mellors, B., Satopää, V., Baron, J., Tetlock, P., Ramos, J., & Swift, S. (2012). *The good judgment project: A large scale test*. AAAI Technical Report. (FS-12-06).

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.

Weber, E. U., & Lindemann, P. G. (2008). From intuition to analysis: Making decisions with our head, our heart, or by the book. In H. Plessner, C. Betsch, & T. Betsch (Eds.), *Intuition in judgment and decision making*, (pp. 191-208) New York, NY: Lawrence Erlbaum Associates, Taylor and Francis Group.

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

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.

Yaniv, I. (2004). The benefit of additional opinions. *Current directions in psychological science, 13*(2), 75-78.

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2017). *Making sense of recommendations*. Unpublished manuscript. Retrieved from: http://scholar.harvard.edu/files/sendhil/files/recommenders55.pdf

FOOTNOTES

[1] Although a blogpost argued that relative to theory of mind for other people, individuals, “don't have a correspondingly intuitive "Theory of Machine"” (Chabris, 2015), this current paper argues that as the tsunami of big data grows, modern people *are currently constructing* a theory of machine.

[2] I paid the first ten participants $0.35 because I thought the task would take longer than it did. I re-posted for ten cents less with a shorter time listed.

[3] When the geopolitical forecasts are not averaged, results hold for the first forecast: People rely more on advice when it comes from an algorithm (M = .54, SD = .37) than another person (M = .27, SD = .33), t(72) 3.38, p = .001, d = .77. For the second forecast, results are in the same direction but do not reach significance, perhaps because of the sample size. People rely more on advice when it comes from an algorithm (M = .59, SD = .42) than another person (M = .42, SD = .38), t(74) 1.78, p = .80, d = .42. Participants may have become tired by the fourth forecast, reflected in some open-ended response at the end of the survey that it was long and they were not interested in the forecast topic.

[4] The first exploratory measure asked participants if they had experience with inaccurate predictions from Microsoft’s age guessing algorithm (how-old.net). Too few participants reported using the algorithm and remembering if it provided accurate results (597 reported that they had not used it at all), so I could not use that measure as a covariate.

[5] I had pre-registered an ANOVA analysis, but determined that a paired t-test as more appropriate prior to analyzing data.

[6] Without controlling for importance, there is a main effect of subjectivity such that participants relied more on algorithmic advice than human advice for the objective decisions than subjective decisions, F(1, 548) = 224.28, p < .001. There is an interaction between subjectivity and expertise, F(1, 548) = 120.12, p < .001. In the non-expert condition, participants relied more on algorithmic advice in the objective than the subjective condition, F(1, 548) = 336.32, p < .001. Yet, people prefer the expert more in the subjective than objective condition, F(1, 548) = 8.01, p = .005 (unlike people preferring the expert equally across the conditions when controlling for importance).

[7] An additional analysis with just the first subjectivity condition showed results redundant with the reported ANCOVA. Controlling for importance, people relied more on the algorithm in the objective condition (M = 3.71, SD = 1.20) than in the subjective condition (M = 3.13, SD = 1.20), F(1, 550) = 15.68, p < .001. Unlike in the ANCOVA, importance itself is not a significant predictor itself, F(1, 550) = 1.85, p = .174. Again, there is an interaction between subjectivity and expertise, F(1, 550) = 14.01, p < .001. In the non-expert condition, participants relied more on algorithmic advice in the objective (M = 4.06, SD = 1.15) than subjective condition (M = 3.11, SD = 1.27), F(1, 550) = 29.95, p < .001. In the expert condition, participants relied more on the expert, regardless of subjectivity (objective: M = 3.34, SD = 1.20, subjective: M = 3.15, SD = 1.23), F(1, 550) = .43, p = .513.

[8] I had pre-registered a primary comparison between the weight estimate and collapsed forecasts, with a secondary analyses comparing across individual forecasts. The results in this section focus on the “secondary analyses.” They are more appropriate to test my prediction which was that participants feel less expertise for both the weight estimate and business forecast. Comparing the business forecast to the geopolitical forecasts is useful. These results as well as any other pre-registered analyses are reported in

the Supplementary Materials. I included more measures in this experiment relative to the others because of the difficulty recruiting such a special sample of experts.

Collapsing across forecasts, there is a main effect of task, such that across samples, people felt they had more expertise for the weight estimate (M = 3.55, SD = 1.63) than for the forecasts (M = 2.66, SD = 1.32), $F(1, 369) = 11.66$, $p = .001$. Although I predicted that lay people and experts report similar expertise for the weight estimate, simple effects show that lay people reported greater expertise (M = 3.67) than the experts (M = 3.01), $p = .002$. But as expected, collapsing across forecasts, the experts (M = 3.40) reported greater expertise than the lay sample (M = 2.49), $p < .001$. Thus, expertise interacted with task, $F(1, 369) = 44.90$, $p < .001$.

[9] When familiarity is not included in the model, results remain with the exception that the interaction between task and expertise becomes non-significant, $F(1, 339) = 3.83$, $p = .051$, $\eta^2 = .010$, $d = .201$. The change in the interaction could be due to a smaller sample size for the expert sample (N = 70), but a Fisher r-to-z transformation shows that the effect sizes of the interactions (including and excluding familiarity as a covariate) do not differ: $z = -.18$, $p = .857$. I found an effect of advisor, $F(1, 339) = 8.88$, $p = .003$, and expertise, $F(1, 339) = 35.85$, $p < .001$.

While experts are more familiar with the term algorithm (M = 5.4, SD = 1.63) than lay people are (M = 4.96, SD = 1.73), $t(369) = 1.96$, $p = .051$, including familiarity as a control helps account for variance in the model.

[10] Neither felt expertise nor forecasting experience acted as mediators between expertise and WOA on any of the tasks, with either all conditions or just those who saw algorithmic advice (the confidence intervals of the indirect effects all include zero). The only exception was in the algorithm condition for the Brexit forecast: there is a significant indirect effect of felt expertise on algorithm reliance through perceived forecasting experience, $ab = .05$, BCa CI [-.10, -.01]. Yet, the mediator accounts for very little of the total effect, $Pm = .15$.

[11] Although I did not pre-register splitting the responses to the "usual source" measure by expertise, it seemed useful to parse out answers in order to better understand the psychology of the expert's decision-making routine.

Table 1

Wording for each task in Experiment 1B

Task and Type	Wording
Weight Guessing (Estimation)	How many pounds does this person weigh?
Longest Ride (Movie Forecast)	How much will The Longest Ride gross over the opening weekend?
HSBC (Geopolitical Forecast)	What is the probability that the HSBC China Services Purchasing Managers' Index will fall to 50.0 or below before 1 June 2015?
SWIFT (Geopolitical Forecast)	What is the probability that before 17 June 2015, SWIFT will restrict any Russian banks from accessing its services?

Table 2

Participant-generated definitions of algorithm

Category of Definition	**Example Definition (Experiment 1B and Experiment 2)**	**% 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 and (sic) consistently. “	14%
Computer	“A series of formulas that will generate an output with the correct inputs. Usually used on the computer.”	.09%
Other	“a way to solve a problem repetitively”	.06%
Predictive Data Analysis	“A model to predict a result based on incomplete data.”	.04%

Table 3

Decision problems and expert advisors in experimental materials for Experiment 6

Decision Problem	Subjectivity	Expert Advisor
Which stock to invest in for retirement	Objective	Financial Advisor
Which surgery to undergo	Objective	Doctor
Which credit card to apply for	Objective	Financial Advisor
How to create your budget for the year	Objective	Accountant
Which job offer to accept	Objective	Career Counselor
Which neighborhood to move to	Objective	Realtor
Which book to buy	Subjective	Book Critic
Whether to get married	Subjective	Marriage Counselor
Which shirt to buy	Subjective	Clothing Stylist
Which joke to use in speech at work function	Subjective	Speechwriter
Whether to end a relationship	Subjective	Therapist
Who to date	Subjective	Professional Matchmaker

Table 4

Wording for each task and predicted difference in expertise in Experiment 7

Predicted difference in expertise between samples	Wording
Weight (Estimation) *Similar Expertise*	How many pounds does this person weigh?
Tesla (Business Forecast) *Similar Expertise*	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?
Cyber (Geopolitical Forecast) *Expert > Lay*	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?
Brexit (Geopolitical Forecast) *Expert > Lay*	What is the probability that the United Kingdom will invoke Article 50 of the **Lisbon Treaty** before July 1, 2017?

Table 5

Percentage of participants reporting each source as their usual source of information for making decisions

Default Source	Percentage of Total Sample	Percentage of Lay People	Percentage of Experts
Group of People	43	36	74
I normally provide the forecasts	32	30	37
Algorithm or statistical model	19	19	17
Other (Including N/A)	16	30	19
Another person	11	28	33

Note: Column does not add to 100 because participants were allowed to mark multiple answers.

Table 6

Moderators in current experiments reconcile prior contradictory results

	Self vs. Algorithm	**Other vs. Algorithm**	**Expert vs. Algorithm**
Subjective domains of individual preferences	*Rely On: Self* Common Sense et al.?	*Rely On: Other Person* Sinha et al., 2001 Yeomans, et al., 2017	*Rely On: Expert* Not previously tested
Objective domains where a standard of accuracy exists	*Rely On: Self* Dietvorst, et al., 2015 Dzindolet, et al., 2002 Keeffe et al., 2005	*Rely On: Algorithm* Dijkstra, et al., 1998* Dijkstra, 1999*	*Rely On: Expert* Promberger & Baron, 2006

Note: *Potentially confounds algorithm with expertise by calling algorithm an “expert system”

Figure 1. The photograph viewed by participants in Experiment 1A.

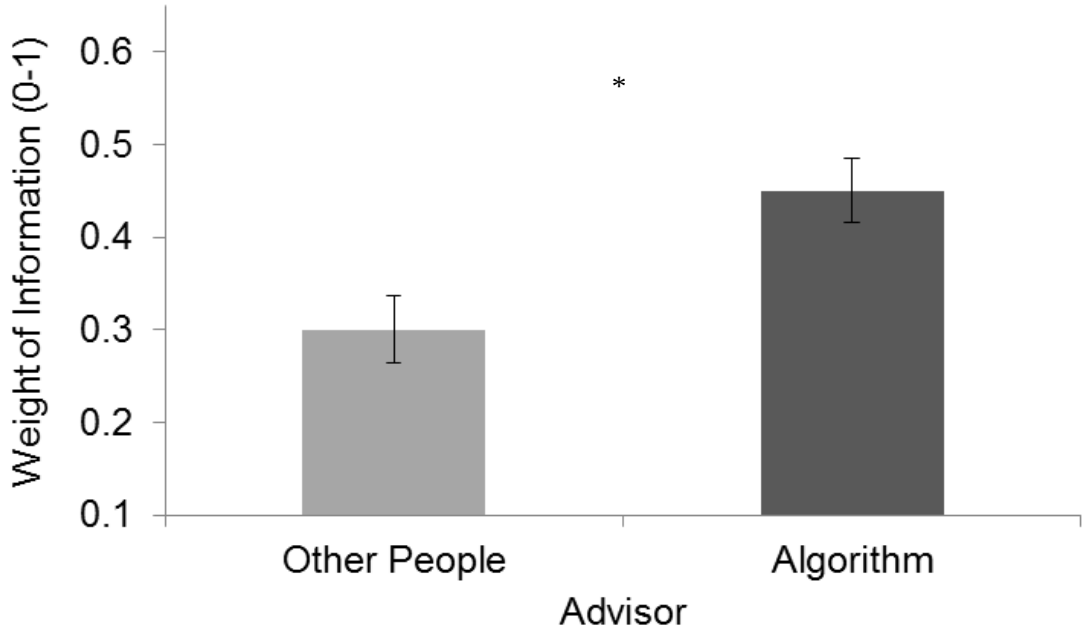

Figure 2. Weighting of Advice (WOA) as a function of experimental advisor (other people vs. algorithm), Experiment 1A. The higher the WOA, the more participants relied on the advice. Error bars indicate standard errors. Note: $*p < .05$.

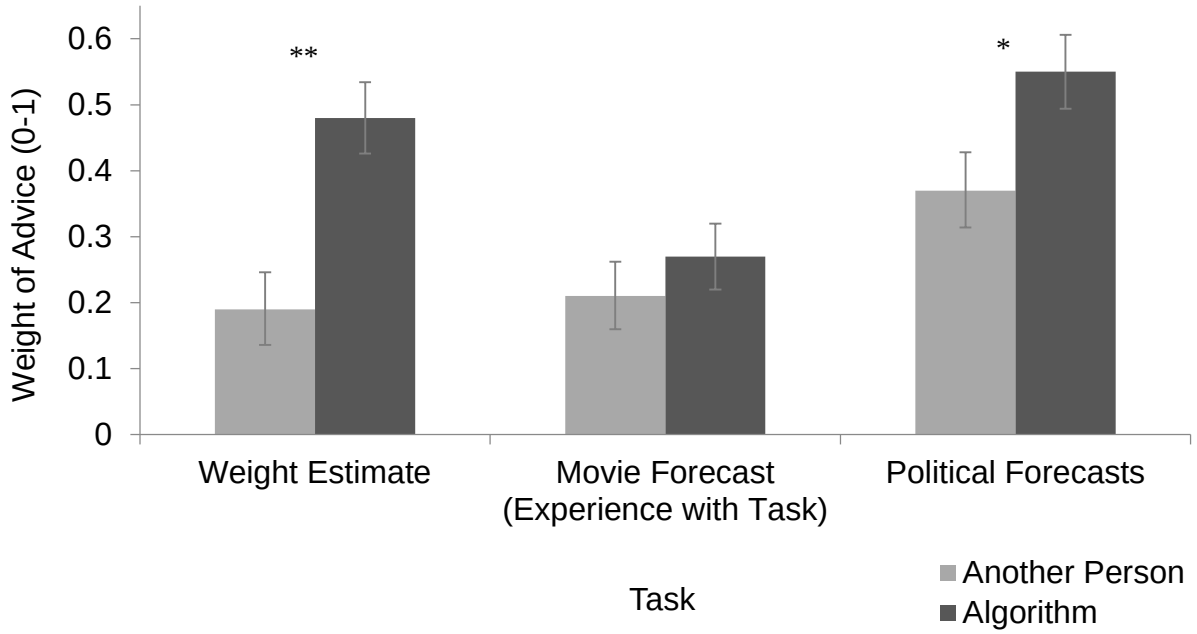

Figure 3. Weight of Advice (WOA) as a function of experimental advisor (another person vs. algorithm), Experiment 1B. Participants estimated someone's weight, forecasted the opening weekend gross for a movie, and forecasted two political world events. The higher the WOA, the greater the change in participants' Time 1 and Time 2 estimates. Note: $^{**}p < .001$, $^{*}p < .05$.

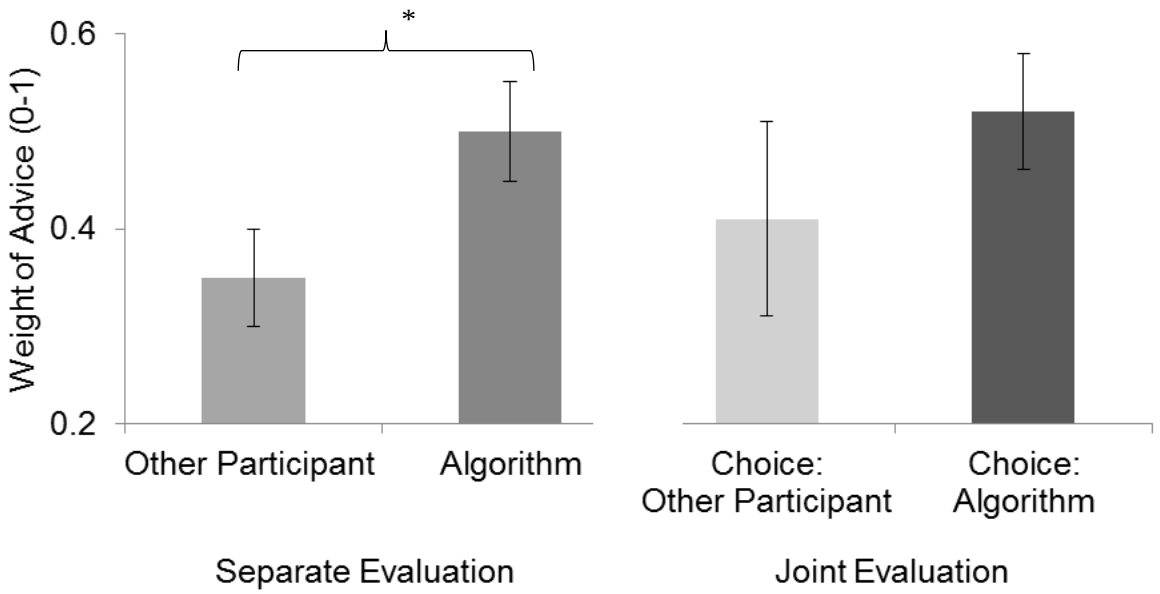

Figure 4. Weight of Advice (WOA) as a function of experimental advisor (person vs. algorithm) in the between- and within-subject designs, Experiment 2. Note: $*p < .05$.

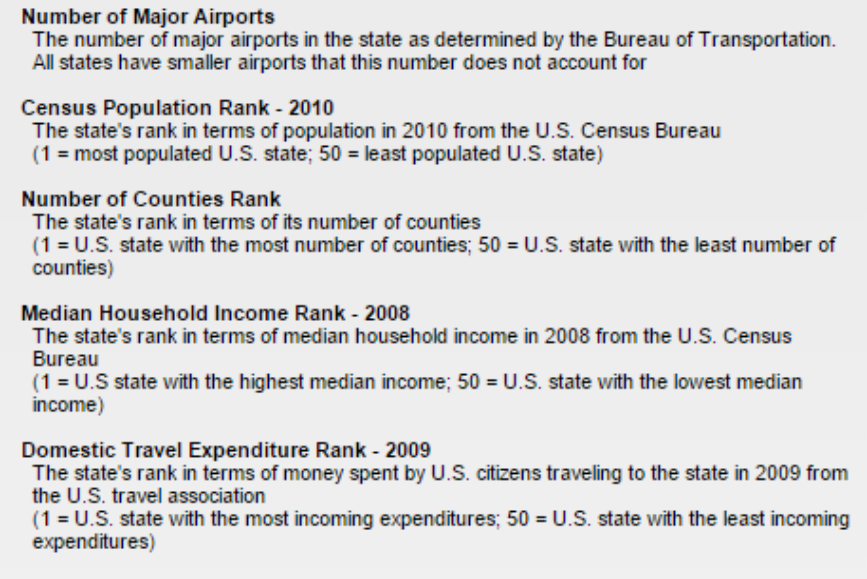

Figure 5. Screenshot of the information participants read they would receive to make their estimate, Experiment 3.

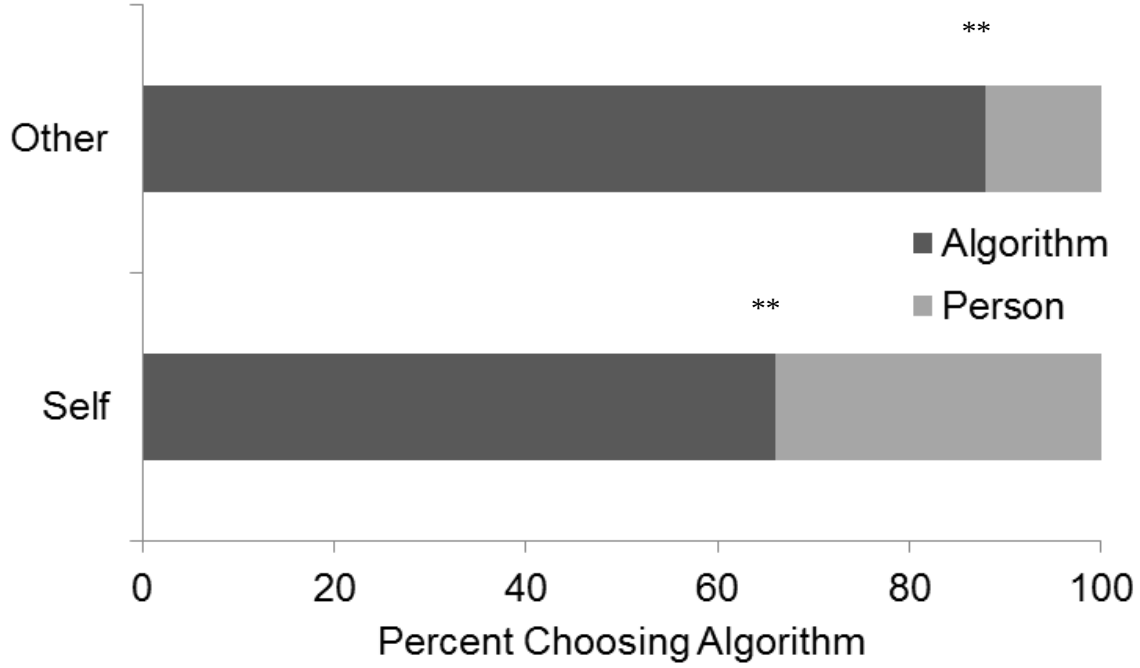

Figure 6. Percent of participants choosing the algorithm as a function of experimental advisor (person vs. algorithm) and self/other (self vs. other), Experiment 3. Note: $^{**}p < .001$. The graph is horizontal because the percentage for the other condition and the self condition each sum to 100%.

Figure 7. Photograph used in Experiment 4.

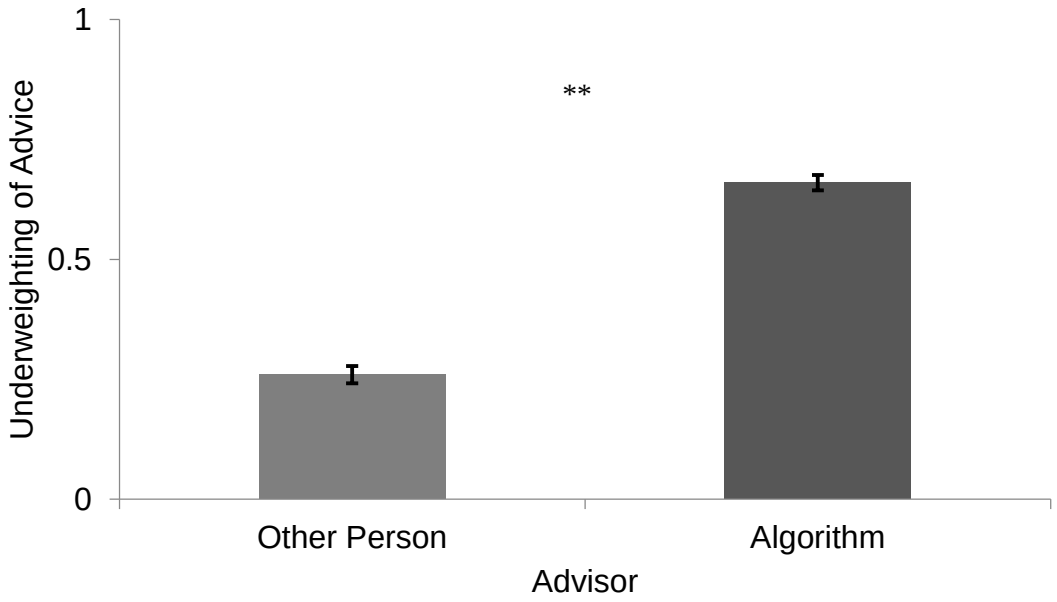

Figure 8. Magnitude of underweighting advice as a function of experimental advisor (person vs. algorithm), Experiment 4. The greater the underweighting, the more participants should have updated to the advice. Note: $^{**}p < .001$.

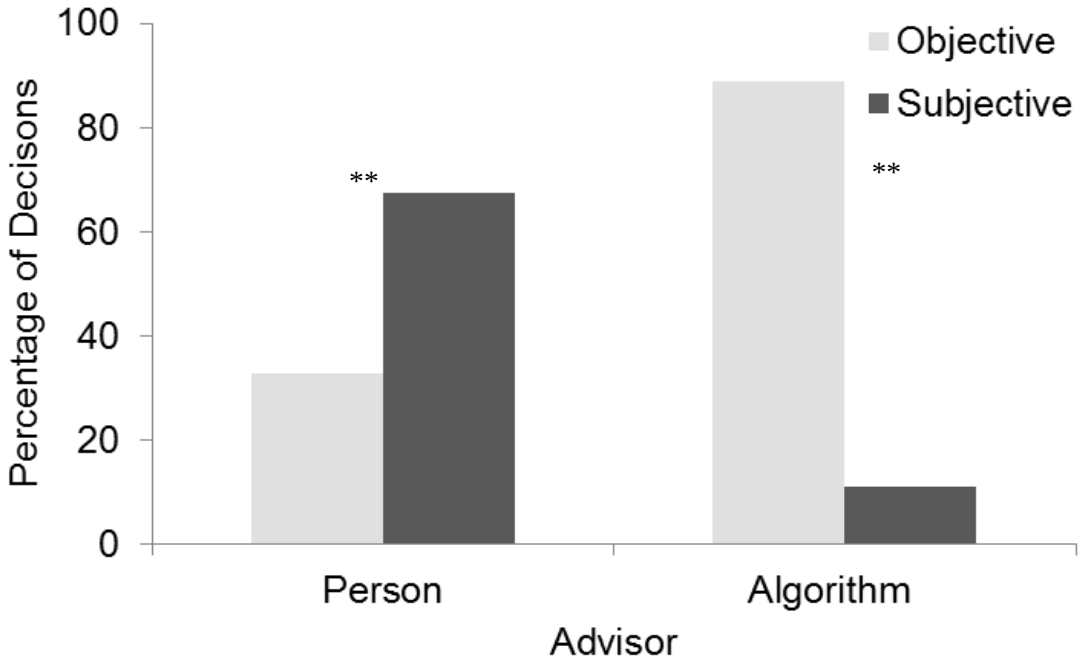

Figure 9. Percentage of decisions coded as objective or subjective as a function of experimental advisor (person vs. algorithm), Experiment 5. Note: $^{**}p < .001$.

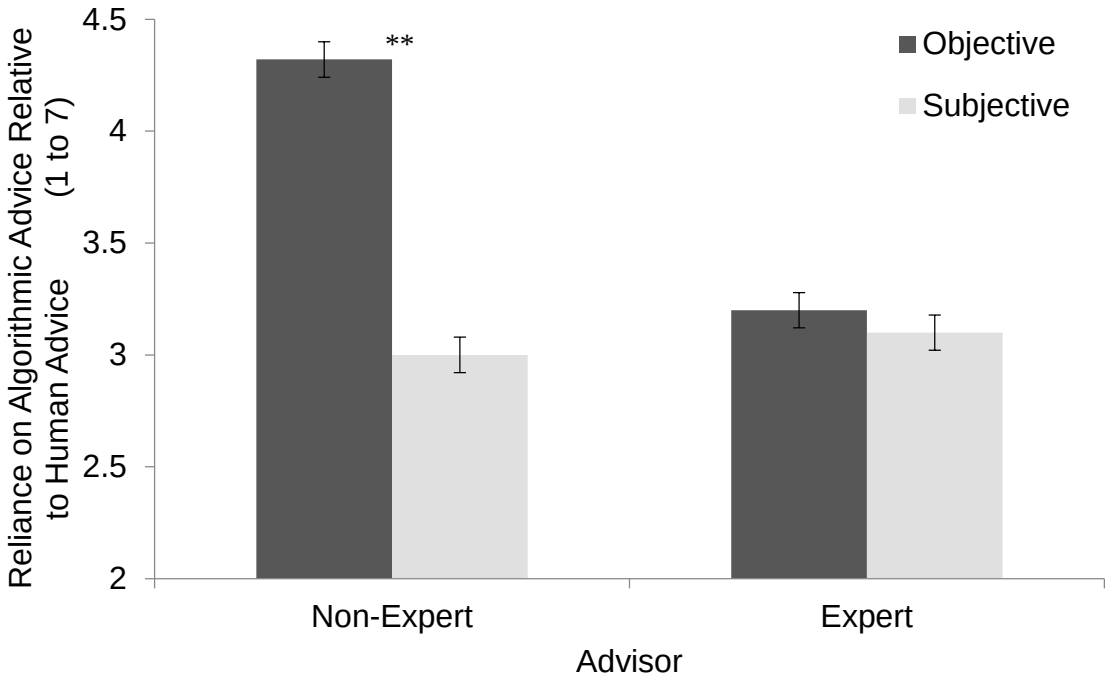

Figure 10. Reliance on algorithmic (versus human) advice as a function of experimental subjectivity (objective vs. subjective) and expertise of the human advisor (non-expert vs. expert), Experiment 6. Note: $^{**}p < .001$.

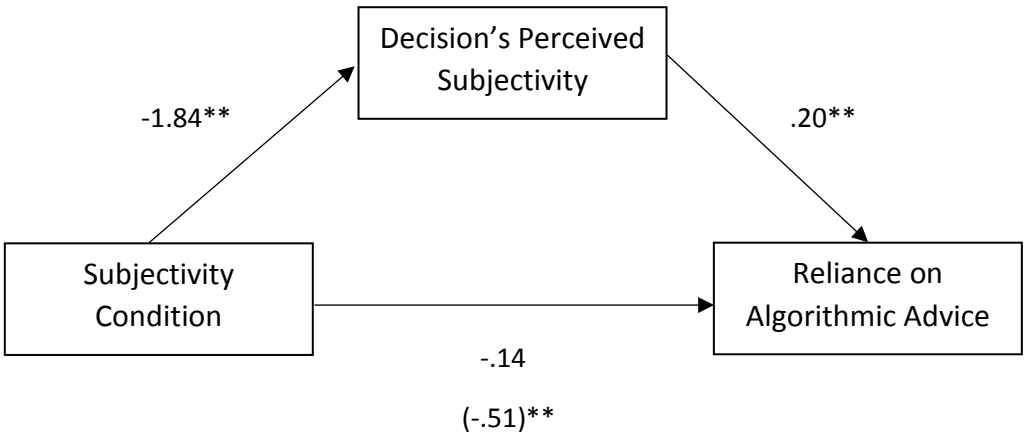

Figure 11. Reliance on algorithmic advice (versus human advice) as a function of experimental subjectivity of decision, mediated by perceived subjectivity, Experiment 6. $^{**}p < .001$.

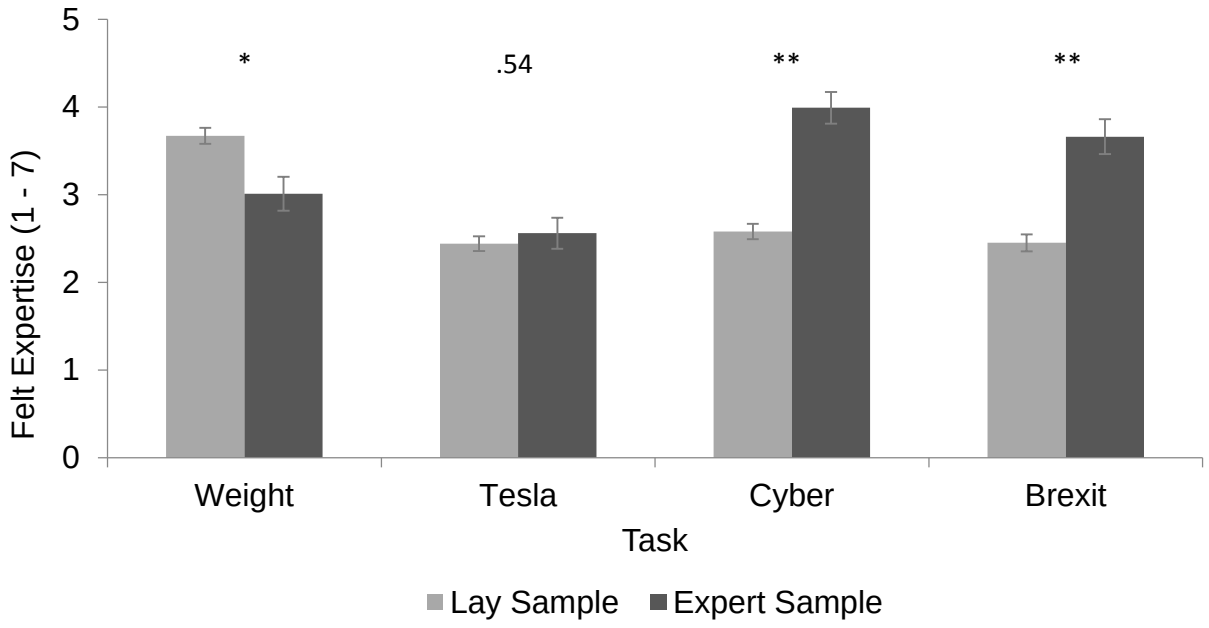

Figure 12. Felt expertise for each task as a function of experimental advisor (another person / forecaster vs. algorithm) and participant expertise (lay vs. expert), Experiment 7. Participants estimated someone's weight, forecasted a business event, and forecasted two political world events. Note: *$p < .01$, **$p < .001$.

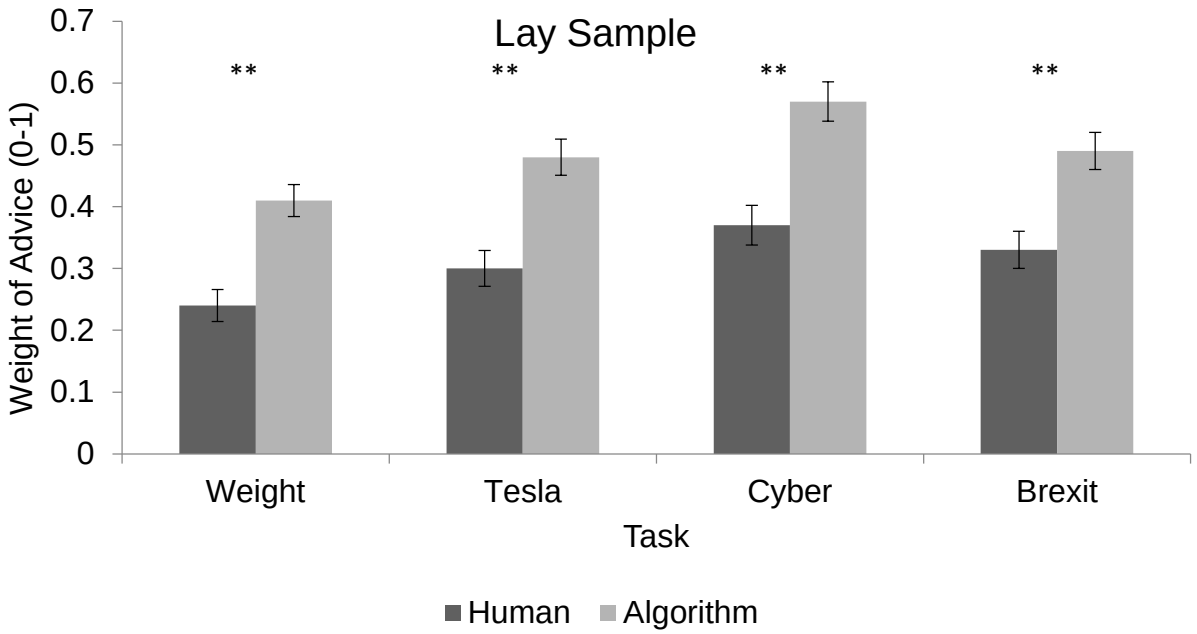

Figure 13. Weight of Advice (WOA) as a function of experimental advisor (another person / forecaster vs. algorithm) for the lay sample, Experiment 7. Participants estimated someone's weight, forecasted a business event, and forecasted two political world events. Note: $^{**}p < .001$.

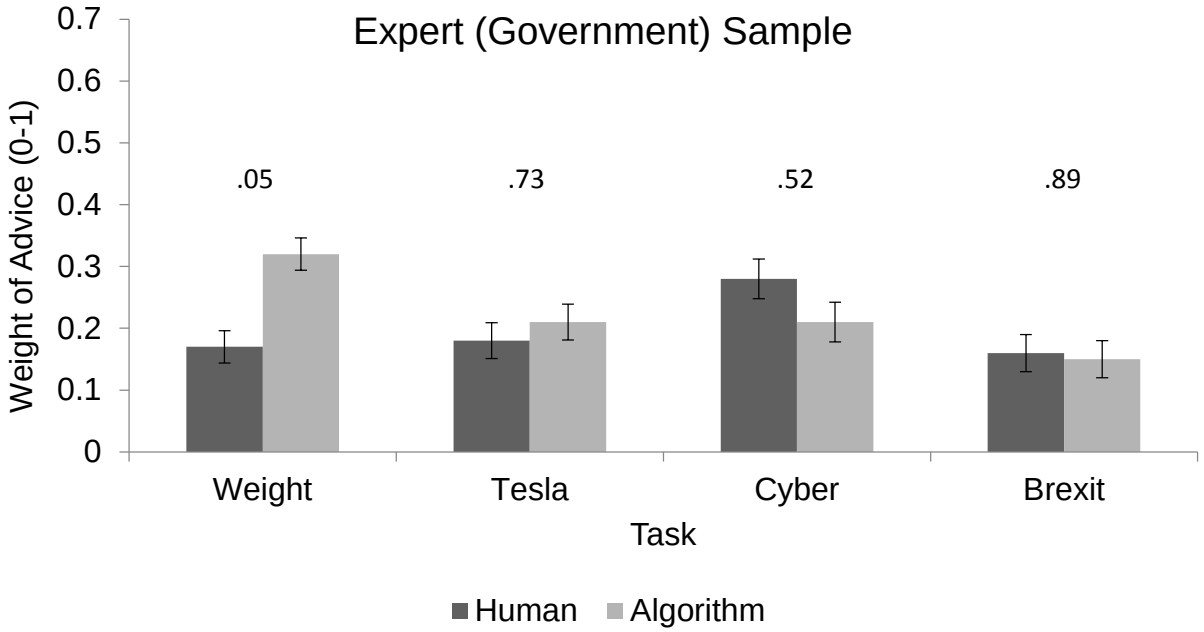

Figure 14. Weight of Advice (WOA) as a function of experimental advisor (another person / forecaster vs. algorithm) for the expert sample, Experiment 7. Participants estimated someone's weight, forecasted a business event, and forecasted two political world events. Note: non-significant *p*-values are denoted.

APPENDIX

Exclusions and Winsorization for Weighting of Advice (WOA)

Experiment 1A
Within the advice literature, moving away from advice rarely occurs (Gino & Moore, 2007). A few participants did move away from the advice, so the analyses include those WOAs as negative.

Following Gino and Moore (2007), I winsorized 11 participants' WOA to 1 because they moved past the advice (WOA greater than 1). I winsorized 1 participant's WOA to -1 because they displayed a WOA less than -1. No one guessed the same weight as the advice (163) on his or her first estimate.

After winsorizing, 68 participants (33.7%) did not change their estimates from Time 1 to Time 2 (WOA = 0), one 131 (64.9%) changed their estimate closer to the advice, and 3 (1.5%) changed their estimate away from the advice (WOA < 0). The experimental conditions do not differ significantly in the number of people with a WOA of 0 or a WOA greater than 0, $\chi 2$ (1, N = 202) = 0.987, p = .320 nor in in perceived difficulty, p = .60. This implies that the experimental manipulation did not affect perceptions of task difficulty. The results remain significant when follow-up analyses exclude participants with negative WOAs. In all other studies, I pre-registered how I winsorized participants' WOA.

Experiment 1B
As pre-registered, for each task, I excluded participants excluded who guessed the same as the advice, moved away from the advice with a magnitude WOA ≤ -1, past the advice with a magnitude WOA ≥ 2 or guessed the actual answer (which I only used for the weight task). I winsorized participants' WOA to 1 if they moved past the advice (WOA > 1) but still displayed a WOA less than 2 and winsorized them to 0 if they moved away from the advice (WOA < 0) but still displayed a WOA greater than -1. I collected data from 77 participants.

Weight estimate. The 2 participants dropped from the weight estimate included 1 participant who guessed the same as the advice (and also did not change from time 1 to time 2) and 1 participant with a WOA < -1. 75 participants had usable WOAs.

Movie. The two participants dropped from the movie forecasting task included 1 participant with a WOA < -1 and 1 participant with a WOA ≥ 2. 75 participants had usable WOAs.

Geopolitical Forecasts. The 3 participants dropped from the first geopolitical forecasting task included 2 participants who guessed the same as the advice (and also did not change from time 1 to time 2) and 1 participant with a WOA ≥ 2. 74 participants had usable WOAs. The 1 participant dropped from the second geopolitical forecasting task displayed a WOA < -1. 76 participants had useable WOAs.

Experiment 2
As pre-registered, I winsorized 8 participants' WOA to 1 because they displayed a WOA greater than 1. No participants moved away from the advice, guessed the same weight as the advice (163) on their first estimate, or guessed the actual weight (164) on their first estimate.

Experiment 3
I collected more completed surveys than planned (479) in order to meet the goal of 400 for the final sample. As pre-registered, I removed 32 participants who had repeat I.P. addresses (only removing their second responses). This left 448 participants. As pre-registered, I removed the 35 participants who reported having possibly seen the task previously and 9 who reported definitely seeing the task. This left the final sample size at 404. As a note, 119 participants did not complete the survey because 42 failed the attention check after the consent form. 26 dropped out and 77 dropped out after answering whether they saw the task before (prior to any manipulation).

Experiment 4
I collected 891 completed surveys and as pre-registered, I removed 48 participants who had repeat I.P. addresses (only removing their second responses), 15 participants who answered with non-response answers in the open ended questions ("NA" or non-sense responses such as "wrinkles"), 55 participants reported that the photograph of the man looked familiar or were unsure, 1 respondent who did not proceed past the English fluency question. Removing these respondents left 772 participants before I winsorized WOAs. As a note, 1 participant did not complete the survey.

As pre-registered, I winsorized 14 participants' WOA to 1 who displayed a WOA greater than 1 (moved towards and past advice) and less than 2. I winsorized 15 to 0 who displayed a WOA less than 0 (moved away from advice) and greater than or equal to -1. I excluded 44 participants who guessed the same as the advice and did not change their answer, 4 who displayed a WOA less than -1 and I excluded 1 who displayed a WOA greater than 2. These pre-registered exclusions brought the final sample size to 671.

Experiment 5
I collected 52 completed surveys. I found no participants associated with a repeat I.P. address. As pre-registered, I removed the 1 participant who answered with nonsense answers such as "yes," "no," and "wait" prior to coding. This left 51 participants with a total of 303 decisions. As a note, 35 participants dropped our when they saw the writing task. The coders showed high inter-rater reliability ($\alpha = .85$).

Breaking 9 ties and removing 27 decisions, from the original 303 decisions, brings the final sample size to 276 decisions coded as objective or subjective (details below). For the 9 ties, where one coder rated the decision a 1 and the other rated it a 3, the author coded the decision, blind to the condition. When one coder rated a decision either a 2 or left it blank (categorizing it as too vague a decision), I used the other coder's response of 1 or 3 (Coder 1 had 43 blank decisions and 70 decisions rated a 2 and Coder 2 had 2 blank decisions and 27 decisions rated a 2). I removed a total of 27 decisions prior to analysis for: 11 decisions left blank by 1 coder and rated a 2 by the other, 15 decisions rated a 2 by both, and 1 decision left blank by both.

Experiment 6
I collected 580 completed surveys. As pre-registered, I removed 25 participants who had repeat I.P. addresses (only removing their second responses) and 5 participants who failed the attention check and tried to complete the survey. This left the final sample size of 550 participants. Including those 5 does not alter the results. As a note, 172 participants did not complete the survey due to failing the attention check directly following the consent form.

Experiment 7
I collected 370 completed survey responses from mTurkers. As pre-registered, I removed 13 participants who had repeat I.P. addresses and 1 who had a repeat mTurk I.D. (only removing their second responses) as well as 36 participants who took the survey in less than 4 minutes or more than 30 minutes. This left 320 mTurkers. As a note, 176 mTurkers did not complete the survey. Most of these people dropped out on the page after the consent form (only 30 dropped out after taking a few of the tasks).

I collected 116 completed survey responses from experts. For the experts, I did not remove any repeat I.P. addresses due to the limited number of available participants in this pool. People may have taken the survey on a work computer which could account for a shared I.P. address. As a note, 76 people did not finish the survey. Most of these people dropped out before providing consent. Only 9 dropped out after a few of the tasks.

I filed an addendum to my original pre-registration after looking only at the open-ended responses of where experts worked, but before analyzing data. This allowed an appropriate exclusion of 33 participants in the expert sample who did not report working in National Security for the U.S. Government (or as a consultant for the U.S. Government). I moved 14 participants who listed a civilian job to the lay sample, and excluded participants who did not list their job as national security for the U.S. government (2 government contractors, 7 government employees in other departments such as agriculture, 8 non-U.S. government employees, and 2 non-profit employees). These exclusions and move of 33 participants brought me to 334 lay participants and 83 experts.

Due to the goal of the experiment, to better understand the role of expertise, I pre-registered more exclusions for this experiment than past experiments. Following my pre-registered exclusions, I removed people who may have had prior exposure to the forecasts: 31 from the 334 mTurkers and I removed 13 from the 83 experts. These people either answered that they had 1. previously participated in the Good Judgment Project or Good Judgment Open forecasting tournament, where they may have seen the same forecast or 2. used the Internet or other sources of information beyond the links provided in the survey to find more information on the tasks. Eight mTurkers and 5 experts reported participation in the Good Judgment Project tournament while 12 mTurkers and 2 experts who were unsure.

Multiple people who used external sources for one forecast also did for a second or all three (Tesla forecast: 12 lay, 6 expert; Cyber forecast: 10 lay, 3 experts; Brexit forecast: 10 lay, 4 expert). These exclusions removed 31 lay people from a sample of 334 for a total of 301 and 13 experts from a sample of 83 for a total of 70.

As pre-registered, by task, I excluded participants who guessed the same as the advice, moved away from the advice with a magnitude WOA ≤ -1, past the advice with a magnitude WOA ≥ 2 or guessed the actual answer (only used for the weight task). 282 lay people and 61 experts answered all tasks.

Weight estimate. I winsorized 10 participants' WOA to 1 who displayed a WOA greater than 1 (moved towards and past advice) and less than 2. I winsorized 2 to 0 who displayed a WOA less than 0 (moved away from advice) and greater than or equal to -1. The 8 participants dropped from the weight estimate included 1 participant who guessed the same as the advice, 1 participant with a WOA ≤ -1, and 6 participants with a WOA ≥ 2.

Tesla Forecast. I winsorized 9 participants' WOA to 1 who displayed a WOA greater than 1 (moved towards and past advice) and less than 2. I winsorized 10 to 0 who displayed a WOA less than 0 (moved away from advice) and greater than or equal to -1. The 10 participants dropped included 9 participants who guessed the same as the advice, 0 participants with a WOA ≤ -1, and 1 participant with a WOA ≥ 2.

Cyber Forecast. I winsorized 20 participants' WOA to 1 who displayed a WOA greater than 1 (moved towards and past advice) and less than 2. I winsorized 3 to 0 who displayed a WOA less than 0 (moved away from advice) and greater than or equal to -1. The 5 participants dropped included 2 participants who guessed the same as the advice, 1 participant with a WOA ≤ -1, and 2 participants with a WOA ≥ 2.

Brexit Forecast. I winsorized 27 participants' WOA to 1 who displayed a WOA greater than 1 (moved towards and past advice) and less than 2. I winsorized 2 to 0 who displayed a WOA less than 0 (moved away from advice) and greater than or equal to -1. The 5 participants dropped included 0 participants who guessed the same as the advice, 1 participant with a WOA ≤ -1, and 4 participants with a WOA ≥ 2.