Title
When do people rely on algorithms?

Author
Logg, Jennifer Marie

Publication Date
2016

Peer reviewed|Thesis/dissertation

When do people rely on algorithms?

Jennifer Marie Logg

A dissertation submitted in partial satisfaction of the

Requirements for the degree of

Doctor of Philosophy

in

Business Administration

in the

Graduate Division

of the

University of California, Berkeley

Committee in Charge:

Professor Don A. Moore, Chair
Professor Leif D. Nelson
Professor Cameron Anderson
Professor Michael A. Ranney

Spring 2016

When do people rely on algorithms?

ABSTRACT

When do people rely on algorithms?

by

Jennifer Marie Logg

Doctor of Philosophy in

Business Administration

University of California, Berkeley

Professor Don A. Moore, Chair

Algorithms, scripts for sequences of mathematical calculations or procedural steps, are powerful. Even though algorithms often outperform human judgment, people appear resistant to allowing a numerical formula to make decisions for them (Dawes, 1979). Nevertheless, people are increasingly dependent on algorithms to inform their decisions on a day-to-day basis. In eight experiments, I tested whether aversion to algorithms is as straightforward a story as past work suggests. The results shed light on the important questions of when people rely on algorithmic advice over advice from people and have implications for the use of algorithms within organizations.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

Thank you to my dissertation committee Don Moore, Leif Nelson, Cameron Anderson, and Michael Ranney. One of the highlights of my research career is having all of your talented brains in the same room during my proposal defense to discuss algorithms. Your insightful feedback is invaluable.

Don, thank you for your mentorship and wisdom. I have learned so much from you; your clever writing and way of thinking about the world is inspiring. I appreciate that you see research as a path towards truth. That perspective has made research a joyful process of discovery for me. You put students first and I am grateful that your door is always open to talk about ongoing projects and new ideas. After meeting with you, I have always felt energized and determined to tackle the next stage of work. You are a grounding force in my life and I am so happy to call you my advisor. I look forward to our continued collaborations.

Leif, thank you for our enlightening discussions. First, I have learned so much more in your weekly journal club than I could have ever expected. Because of what I've learned from you, I not only consume new research differently but have used it to improve my own research. Second, I appreciate how you have encouraged me to think about my dissertation work from new perspectives. You taught me that if I have a question, to go for it and find the answer.

Cameron, thank you for sharing your genius about framing research ideas. I admire your thought process and astute observations about both theory and human nature. Our research discussions have been a great source of motivation.

Michael, thank you for your creative feedback and ideas on my dissertation. I greatly admire your contagious enthusiasm for research. I have immensely enjoyed our discussions on algorithms. Conversations with you are richly generative and your insights have challenged me to consider new and exciting directions.

Thanks to Poonam Arora for your mentorship throughout my research career; you are my friend and role model. I admire your strength of character, passion for research, and genuine kindness.

Thanks to the faculty at Haas, especially the entire Management of Organizations faculty. I appreciate your generosity of time. You have all taught me so much through your questions and comments in each colloquium talk. Thanks especially to Jenny Chatman, Ming Leung, Dana Carney, Laura Kray, and Sameer Srivastava for your thoughtful feedback on research ideas and presentations and for career advice. Toby Stuart, your questions during talks are often the highlight of that talk for me. Thanks to Jo-Ellen Pozner for your continued encouragement. Clayton Critcher, thank you for your statistical advice, for leading behavioral lab, and for your fascinating questions and comments during talks. Thank you, Juliana Schroeder and Ellen Evers, for your friendship and support in my last year of graduate school.

Thanks to my collaborators. Liz Tenney, thank you for helping me develop as a researcher and writer - I will always strive to utilize your writing tips. I have also enjoyed our

time together discovering The Bay. Uriel Haran, I admire your ability to develop fascinating new research ideas and value our collaboration. Thank you Rick Larrick, for imparting your wisdom to our collaboration. Thank you also for your insights on this dissertation work.

Thanks to those who have made Berkeley my home. Donatella Tauasi, thank you for your level-headed outlook on life; our talks keep me centered. Muping Gan, your ever-positive outlook and joy for life have kept me energized and enthusiastic. Wei Ng, thanks for sharing your curious mind and beautiful graphs with me and for your wonderful dinner parties. Minah Jung, thank you for your sage advice and for our California adventures. Jon Cowan, thank you for great conversations on our hikes and on your deck. Laura Howland, thank you for our talks and sharing research ideas over ice cream. Katherine Ullman, thank you for making the office a brighter place – you are an absolute delight. Alice Moon, I have so much fun with you and have always appreciated your great comments in journal club. Stephanie Eistetter, I wish we could take tea time more often. Phoebe Wong, thanks for being you; please do not change. Silva Kurtisa, thank you for your calm presence and incisive wit. Kate Ashley, Korcan Ak, Scott Roeder, and Mike Rosenblum, thank you for your friendship and for taking breaks to play tennis. Alex Van Zant, Eliot Sherman, Sanaz Mobasseri, Angus Hildreth, Brian Reschke, Daron Sharps, Mike O'Donnell, Fausto Gonzalez, and Hannah Perfecto, I have appreciated your friendship and feedback on ideas over the years. Thanks to the impressive lab managers and research assistants who make the Accuracy Lab a motivating and productive place. You are great thinkers and great people. Thanks to the Haas Ph.D. Program office, Kim Guilfoyle, Bradley Jong, and Melissa Hacker for your smiling faces, knowledge, and kind words.

Thanks to my friends and family. Carolyn Rovee-Collier, you were my first female scientist role model and I owe my research career path to you. I hope to one day be half the woman you were in life and in science. Thank you to the entire CRED family at Columbia, especially to Dave Krantz, Elke Weber, and Ben Orlove. The poise with which you lead researchers to create something bigger than their individual work served as an amazing introduction to research. Thanks to Julie Smith Holterhaus, Katherine Fox-Glassman, Kirstin Appelt, Dave Hardisty, Ray Crookes, Lisa Zaval, and Shahzeen Attari who welcomed me into the research family.

Thanks especially to Dad and Lynn for being there every step of the way. Dad, you have shown me what it means to build a good life. You have made me who I am by fostering my curiosity about the world. Thank you for a call whenever I need some sunshine. Lynn, your strength and immense capacity to care for others humbles me. I am so grateful for your warm support. Katie, thank you for being my family.

Malachy English, thank you for your constant support and making me smile every day. When I met you, I knew that you would be a special person in my life. I am so lucky and proud to know you.

Thank you also to the UC Berkeley Haas Dissertation Fellowship, the Intelligence Advanced Research Projects Activity (IARPA), and the Behavioral Lab at Haas for their generous financial support.

PREFACE

My dissertation is organized as follows. In Chapter 1, I discuss the potential of algorithms to improve the accuracy of human judgment, anecdotal evidence for both distrust of and reliance on algorithms, and conflicting empirical evidence. I situate my work relative to the literature on clinical versus actuarial decisions, wisdom of crowds, overconfidence, and advice-taking. This work shows that algorithms can provide more accurate judgments than people and also suggests that people are unwilling to listen to algorithmic advice (Dawes, 1979).

In eight experiments, I examined how much people are willing to rely on advice from an algorithm relative to advice from people. These experiments are organized into three categories. First, I tested if people are willing to rely on algorithmic advice and attempt to avoid limitations found in prior empirical work that presents conflicting evidence. In Experiments 1A and B (Chapter 2 and 3), participants were more influenced by advice when they thought it came from an algorithm than when they thought it came from other people.

Second, I tested *when* people are willing to rely on algorithmic advice by reconciling my results with work that finds aversion. Experiment 2 (Chapter 4) found that reliance on algorithmic advice was robust to different presentations of the human and algorithmic advisers (joint versus separate evaluation). Experiment 3 (Chapter 5) showed that overprecision, excessive confidence in one's knowledge, attenuated reliance on algorithms.

In Experiment 4 (Chapter 6), subjectivity of the decision domain moderated expectations of reliance on algorithmic advice. Participants expected greater reliance on algorithmic advice for more objective than subjective decisions. Experiment 5 (Chapter 7) tested why people relied on algorithms and found that participants expected to rely more on expert advice but relied similarly on algorithmic and expert advice. Experiment 6 (Chapter 8) examined the interaction between the subjectivity of decision domains and the expertise of the human advisor with different manipulations than Experiments 4 and 5. Participants relied more on algorithmic advice than non-expert human advice for objective than subjective decisions but relied more on expert advice, regardless of subjectivity.

Third, I tested if people rely on algorithmic advice as much as they should. In Experiment 7 (Chapter 9), participants were provided with normative information in order to benchmark how much they should rely on the advice. Although reliance on algorithms replicated, participants underweighted the algorithmic advice more than human advice.

In Chapter 10, I summarize my research findings for when algorithms are most likely to help improve people's accuracy. I discuss the theoretical implications, practical implications, and future directions for this program of research. These results shed light on the important questions of when people rely more on advice from algorithms than advice from people and have implications for the use of algorithms within organizations.

CHAPTER 1

Theoretical Background and Literature Review

In this chapter, I review evidence on algorithms' potential to improve on human judgment and whether people are willing to listen to them. This dissertation seeks to examine the apparent distrust in algorithms and widespread dependence on them. It seeks to answer when people rely on algorithmic advice more than advice from people.

Algorithms, scripts for sequences of mathematical calculations or procedural steps, are powerful. They can complement human judgment and are used widely to inform it. Professionals routinely extract information from the Internet using Google Search, perhaps while listening to music from Pandora. Algorithms inform many day-to-day decisions from the mundane to the consequential. They can help us travel from point A to point B with Google Maps while using Global Positioning System (GPS). Many people even use algorithmic financial advisors such as Betterment for their financial decisions. More and more companies 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. This dissertation focuses on more complex and opaque algorithms than the most basic form of an algorithm, which could be a simple procedure such as "When my alarm wakes me up, I turn it off" or "After I put on running shoes, I lace them up."

The Power of Algorithms

Although organizations have traditionally relied on humans to forecast future events (Hartford, 2014), algorithms often outperform human judgment in many domains. When attempting to accurately forecast future events, aggregation of individual forecasts can outperform forecasts made by individuals (Hastie & Kameda, 2005; Larrick & Soll, 2006). In addition, more complex algorithms can outperform simple averaging in terms of accuracy (Baron, Mellers, Tetlock, Stone, & Ungar, 2014).

Literature dating back to the 1950s compares clinical (human) versus actuarial (statistical) predictions and shows that algorithms often make more accurate forecasts than experts. Simple aggregation can outperform expert forecasts of many consequential outcomes: survival of cancer patients (Einhorn, 1972), recidivism of parolees (Caroll et al., 1982), and even geopolitical events including terrorist attacks and the outbreak of war (Ungar, Mellers, Satopaa, Baron, Tetlock, Ramos, Swift, 2012). Algorithms can also outperform human judgment in assessments of different kinds of pathologies (Beck, Sangoi, Leung, Marinelli, Nielsen, van de Vijver, & Koller, 201; Goldman, Caldera, Nussbaum, Southwick, Krogstad, Murray, & Slater, 1977; Heden, Öhlin, Rittner, & Edenbrandt, 1997), operational risk (Tazelaar & Snijders, 2013), and answering trivia questions (Tesauro, Gondek, Lenchner, Fan, & Prager, 2013).

Algorithms can surpass human judgment in accuracy for a few reasons. First, work on the wisdom of crowds shows that even the simplest algorithm (averaging across people) improves accuracy because it cancels out errors from individuals (Galton, 1907; Hastie & Kameda 2005; Soll & Larrick 2009; Surowiecki, 2004). Second, many algorithms are not based

on simple aggregation of human judgment and are instead presented with the same cues as people are. In these instances, algorithms can weigh the cues more appropriately than people do (Dawes, 1979). For example, when predicting heart attacks from cues within test results, an algorithm relied on different cues than a widely used medical program and expert cardiologists. The cues used by the algorithm better predicted heart attacks than the cues used by both the program and experts (Heden et al., 1997). In fact, algorithms can even *identify* more predictive cues. When pathologists and an algorithm were given the same biopsy images, the algorithm predicted the severity of breast cancer better than pathologists partly because it identified cues other than those the pathologists were trained to use (Beck et al., 2011).

Apparent Distrust of Algorithms

Despite the helpfulness of algorithms, people appear resistant to allowing a numerical formula make decisions for them (Bazerman, 1985; Dawes, 1979). The consequences of distrust can prove substantial. To pick one tragic example, in 2004, Flash Airlines Flight 604 crashed into the Red Sea, resulting in the largest death toll in Egypt's aviation history (Sparaco, 2006). The captain, who was experiencing a condition known as spatial disorientation, trusted his own flawed judgment over the aircraft's instruments. In this case, reliance on human judgment over calculations by a machine resulted in disaster. Another domain in which decisions determine life or death is medicine. For decisions such as changes to medications, cardiologists ignored recommendations from statistical algorithms more than half the time (Keeffe et al., 2005). Even in industries where computational algorithms have won widespread acceptance, such as Silver's Player Empirical Comparison and Optimization Test Algorithm (PECOTA) in baseball, acceptance can take time and face substantial resistance (Silver, 2012). People appear to prefer their own judgment to algorithmic advice, especially when those people are considered experts and the algorithmic advice threatens their job security by doing their job for them.

Framing algorithms as competitors to human judgment, rather than complements, may exacerbate distrust. In 1997, IBM's Deep Blue computer beat the world chess champion, Garry Kasparov. IBM's Watson successfully competed against the best human contestants in the trivia contests on "Jeopardy!" In March 2016, Google's DeepMind AlphaGo beat the grandmaster of Go, Lee Sedol, in 4 out of 5 games. Go presented a more complex game than chess because of the impossibility of winning through brute-force calculations. AlphaGo was trained on old matches of the game as well as its own independent simulations, the aim of which was to mirror human intuition (Tran, 2016). Popular books explain algorithms in terms of a threat to human agency (Steiner, 2012) and researchers have predicted which jobs computers might usurp from humans (Frey & Osborne, 2013). Controversy surrounds high frequency trading (HFT), financial trading based on algorithms, in both the headlines (Koba, 2013) and popular books (Lewis, 2014).

Viewing machines as competition to humans is not new. As far back as the 19th century, the Luddites, a group of textile artisans, protested the use of machines to replace standard labor practices. History has simplified this movement into a battle of humans against machines (Conniff, 2011). This framing encourages an out-group bias (Brewer, 1979; Harvey, White, Hood, & Sherif, 1961; Sumner, 1906) against computers and algorithms in favor of human judgment. Although most innovations in machines and robotics face initial backlash, slow warming to the idea often follows. The media has become more accepting of self-driving cars

(O'Toole, 2014), delivery drones (Weinstein, 2013), robot butlers, and robot house cleaners (Markoff, 2014).

Changing Boundaries of Technology and the Rise of Machines

There are always limits to what algorithms can achieve. Yet, the boundaries on what algorithms, and machines driven by algorithms, can do are rapidly expanding. What seemed impossible for technology to achieve a few years ago has already become reality. Perhaps only an algorithm could have predicted some recent innovations like robot chefs (Murphy, 2015) and robot farmers who milk cows (McKinley, 2014).

Previous work focused on forecast tasks on which algorithms performed better than humans. Although there is no empirical work that shows where algorithms perform worse than humans, to the author's knowledge, there are obviously instances where algorithms and automated machines fall hilariously flat on their (figurative) faces. Although there is limited work on people's perceptions of algorithms, there is work on expectations of computers and robots. Not many people would expect a computer to understand things like culture (Copeland, 2013), to sing, to dance, or to tell jokes. A few years ago, people thought that computers were worse than humans at providing comfort to people, expressing emotion, performing less routine physical tasks, and recognizing complex patterns (Kamenetz, 2013).

Automated machines, which are often run by algorithms, are improving. Improvements in pattern recognition already allow organizations to save lives and conserve other valuable resources. The US. Government invested in robots in order to address the biggest threat to troops in Afghanistan in 2009 ("Roadside bombs 'No. 1 threat' to troops in Afghanistan", 2009), roadside bombs (improvised explosive devices or IEDs). By 2012, it had utilized an estimated 3,500 robots (Hodge, 2012). Among other duties, robots helped detect IEDs and inspect vehicles at checkpoints (Axe, 2011). In other preventative measures, predictive policing is more common, where algorithms combine historical crime maps with geographical areas vulnerable to crime and weather maps in order to increase patrols in flagged areas (Lynch, 2016). Google's driverless cars are designed to recognize and respond to changes in traffic patterns. In fact, the largest hurdle in trials is the transition of control from the car to the passenger, because the passenger is often so comfortable that they fall asleep (Gross, 2015). Algorithms in Intensive Care Units (ICUs) monitor and ping remote doctors who are on-call overnight when they detect that patients in hospital rooms need care (Mullen-Fortino et al., 2012). And shortly, the examples provided here will sound charmingly outdated.

Widespread Dependence on Algorithms

It may take a few more years before robo-chefs are a common household member (or appliance) but many people already rely on algorithms in their day-to day lives. Algorithms inform our decisions when they serve as our secretaries, travel agents, headhunters, matchmakers, movie critics, travel agents, D.J.s, clothing stylists, beauty specialists, and sommeliers (from Siri to Kayak, LinkedIn, OkCupid, Netflix, Kayak, Pandora, Stitch Fix, Birchbox, and Club W respectively).

Improved accessibility is one potential reason for widespread dependence on algorithms in our day-to-day lives. Before the personal computer, mainframe computers played an important role improving decisions but were only accessible to large organizations with deep pockets and enough physical space to house them. In World War II, the British government relied on Alan Turing's "bombe" machines, algorithms that cracked the German codes created by their encryption machine, Enigma. The improvement of technology from mainframe, to personal computer, to smartphone has provided more people access to algorithms. The original mainframes took up an entire room. In order to use them, people needed to stay in the same room as the machine. Today, users of personal computers and smart phones use algorithms daily. As those technological platforms become more common across the world, algorithms become more seamlessly integrated into their use. Multiple algorithms quietly power the (often free) applications we use on our phones and even our watches (Apple Watch). Algorithms previously required us to relocate in order to utilize them. Now, they move with us.

Empirical Evidence for Algorithm Aversion

The enormous strength of algorithms has prompted speculation as to whether people use them as they ought to. This paper seeks to explore the tension between the apparent distrust in, and widespread dependence on, algorithms. Experiment 1A was designed to test two competing predictions. The first is that people rely on advice from an algorithm less than they rely on advice from other people. Anecdotal evidence on the accuracy of algorithms would have us believe so (Dawes, 1979; Dawes, Faust, & Meehl, 1989; Kleinmuntz, 1990; Meehl, 1954; Meehl, 1957; see also Kleinmuntz & Schkade, 1993). A handful of empirical experiments from the psychological and computer science literatures have examined people's perceptions of algorithms but this work has limitations that are discussed shortly. More recent experimental evidence suggests that after seeing an algorithm err, participants relied more on themselves than the algorithm (Dietvorst, Simmons, & Massey, 2015; Dzindolet, Pierce, Beck, & Dawe, 2002). In another experiment, participants who imagined they were medical patients chose a doctor's diagnosis over an algorithm's diagnosis (Promberger & Baron, 2006). 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, working).

Empirical Evidence for Reliance on Algorithms

The second prediction, that people are willing to rely more on advice from an algorithm over advice from a person, is not without evidentiary support. This evidence also has limitations that are discussed below. Results from the computer science literature show that when solving a logic case, 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. and Dzindolet et al. Other work shows that people happily rely on algorithmic search engines to remember known information for them (Sparrow, Liu, Wegner, 2011; Wegner, Ward, 2013), although it does not examine reliance on algorithms for prospective events.

There is surprisingly little experimental evidence testing people's perceptions of algorithmic advice. Moreover, the evidence is contradictory. As more and more organizations utilize algorithms to sort through big data, they create more opportunities to provide algorithmic advice. More experimental evidence is needed to better understand when people rely on algorithmic advice.

Limitations of Existing Evidence

I discuss the limitations to both evidence that suggests algorithm aversion and evidence that suggests reliance on algorithms and how my dissertation work addresses them. First, I discuss limitations regarding the subjective domains and repeated interactions found in past work. Second, I discuss potential confounds of past work: confounding human judgment with the self and confounding human judgment with expertise. The limitations of the existing empirical work leaves open the question as to whether people accept or reject algorithmic advice.

Subjective Domain. Participants relied on other people for recommendations of products that are subjectively judged, including books, movies, and jokes (Sinha, & Swearingen, 2001; Yeomans, et al., 2016). Even critics differ in stylistic preferences. Preferences are an interesting domain to consider, but one of its weaknesses is that it does not allow for an objective standard of accuracy across an entire sample. In order to understand when people are willing to improve the accuracy of their decisions by listening to algorithms, this dissertation first focuses on more objective domains (estimates and forecasts of future events) so that greater updating to advice means greater accuracy. Experiments 4 and 6 examine more subjective decisions in order to understand how people respond to algorithmic advice across a broader range of situations.

Repeated Interactions. Prior work focusing on algorithm aversion examined situations in which participants saw an algorithm make errors (Dietvorst et al., 2015; Dzindolet et al., 2002). Examining perceptions of algorithms prior to feedback are important because there are many situations in which people have limited experience with that particular algorithm. For instance, any forecast that includes a long time horizon will force the forecaster to decide whether they need to rely on an algorithm prior to feedback on the algorithm's performance. We may not know whether many forecasts are correct until years or even decades down the line: outcomes of climate change, political events such as conflicts between nations, economic events, relationship outcomes, etc. The experiments in this dissertation focus on willingness to rely on algorithms prior to seeing the performance of that algorithm.

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. A robust literature on advice-taking shows that people tend to discount advice from others, partly because they have access to their own reasoning and not to others' (Bonaccio & Dalal, 2006; Yaniv & Kleinberger, 2000; Yaniv, 2004). Thus, people consider less evidence when they attempt to justify an advisor's judgment compared with their own judgment.

Indeed, work on overconfidence suggests that people rely on their own judgment over others' judgments. People often show excessive confidence in the quality and accuracy of their own knowledge and beliefs, overprecision (Moore & Healy, 2008; Moore, Tenney, and Haran,

2016), which should encourage them to discount advice from others. Dzindolet (2002) attempted to address excessive confidence in the accuracy of one's own beliefs by including 200 iterations of a task. However, it is unclear how repeated exposure to the task addresses excessive confidence.

Anecdotal observations of algorithm aversion similarly conflate human judgment and the self, where people, especially experts, reject algorithms. Anecdotal evidence suggests that people, including the doctors (Keeffe et al., 2005) and talent scouts (Silver, 2012) previously discussed, reject algorithms that try to do 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 answers a different question than if they prefer to rely on advice from a human or algorithm. The first question conflates one's own knowledge and human judgment.

Although people may distrust algorithms, especially relative to their own knowledge, they may still rely on algorithmic advice more than advice from other people. For example, people may prefer not to admit that they are lost in the first place, often trying to figure out how to "find" where they are before consulting Google Maps. Many people will only ask a stranger for directions in extenuating circumstances: in a place with limited cell service or are in an area such as a park or college campus where Google Maps does not cover. The experiments in this dissertation first examine perceptions of algorithms by comparing reliance between two external advisors, an algorithm and another person, rather than between an algorithm and one's own knowledge. Experiment 3 manipulates whether participants choose between another person and an algorithm or choose between themselves and an algorithm to directly test the role of overconfidence in participants' perceptions of algorithmic advice.

Confound of Expertise. Prior work that suggests algorithm aversion sometimes confounds human judgment with expertise and work that suggests reliance on algorithms confounds the algorithm with expertise. Work in which participants relied on a doctor more than a computer for a medical decision (Promberger & Baron, 2006), potentially confounds human judgment and expertise. In other work, participants relied more on an "expert system" than another person (Dijkstra et al., 1998; Dijkstra, 1999), which leaves open the possibility that the algorithm condition was confounded with expertise. In these instances, participants may have merely trusted the computer because they expected the expert advisor to produce superior advice to a non-expert advisor.

Experiments 1 through 4 control for expertise by presenting the human advisor as another person or other people. This comparison also has higher external validity than comparing a human expert and an algorithm: expert advice is often costly and thus not readily available. An externally valid comparison to an algorithm is another person, not an expert. When people are lost, they often consult Google Maps, or if pressed, ask a stranger on the sidewalk. People do not think to call a travel agent or cartographer. In order to examine whether people infer expertise from algorithmic advice, Experiment 5 and 6 orthogonally manipulate expertise of the person and algorithmic versus human advisor.

Overview of Experiments

Given the power of algorithms and the conflicting evidence as to whether people rely on them or not, this dissertation tests the psychological processes behind people's response to algorithmic and human advice. The goal of the experiments in this dissertation is to

1. Test if people are willing to rely on algorithmic advice while avoiding limitations of the prior empirical evidence which is contradictory and limited,

2. Test when people are willing to rely on algorithmic advice by reconciling my results of reliance on algorithmic advice with work that finds aversion, and

3. Examine if people rely on algorithmic advice as much as they should

First, I test: are people willing to rely on algorithmic advice? Experiments 1A and 1B sought to test whether participants relied on or were averse to algorithmic advice compared with human advice while avoiding confounds of prior work. The results 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. The results held when participants made objective estimates as well as forecasts.

Experiment 2 tested whether presenting the advisors separately or jointly moderated reliance on algorithmic advice and found that reliance is robust to a direct comparison between advisors. Experiment 3 tested confidence in participants' own knowledge as a moderator of reliance on algorithmic advice. Reliance on algorithmic advice was attenuated when participants had the option to choose between their own knowledge (rather than another person's) and algorithmic advice.

Experiment 4 examined whether the subjectivity of a decision moderated expectations of reliance on algorithmic advice. Participants expected people to rely most heavily on algorithmic advice for more objective than subjective decisions. They expected greater reliance on advice from other people for more subjective than objective decisions. Experiment 5 tested a mechanism that may drive reliance on algorithms: viewing them as experts. Participants relied similarly on advice from an algorithm and a human expert, suggesting that they view them as equivalent. Experiment 6 examined the interaction between subjectivity of the decision domain and expertise of the human advisor. Expertise of the human advisor moderated the effect that subjectivity had on algorithm reliance. Reliance on algorithmic advice was greater for objective than subjective decisions, but reliance on an expert was greater, regardless of subjectivity.

Experiment 7 compared reliance on algorithms over other people to a normative benchmark for an objective task. Again, participants relied more on algorithmic advice than human advice. However, relative to the normative benchmark of how much they should have weighted the advice, participants underweighted algorithmic advice more than they underweighted advice from another person.

CHAPTER 2

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

Experiment 1A tests how much people rely on advice from an algorithm compared with advice from other people for an estimate with an objective standard. The goal of this experiment was to create a simple, clean test in order to avoid some of the potential confounds in prior work: human judgment with the self and expertise with either advisor.

All participants received the same advice but the experimental manipulation varied the label of the advisor. In fact, the advice was based on estimates of participants from a past experiment in both conditions (Moore & Klein, 2008). Experiment 1A's paradigm, adapted from the judge-advisor system (JAS) by Sniezek & Buckley (1995), allowed participants to decide how much they want to weigh the advice relative to their first estimate, in both the human and algorithm conditions. This paradigm avoids confounding human judgment with the self by measuring reliance on one's own knowledge relative to an external advisor's advice in both conditions.

Another important feature of Experiment 1A is that the algorithmic advice was presented to participants as a "black box." 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.

In this and all other experiments, I report how sample sizes were determined, pre-registered data exclusions, and all conditions. The links to materials, which include all measures, are provided for each experiment and all data will be posted online prior to publication. (Simmons, Nelson, Simonsohn, 2012). Analyses of Experiments 1B, 2, 3, 4, 5, 6, 7, and 8 were pre-registered at Open Science Framework. Pre-registrations can be found in this link: http://www.jennlogg.com/dissertation.html

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.[1] Participants were instructed that accurate responses would increase their chances of winning a $10 bonus. For this and all other experiments where participants provided estimates and forecasts, their final answers were incentivized. For this and all other experiments, I pre-registered exclusion criteria to remove any repeating occurrence of an I.P. address, keeping only the first instance of a survey response associated with that I.P. Final sample sizes include the number of participants after removing survey responses with repeated I.P.s and other pre-registered exclusionary criteria. I describe any other exclusionary criteria beyond repeat I.P. addresses.

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. They did so twice and received advice before making their second estimate, either an estimate from other people or from an algorithm based on an estimate from other people. The main dependent variable was the amount that participants relied on the advice they received: Weight 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 (163 pounds), and made a final estimate. The instructions asked participants to make all estimates in pounds and reminded them of the conversion from kilograms to pounds. In all conditions, the advice was 163 pounds, which came from averaging weight estimates provided by participants in another experiment (Moore & Klein, 2008). The person in the photograph actually weighed 164 pounds.

To view the online questionnaire as a participant, follow this link:

https://s.qualtrics.com/SE/?SID=SV_8ApVS9QrtzdlTbD&Preview=Survey&BrandID=berkeley

Advisor manipulation. Prior to their second estimate, all participants received the same advice, 163 pounds, described as an estimate from either other people or an algorithm. The advice was the average estimate from 415 participants in the past study. The average was actually quite accurate and only one pound off from the person's actual weight (164 pounds).

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."

Main dependent measure: Weight of Advice (WOA). The main dependent variable was how much participants relied on the advice. Consistent with the JAS literature, I refer to this as Weighting of Advice (WOA). Specifically, WOA measures the degree to which participants move their estimates toward the advice from Time 1 to Time 2. The measure is continuous and provides more information about participant's reliance on the advice than a binary choice measure can capture. In addition to the options of fully discounting or fully updating to the advice, participants can rely on the advice as little or as much as they would 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. A WOA of 0 means that the final estimate the participant gave was the same as his or her initial estimate and

connotes 100% discounting of the advice. A WOA of .5 means that the participant averaged the advice with his or her initial estimate.

Confidence measures. After each estimate, participants indicated how confident they were about the estimate, "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 of math questions (Schwartz, Woloshin, Black, & Welch, 1997). The higher a participant's numeracy score, the greater their comfort with numbers and mathematical literacy (0 to 11).

Results

Did participants rely more on advice when it came from an algorithm or from other people?[2] 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), $F(1, 200) = 8.86$, $p = .003$, $d = .39$. Results hold when controlling for gender, numeracy, and Time 1 confidence, $F(1, 197) = 9.02$, $p = .003$. See Figure 2. There are no main effects of these three variables (Fs < .39, ps > .52). Additionally, participants' increased confidence at Time 2 was greater in the algorithm (M = 7.79, SD = 10.90) than human condition (M = 4.48, SD = 8.83), $t(200) = 2.37$, $p = .019$, $d = .33$. This change in confidence parallels the greater reliance (WOA) in the algorithm condition.

Did numeracy correlate with reliance on advice (WOA)? For participants in the human condition, the correlation between WOA and numeracy did not attain statistical significance, r = -.12, $p = .225$. However, for participants in the algorithm condition, higher numeracy correlates with higher WOA, $r(100) = .21$, $p = .037$.

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 with the conclusion from the literature that people are averse to algorithms. Interestingly, the higher participants' numeracy, the greater their reliance on the algorithmic advice. This result suggests that familiarity and knowledge with math is one mechanism for reliance on algorithms. Familiarity with the types of procedures that might commonly drive algorithms could increase reliance on them.

Despite uncertainty about the black box algorithm's computations or process, participants were willing to rely on its advice. A black box operationalization of algorithmic advice creates a conservative test of reliance on algorithmic advice for two reasons. First, without an equation, the algorithm is more similar to the human advice. Second, distrust of algorithms may arise for many reasons, including the opacity of the algorithm's process. Uncertainty in the algorithm's process increases when people cannot access the calculations and processes of the algorithm.

Regardless of whether the process of the human advice is opaque, people should have a better understanding of how people arrive at their advice from their own experience with their own deliberation regarding the judgment. 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.

Two other important features of this experiment are the objective nature of the task and the operationalization of human judgment. This task has an objectively correct answer that allows for a measure of accuracy. The advice was an average of many estimates, and was fairly accurate in this experiment and the other experiments with a similar paradigm (Experiments 1B, 2, 5, and 6). This means that greater reliance on advice means greater accuracy. Second, participants in the human advice condition received advice identified as representing the average opinion from a number of others. By contrast, past work that found reliance on algorithms asked participants to choose between advice from an "expert system" and a single individual (Dijkstra, 1999; Dijkstra, Liebrand, & Timminga, 1998). One possible explanation for reliance on algorithms in prior work is that participants assumed that an algorithm 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).

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 to the results of Experiment 1A because it uses different tasks and a different operationalization of human judgment.

CHAPTER 3

Experiment 1B: Replication of Experiment 1A with MBA students and forecasts

Experiment 1A found that participants relied on algorithmic advice more than advice from other people. Experiment 1B seeks to replicate these results while ruling out potential alternative explanations and increasing their generalizability by using a different sample of participants, MBA students. In Experiment 1B, only the labels of the advisors ('person' and 'algorithm') differ between conditions. Conditions that differ only in the label of the advisor ensures that the conditions do not differ in number of words and allows participants to use their own default interpretations of human judgment and algorithmic advice.

Experiment 1B operationalizes human advice differently than Experiment 1B for a few reasons. Experiment 1A created a conservative test between human and algorithmic advice by basing them both on estimates from multiple people. Such a comparison gives the human advisor a fighting chance (as opposed to presenting an algorithm as one that scans the photo and uses complex equations). Although unlikely in a between-subjects design, a potential alternative explanation for the results in Experiment 1A is that participants thought that the average in the human condition was a second-rate algorithm. It is also a possibility that people may find the word "average" aversive because it can connote 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 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 aims to rule out this potential explanation by including an additional task, forecasts. The researcher cannot know the future, so participants cannot assume that she knows the correct forecast estimate. A different task also helps to increase the generalizability of the results.

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. Participants were instructed that accurate responses would increase their chances of winning a $10 bonus.

Design

The design and main dependent variables were the same as in Experiment 1A: a 2-cell (advisor: person vs. algorithm) between-subjects design than manipulated the source of advice participants received with the amount that participants relied on the advice as the dependent variable.

Procedure and Materials

Overview. The procedure was similar to Experiment 1A with one main change: the addition of forecasts. This and the remaining experiments did not include the difficulty and numeracy measures from Experiment 1A as they did not affect the results in Experiment 1A. Participants estimated the person's weight, forecasted what a movie would gross on its opening weekend, and forecasted two political events:

For the movie forecasts, participants answered, "How much will The Longest Ride gross over the opening weekend?"

For the political forecasts, participants answered, "What is the probability that the HSBC China Services Purchasing Managers' Index will fall to 50.0 or below before 1 June 2015?" and "What is the probability that before 17 June 2015, SWIFT will restrict any Russian banks from accessing its services?"

The order of tasks was held constant to ensure that the weight-guessing task was a direct replication of Experiment 1A. Prior to the time of the survey, the students had researched and forecasted the gross opening weekend for different movies as a weekly assignment for class, which allowed for a test of algorithm reliance when participants had experience with the task. Advice for the movie forecast came from a website that made forecasts for each movie opening the upcoming weekend.

The other two forecasts were political forecasts, a topic participants had not researched previously for class. The advice came from the forecasting tournament (including thousands of participants) hosted by the federally funded Good Judgment Project (GJP), on which the author was a collaborator. Prior to forecasting the two world events, participants were given 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."

The tournament was referenced in order to better understand how participants would respond to advice produced by the forecasters taking part in the tournament relative to advice produced by an algorithm. At the time of the experiment, the tournament forecasts were being fed to the CIA but it was unclear how much decision makers there might rely on the information. See Chapter 10 (section "Future Directions for Theory of Machine") for details on the tournament and more work related to it. The dependent variable was how much participants relied on the advice, Weighting of Advice (WOA).

To view the online questionnaire as a participant, follow this link:
https://berkeley.qualtrics.com/jfe1/preview/SV_ekXjWaBDAtTmnm5

Advisor manipulation. Prior to their second estimate for the weight-guessing task, all participants received the same advice (163 pounds for the weight task, 14.40 million for the movie forecast, 25% for the first political forecast, and 8% for the second political forecast) described as an estimate from either another person or an algorithm.

In the human condition for the weight task and movie forecast, participants read, "The estimate of another person was: X." For the two political forecasts, participants in the human

condition read, "The estimate of another forecaster was: X%." Another forecaster, rather than another person, was referenced 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, the main dependent variable was how much participants relied on the advice.

Definition of Algorithm. After Experiments 1B and 2, participants answered what they think an algorithm is.

Results

Did participants rely more on advice when it came from an algorithm or from other people? The subsequent analyses include data from 74 participants who produced usable WOA for all four tasks.[3] I averaged WOA measures for the two political forecasts. Then I submitted the weight estimation, movie forecast, and average political forecasts to 2-cell (advisor: person vs. algorithm) MANOVA. Replicating results from Experiment 1A, participants relied more on the advice when it came from an algorithm than from another person across tasks, as evidenced by the significant difference in WOA based on advisor, $F(3, 79) = 5.59$, $p = .002$. See Figure 3. Reliance was greater on algorithmic advice than advice from another person for the weight estimate, $F(1, 72) = 13.21$, $p = .001$, non-significantly for the movie forecast (the one task participants had experience researching in past weeks), $F(1, 72) = .50$, $p = .481$, and for the political forecasts, $F(1, 72) = 5.52$, $p = .022$.[4] Overall, participants relied more on advice from an algorithm than from another person.

Participants in Experiment 1B and Experiment 2 merely read the label "algorithm" and thus were requred to rely on their own default interpretations of the term. What do people think of when they think of an algorithm? Participants' open-ended questions at the end of the surveys were coded by a research assistant. The definitions fell into five broad categories. Although the answers were fairly general, the definitions alligned with the definition of the construct used in this dissertation, as a series of mathematical calculations or procedural steps. See Table 1.

Discussion

Experiments 1A and 1B found that participants relied more on algorithmic advice than advice from other people. The forecasting tasks in Experiment 1B ruled out the possibility that participants assumed the algorithm was a proxy for the researcher, who had access to the correct estimate. Participants relied more on algorithmic than human advice to forecast the world events, where the answer was unknowable at the time of the survey. Interestingly, there was no significant effect of reliance for the movie forecast; Participants' experience with the task may have lead them to discount all advice in favor of their own estimate. In the future, I plan to test the role of expertise as well as experience with the task as a moderator to reliance on algorithmic advice. Across the tasks, the effect of reliance on algorithmic advice was substantial.

These results also cast doubt on the alternative explanation that participants relied on the algorithm because they viewed the human operationalization (an average) as a second-rate algorithm. Reliance on the algorithm did not depend on whether human judgment was conceptualized as an average estimate or as an estimate from a specific person.

The results of Experiments 1A and 1B contradict the idea of algorithm aversion. Experiment 2 attempts to reconcile reliance on algorithmic advice in Experiment 1A and 1B with the prior results of algorithm aversion by varying the presentation of the advisors. Experiment 2 tests a major difference between this work and virtually all of the work that finds aversion: joint versus separate evaluation. 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). Experiment 2 tests joint versus separate evaluation as a moderator to reliance on algorithmic advice.

CHAPTER 4

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

Experiment 2 tests how advisors are evaluated as a possible moderator to reliance on algorithmic advice. A robust literature on preference reversals between joint versus separate evaluations shows that preferences can differ depending on how options are presented. Attributes which are otherwise difficult to evaluate without a comparison option 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).

Reliance on algorithmic advice may depend on whether the options are considered separately (as in Experiments 1A and 1B) or jointly (as in prior work) such that participants rely more on algorithmic advice when evaluating one advisor but prefer advice from a person when choosing between the advisors. Experiment 2 also uses a more specific operationalization of the human advice than "average of past participants" and "another person" in Experiments 1A and 1B: "another participant." The prediction was that participants would rely more on algorithmic advice when evaluating the advisors separately but that they would rely more on human advice when evaluating the advisors 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 my three-cell design.[5] I needed one hundred and ninety-nine participants, 50 per cell, to detect a medium effect size (Cohen's d = .4, f = .2) for an interaction in 2 x 2 ANOVA at 80% power. This study had 3 cells and needed 150 participants. Participants were instructed that accurate responses would increase their chances of winning a $10 bonus.

Design

The experiment had a design similar to that of Experiment 1A, with the addition of a third joint-presentation condition. 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 individually in two of the conditions or jointly in the choice condition. Participants were randomly assigned to see advice from another participant, see advice from an algorithm, or chose between seeing advice from another participant or algorithm. There were two main dependent variables: WOA (reliance on the advice) and choice of advisor.

Procedure and Materials

Overview. The materials differed from Experiment 1A in two ways: the addition of a third condition and the removal of the difficulty measure and numeracy scale. Participants were randomly assigned to receive advice from another participant, receive advice from an algorithm, or chose between advice from another participant and algorithm. Then, they read the advice. In the choice condition, the order in which the advisors were presented was counterbalanced.

To view the online questionnaire as a participant, follow this link:
https://berkeley.qualtrics.com/jfe/preview/SV_cXZY9OefEOzz3sp

Advisor manipulation.

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 the same sentences as above, based on the advisor they chose.

Main dependent measure: Choice. The other main dependent variable was whether participants chose to receive advice from the person or algorithm.

Secondary measure: Weight of Advice (WOA). The secondary dependent variable was how much participants relied on the advice.

Results

Replicating 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), $t(100) = 2.19$, $p = .031$, $d = .44$.[6] 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

Experiment 2 replicated reliance on algorithmic advice with a third population, undergraduate students, and used different operationalizations of both advisors. The results suggest that reliance on algorithmic advice is 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 and for environmental issues, support for social issues, and voter preferences (Hsee, 1996; 1998; Irwin et al., 1993; Nowlis & Simonson, 1997). In order to control for overconfidence, Experiments 1A, 1B, and 2 asked participants to choose between two external advisors. A choice between two external advisors allowed participants in both conditions to

consider their own knowledge relative to the advice. This choice contrasts past work which asks participants to choose between their own estimate and an algorithm's advice, which confounds human judgment with the self. Experiment 3 seeks to reconcile the results of Experiments 1A, 1B, and 2 with prior work that finds algorithm aversion by manipulating whether participants choose their own estimate or another person's advice relative to an algorithm's advice.

CHAPTER 5

Experiment 3: The effect of overconfidence on algorithm reliance

Experiment 3 tests a moderator of reliance on algorithmic advice: overconfidence. Experiments 1A, 1B, 2 and 3 purposefully controlled for overconfidence in participants' own knowledge and found participants relied more on advice from an algorithm than on advice from other people. Prior work suggests that people are resistant to allowing algorithms to make decisions for them (Dzindolet, Pierce, Beck, & Dawe, 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 and Larrick, 1999). Experiment 3 orthogonally manipulates human judgment (either choose one's own estimate or another person's advice) and the human versus algorithmic judgment to test whether reliance on algorithmic advice is moderated by overconfidence.

While Experiments 1A, 1B, and 2 measured reliance as a single interaction[7], other work has focused on why people are averse to algorithmic advice after they see it err (Dietvorst et al., 2015). When the control condition in this work is examined, prior to seeing performance feedback, participants relied more on the algorithmic advice than another participant's estimate. These results are consistent with the results from Experiments 1A, 1B, and 2. They even relied more on the algorithmic advice than *themselves*. In one Dietvorst et al.'s Experiment 3A, participants were indifferent between their own estimate and an algorithm's. These results suggest that the option to choose one's own knowledge over algorithmic advice may moderate reliance on algorithmic advice.

The goal of Experiment 3 was to replicate reliance on algorithmic advice over advice from other people ('other' condition) and moderate reliance when participants could choose between their own estimate ('self' condition) and an algorithm's estimate. The prediction was that reliance on algorithms is moderated by overconfidence; specifically that participants would rely more on algorithmic advice than advice from other people but that people would rely on themselves more than an algorithm.

Method

Participants

The final sample included 403 participants (177 women; 226 men; *Mdn*age = 32).[8] Participants on Amazon's Mechanical Turk were paid $0.80 to take the survey and instructed that the accuracy of the estimate they chose to submit (their own/another participants or an algorithm's estimate) determined their bonus pay. I pre-registered a sample size of 400 so that I could detect an attenuated effect on reliance. I used the effect size from Experiment 4, where 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 from the control condition in Experiment 4 (Simonsohn, 2014), which produces 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 that the experiment was well-powered.

Design

The experiment had a 2-cell (self/other: self vs. other) design. The between-subjects manipulation varied whether the human estimate was their own or another participant's estimate. The within-subjects factor varied the choice of an algorithm or human. The choice determine participants' incentive compatible bonus pay. Participants received $1 bonus if they made an accurate estimate. This bonus decreased by $0.15 for each rank their estimate was off from the actual rank. The scoring is detailed in the materials section. The main dependent variable was the percent of people who chose to have the algorithm's estimate determine their pay. People chose between an algorithm's estimate and a person's estimate (either their own or another person's).

Procedure and Materials

Overview. The materials and procedure were direct replications of the control condition of Experiment 3A in Dietvorst et al. (2015) but for one change, the addition of a new "other" condition, detailed below. All participants read about the estimation they would make:

> …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 they would receive in order to make the estimate. See Figure 5.

Prior to making their own estimate, participants chose how their bonus pay was determined, either by an algorithm or a person (themselves or another participant). Participants were randomly assigned to either the self condition (replicating Dietvorst et al.) or the other condition. In the self condition, they chose whether their own estimate or an algorithm's estimate determined their final pay. The one change to the materials was the addition of a condition, the other condition. In the other condition, participants chose whether another participant's estimate or an algorithm's estimate determined their final pay:

> For the official estimate, you can choose to have either your estimated rank or the statistical model's estimated rank determine your bonus. In other words, you can choose to be paid based on your accuracy, or you can choose to be paid based on the model's accuracy. You will make an estimate no matter which option you choose.
> The bonus will be determined 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

Would you like your estimated rank or the model's estimated rank to determine your bonus?

This choice was similar to the choice in Experiment 2 where participants how they wanted to determine the advisor that would determine their bonus before they saw the advisor's estimate.

Then, participants made their own estimate, regardless of how their final pay was determined. All participants read that the algorithm was developed by "experienced transportation analysts" and that it used the same information that they would receive. The main dependent variable was the percent of people who relied on the algorithm to determine their pay.

To view the online questionnaire as a participant, follow this link:

https://berkeley.qualtrics.com/jfe/preview/SV_eQm1RSUUORJfLP7

Self / Other manipulation. Participants either chose between their own estimate and an algorithm's estimate or between another participant's estimate and an algorithm's estimate.

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 the estimate they wanted to determine their pay, "Would you like your (the other participant's) estimated rank or the model's estimated rank to determine your bonus?"

Results

Participants relied more on an algorithm than another person and even more than themselves. More participants chose to have their bonus pay determined by the estimate from the algorithm than another person, χ^2 (1, N = 206) = 118.14, $p < .001$, $r = 0.76$, consistent with reliance on algorithms. They even chose the algorithm's estimate over their own estimate, $\chi 2$ (1, N = 197) = 20.15, $p < .001$, $r = 0.32$. The option to choose one's own estimate attenuates reliance on algorithmic advice, as shown by the difference in effect sizes between the self and other conditions, $z = 6.62$, $p < .001$. See Figure 6. The results are consistent with the overconfidence literature, attesting to the excessive faith that people have in the quality of their own judgment.

Overall, there was a significant association between whether participants chose the algorithm and whether they had the option to choose themselves, $\chi 2$ (1, N = 403) = 27.35, $p < .001$, $r = 0.26$. The odds of choosing algorithmic advice over another person's advice is 3.73 times higher than choosing algorithmic advice over one's own estimate.[9] Results hold when choice is regressed on self/other in a logistic regression, $\chi 2(1) = 28.10$, $p < .001$.

Discussion

The results show that reliance on algorithmic advice is reduced when overconfidence is introduced to the estimation process. Nevertheless, more people chose to rely on the algorithm than their own judgment. In sum, Experiment 3 replicated reliance on algorithms using a different estimation task and showed that overconfidence attenuated the effect of reliance on algorithms. Replicating past work decreased the number of factors that varied between Experiment 3 and past work. However, the materials came from a condition that produced indifference between the algorithm's estimate and participant's own estimate, which could explain why there was not a complete reversal of the effect.

These results suggest that prior work may have found algorithm aversion partly because participants were asked to choose between their own estimate and an algorithm's, as a robust literature on overconfidence in judgment 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 these results suggest that advice from an algorithm may help. Experiment 4 moves beyond purely objective domains and examines another potential moderator to reliance on algorithmic advice: the subjectivity of the decision domain.

CHAPTER 6

Experiment 4: The effect of subjectivity on algorithm reliance

Experiment 4 tests the subjectivity of decisions as a moderator to reliance on algorithmic advice. Experiments 1A, 1B, 2, and 3 showed that participants relied more on an algorithm than other people, and even themselves. These results contrast aversion to algorithms found in more subjective domains such as book, movie, and joke recommendations (Sinha, & Swearingen, 2001; Yeomans, Shah, Mullainathan, & Kleinberg, 2014).

They also contrast results that suggest aversion to algorithms relative to doctors. Participants may have relied on a doctor more than an algorithm in Promberger & Baron (2006) because they saw the decision to rely on an advisor for a medical diagnosis as more subjective than objective. Prior to deciding which advisor to follow, 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 of the results may make the decision sound more subjective than objective, relative to descriptions that include the numeric test results. Doctors in the real world often provide numeric information and sometimes follow up with qualitative descriptions. For instance, patients may often ask a doctor for a translation of what numerical results mean, asking how their result compare to a "normal" result.

Prior work shows that participants viewed "expert systems" as generally more objective and "rational" than humans (Dijkstra, Liebrand, & Timminga, 1998), and this should affect the contexts in which people are more likely to trust them. As noted in the introduction, this finding is 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. The prediction was that in an objective domain like a financial decision, where people prefer more rational decision making (Hsee, Zhang, Yu, & Xi, 2003), they will rely more on algorithmic advice, and that in a subjective domain like dating, where people think intuition plays a role and prefer more affect-based decision making (Weber & Lindemann, 2008), they will rely more on advice from people.

Method

Participants

The final sample includes 303 decisions from 51 participants (27 women; 24 men; *Mdn*age = 28). Although I pre-determined a sample size of 50, the power comes from the number of decisions produced by the participants. Participants were asked to provide 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. I estimated that a total of 300 decisions would provide enough power for a chi square. Nonsense answers such as "yes," "no," and "wait" from one participant were removed from the decisions prior to coding for a total of 303 decisions.

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. These decisions were coded by research assistants blind to the conditions as subjective or objective. The dependent variable was the number of decisions in each condition coded as subjective or objective.

Procedure and Materials

To view the online questionnaire as a participant, follow this link:
https://berkeley.qualtrics.com/SE/?SID=SV_0uKhh3dBZE9bdd3&Q_JFE=0&Preview=Survey

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. The decisions were randomized using a Qualtrics survey. Two research assistants coded the participant-generated decisions as 1 = *very subjective*, 2 = *neither subjective nor objective* or 3 = *very objective*.

Dependent variable: Number of subjective and objective decisions. The dependent variable was the number of subjective decision and number of objective decisions.

Results

Inter-rater reliability for the coding of participant generated decisions by two coders was high ($\alpha = .85$). After breaking ties [10] and removing 27 decisions,[11] the final sample size for analysis was 276 decisions coded as objective or subjective. In order to test whether subjectivity moderated reliance on algorithmic advice, ratings were submitted to a 2 (subjective, objective) X 2 (person, algorithm) chi-square test.

There was an association between whether the decision was listed for a human or algorithm advisor and subjectivity, χ^2 (1, N = 276) = 92.66, $p < .001$, $r = 0.58$. See Figure 7. The odds of a decision being rated as objective (versus subjective) were 16.67 times higher when the decision was listed for an algorithmic rather than human advisor. The odds of the decision being rated as subjective (vs. objective) for the decisions were 16.56 times higher when the decision was listed for a human rather than algorithmic advisor. These results suggest that participants expected reliance on algorithmic advice for more objective decisions.

People listed a greater number of objective than subjective decisions for an algorithmic advisor, χ^2 (1, N = 144) = 87.11, $p < .001$, $r = 0.78$. People listed a greater number of subjective

than objective decisions for human advisors, χ^2 (1, N = 136) = 16.03, p < .001, r = 0.34, though this difference was smaller than for the algorithmic advisor, z = 5.72, p < .001. This result suggests that people may expect reliance on human advisors to vary less by domain than reliance on algorithmic advisors.

Discussion

In Experiment 4, subjectivity moderated expectations of reliance on algorithms. Participants expected greater reliance on an algorithmic advisor for more objective than subjective decisions and greater reliance on a human advisor for more subjective than objective decisions. The difference in expected reliance between subjectivity conditions was greater when people considered algorithmic advice than human advice. These results help explain why prior work finds algorithm aversion within subjective domains and Experiments 1A, 1B, 2, and 3 find that participants rely more on algorithms than other people and even themselves. Experiment 5 examines a potential mechanism for *why* people rely on algorithmic advice in Experiments 1A, 1B, 2, 3, and 4: the assumption that an algorithm is equivalent to expert opinion.

CHAPTER 7

Experiment 5: The effect of human expertise on algorithm reliance

Experiment 5 examines why people trust algorithms more than other people. It is possible that participants in Experiments 1A, 1B, 2, and 3 viewed advice from other participants as redundant with their own. In contrast, they may have viewed the algorithm as more analogous to an expert. If people perceive an algorithm as the distillation of expert opinion, they are likely to increase their reliance on it (Birnbaum & Stegner, 1979). Some work finds that people prefer to receive a medical diagnosis from a doctor (an expert) than an algorithm (Promberger & Baron, 2006). This work does not clearly distinguish between the separable influences of human vs. expert guidance. Conflicting evidence finds that people rely more on algorithmic "expert systems" than on a person (Dijkstra, Leibrand, Timminga, 1998; Dijkstra, 1999). Likewise, this work does not clearly distinguish between the separable influences of algorithmic vs. expert guidance. Experiment 5 seeks to explain why people rely on algorithmic advice by doing so.

Method

Participants

The final sample included 469 participants (195 women; 274 men; *M* age = 32). I determined the sample size a priori with the goal of including 508 participants, for 127 participants per cell. As a conservative estimate for a 4-cell ANOVA with planned contrasts, I needed 505 participants to detect an *interaction* in an ANOVA with an effect size $d = .25$ ($f = .125$) at 80%,[12] Participants were instructed that accurate responses would increase their chances of winning a $10 bonus. I collected data until I had 508 participants. Due to more people guessing the actual weight than participants had in past experiments, the sample size is marginally smaller (39 participants less than the planned 508).

Design

The experiment had a 4-cell (person vs. algorithm vs. expert vs. non-expert) between-subjects design. This experiment used the same paradigm and main dependent variable as Experiments 1A and 1B, WOA (weighting of advice) but included a photograph of a different person and an additional measure, expected usefulness of advice. A different photograph helps increase the generalizability of the results. I included an explicit measure, expected usefulness, to better test participants' expectations of advice prior to receiving any advice.

Procedure and Materials

Overview. The procedure was similar to Experiment 1A with three main changes: the addition of "expert" advisor conditions, a different photograph (see Figure 8), and an additional measure of perceived usefulness prior to receiving advice. Participants estimated the person's weight, read that they were about to receive advice an advisor and rated the expected usefulness of the advice. Then, they saw the advice and gave their final answer. In all conditions, the advice was 134 pounds, which came from averaging weight estimates provided by participants in the same past experiment with the Figure 1 photograph. The person in the photograph actually weighed 135 pounds.

To view the online questionnaire as a participant, follow this link:
https://s.qualtrics.com/SE/?SID=SV_8ApVS9QrtzdlTbD&Preview=Survey&BrandID=berkeley

Advisor manipulation prior to advice. Prior to seeing the advice, participants read that they were about to receive advice from either a "person," "algorithm," "person who is a weight-guessing expert," or "weight-guessing expert system."

Secondary dependent measure: Usefulness. I measured how useful participants expected the advice would be, prior to seeing the advice: **"**How useful do you think you will find the advice from the (person / algorithm / expert / expert system) in helping you make your final estimate?" ranging from 1 = *Not at all useful* to 7 = *Extremely useful*.

Advisor manipulation. Participants received the same advice, 134 pounds, described as an estimate from either another person, an algorithm, a human expert, or an expert system.

In the human condition, participants read, "The estimate of the person was: 134 pounds."

In the algorithm condition, participants read, "The estimate of the algorithm was: 134 pounds."

In the expert condition, participants read, "The estimate of the expert was: 134 pounds."

In the expert condition, participants read, "The estimate of the expert system was: 134 pounds."

Main dependent measure: Weight of Advice (WOA). As in Experiment 1A, 1B, and 3 the main dependent variable was how much participants relied on the advice.

Results

Weighting of advice (WOA). Do people equivocate algorithmic advice with expert advice? WOA was submitted to a one-way ANOVA with one four-level between-subjects factor.[13] Replicating prior results, participants relied more on an algorithm (M = .38, SD = .32) than another person (M = .30, SD = .31), p = .042, d = .25, as planned contrasts show.[14] See Figure 9. Yet, when the person was an expert, participants relied similarly on the advice from the algorithm (M =.38, SD = .32) and human expert (M =.35, SD = .29), p = .452. They also relied similarly on the expert (M = .35, SD = .29) and expert system (M =.34, SD = .31), p = .761. The omnibus test was not significant, $F(3, 465) = 1.43$, p = .235. See Figure 10 for a percentage histogram of participant's weighting of advice.

Usefulness. Similar to the WOA measure, participants expected the algorithm to provide more useful advice (M = 4.87, SD = 1.23) than the person (M = 4.47, SD = 1.47), p = .021, before receiving advice. See Figure 11. Oddly, participants also expected the expert to provide more useful advice (M = 5.23, SD = 1.20) than the algorithm (M = 4.87, SD = 1.23), p = .036. Participants seemed to focus on the label "expert" because they expected similar usefulness between the expert (M = 5.23, SD = 1.20) and expert system (M = 5.15, SD = 1.31, p = .621. The conditions differ marginally in usefulness overall, $F(3, 465) = 3.95$, p = .047. In exploratory analyses, usefulness ratings correlate with WOA, $r(467) = .13$, p = .005. Oddly, the correlation

is driven by correlations in the expert, $r(114) = .21$, $p = .026$, and expert system conditions, $r(119) = .23$, $p = .011$ (p-values for the person and algorithm conditions are .825 and .645 respectively).

Discussion

In addition to replicating algorithm reliance, Experiment 5 showed that participants relied similarly on algorithmic and expert advice. Figure 10 shows that reliance on algorithms is not driven by a few people updating completely to the advice. If anything, reliance is driven by fewer participants fully discounting the algorithmic advice (WOA = 0). This pattern is consistent across experiments. It also suggests that participants' reliance on expert advice is partly driven by people averaging between their original estimate and the expert's advice. Figure 10 shows that about 20% of participants who saw the expert advice averaged between their estimate and the advice.

The different pattern of results between WOA and usefulness is puzzling. If the explicit usefulness measure showed the same pattern as the WOA measure, then the results could explain why participants in Experiments 1A, 1B, 2, and 3 relied more on algorithms than other people: they may have equated algorithmic advice with expert advice. One potential, though unlikely, explanation is that participants' confidence in their original estimate created a ceiling effect for WOA, such that WOA was no different in the expert and algorithm conditions.

A ceiling effect is unlikely because WOA and usefulness are correlated in the expert condition but not in the person or algorithm conditions. If there was a ceiling effect, usefulness and WOA should not correlate in the expert condition, where usefulness is higher than the other conditions. Furthermore, the WOA in the algorithm condition is similar to the other experiments, perhaps even lower. See Table 2.[15] These results call for further testing to ensure the robustness of the results. Experiment 6 was created to further test the role of both expertise and subjectivity on algorithm reliance.

CHAPTER 8

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

Experiment 6 tests the interaction of a decision's subjectivity and the expertise of a human advisor on algorithm reliance. Experiment 6 moves from Experiment 4's focus on lay perceptions of reliance to participants' own reliance and seeks to better understand whether participants infer expertise from an algorithmic advisor. Experiment 6 serves to improve on Experiments 4 and 5 and to examine the interaction between subjectivity and expertise. Furthermore, it tests perceived subjectivity of a decision as a mediator to reliance on algorithms.

The decisions in Experiment 6 were based on decisions from prior work as well as decisions that people make in the real-world with the help of algorithms. The materials include decision problems similar to those found in prior work in order to better reconcile the results of algorithmic reliance with prior work that finds aversion. For instance, materials included a medical decision in the objective condition and a decisions about books and movies in the subjective condition.

The materials included decisions based on widely used, real-world algorithms: deciding which stock to invest in, who to date, and what to wear. In these decisions, companies rely on algorithms to provide their customers with advice. Betterment flouts its financial advice as algorithmic, OkCupid does not explicitly label the date recommendations as algorithmic, and Stitch Fix labels their clothing recommendations as from a "personal stylist." The prediction was that subjectivity would moderate reliance on algorithmic advice relative to another person and that regardless of subjectivity, participants would rely on expert advice.

Method

Participants

The final sample included 555 participants. In order to detect an attenuated interaction, I first estimated an effect of subjectivity as $d = .35$.[16] In order to detect that effect size at 80% power, a total of 260 participants were needed for a 2-cell design. Per Simonsohn (2014), I 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. Subjectivity was within-subjects and expertise was between-subjects. The main dependent variable was how much participants relied on advice from an algorithm relative to another person (or an expert).

Procedure and Materials

Overview. Participants were asked to consider a total of twelve decisions and to "imagine you are about to make each decision." See Table 3. They read six subjective decision

problems and six objective decision problems where the order of subjectivity was counterbalanced.

The dependent variable was how much they thought they would rely on advice from an algorithm or another person. At the end of the survey, participants rated first the subjectivity and then the importance of each decision. Perceived subjectivity was used as a mediator to reliance on algorithms. Measuring importance allowed me to statistically control for a potential confound between the subjectivity conditions.

To view the online questionnaire as a participant, follow this link:
https://berkeley.qualtrics.com/jfe/preview/SV_51NyEocyv4WEVAF

Subjectivity manipulation. Participants were asked about both subjective and objective decision problems that were pre-tested with another sample of mTurkers.

Subjectivity manipulation. Participants read about either non-experts (another person) or an expert. Instead of using the label expert, participants read about specific professional roles that signaled expertise (that the individual made the decision for a living or gave advice on the decision for a living). See Table 3.

Dependent variable. For each decision problem, participants imagined that they were about to make the decision and rated how much they expected to rely on advice from either an algorithm or another person (or expert), "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.*

Manipulation checks. At the end of the survey, participants rated how subjective/objective they thought each decision was. As a clarification of what it means for a decision to be subjective/objective, the participants were told, "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 measure. At the end of the survey, participants rated how consequential they thought each decision was, "How important do YOU think the following decisions are?" on a scale from 1 = *not at all important* to 7 = *very important.*

Results

In line with my pre-registered analysis, I included all decision problems in the analyses except for the moving decision, as the perceived subjectivity rating for it was no different than the mid-point of the scale. All of the subjective decisions were appropriately rated less than the mid-point of the scale (more subjective) and all of the other objective decisions were rated greater than the mid-point of the scale (more objective).

Reliance on Algorithmic Advice. I averaged reliance on algorithmic advice relative to human advice across decisions and submitted them to a 2 (subjectivity: subjective vs. objective) X 2 (expertise of person: expert vs. non-expert) mixed ANOVA with expertise as a between-subjects factor and subjectivity as a within-subjects factor. There is a main effect of subjectivity such that participants relied more on algorithmic advice than human advice for the objective decisions (M = 3.83, SD = 1.39) than subjective decisions (M = 2.99, SD = 1.24), $F(1, 553) = 225.81$, $p < .001$. A main effect of expertise was not of interest to my predictions.

There is an interaction between subjectivity and expertise, $F(1, 553) = 122.83$, $p < .001$. In the non-expert condition, participants relied more on algorithmic advice in the objective (M = 4.39, SD = 1.24) than the subjective condition (M = 2.93, SD = 1.27), $F(1, 553) = 342.71$, $p < .001$. Yet, reliance on algorithmic advice decreases in the expert condition (objective: M = 3.26, SD = 1.29; subjective: M = 3.04, SD = 1.21), $F(1, 553) = 7.74$, $p = .006$.

Controlling for Importance of Decisions. Participants rated the objective decisions as more important (M = 5.66, SD = 0.84) than the subjective decisions (M = 4.57, SD = 0.97), $t(554) = 28.47$, $p < .001$,[17] which suggests that subjectivity and importance are not independent, and analyses should control for importance. The above results become stronger when controlling for importance. I averaged importance across decisions and submitted reliance ratings to the same mixed ANCOVA with importance as a within-subjects covariate.

There is an effect of importance, $F(1, 1104.01) = 8.34$, $p = .004$, and when 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, $F(1, 550.98) = 124.98$, $p < .001$. See Figure 13. Again, 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), $F(1, 719.23) = 342.71$, $p < .001$. The magnitude of simple effect in the expert changed to non-significant, as suggested by the larger interaction: participants relied *similarly* on algorithmic advice in the objective (Adjusted M = 3.19, SE = 0.08) than subjective condition (Adjusted M = 3.12, SE = 0.08), $F(1, 719.23) = 0.87$, $p = .352$.

Within-Subjects Mediation Analysis. I tested whether perceived subjectivity mediated the relationship between the subjectivity conditions and reliance on algorithmic advice relative to advice from another person in the non-expert condition. 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 & Dunning (2009) based on Judd, Kenny, & McClelland (2001).

First, I tested that the subjectivity condition 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 = 5.07, SE = 0.08) than in the subjective condition (Adjusted M = 3.21, SE = 0.08), $F(1, 455.25) = 43.46$, $p < .001$, controlling for importance.

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

conditions, perceived subjectivity correlates with relative reliance on algorithmic advice, controlling for importance, objective: $pr(276) = .168$, $p = .005$, subjective: $pr(276) = .492$, $p < .001$.[18] The more participants perceive a decision as objective, the more they will rely on algorithmic advice over a non-expert.

I then examined participants' ratings of importance, relative reliance on algorithmic advice, and perceived subjectivity within the subjective decisions relative to the objective decisions along three difference scores (objective minus subjective ratings). If the effect of the 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, $pr(275) = .367$, $p < .001$.

Between-Subjects Mediation Analysis. The nature of this experimental design allowed me to run a follow-up mediation analysis by setting aside the second subjectivity condition participants saw and running a between-subjects analysis on the first subjectivity condition participants saw (Critcher & Dunning, 2009).[19] At the individual level, perceived subjectivity correlates with reliance on algorithmic advice (relative to advice from another person), controlling for importance, $pr(276) = .305$, $p < .001$.

At the aggregate level, subjectivity condition predicts relative reliance on algorithmic advice, controlling for importance $\beta = .89$, $t(x) = 5.38$, $p < .001$. Importance itself is not significant, $\beta = -.05$, $t(x) = -.65$, $p = .519$. When I entered perceived subjectivity into the model, perceived subjectivity predicts relative reliance on algorithmic advice, $\beta = -.167$, $t(x) = -3.03$, $p = .003$, but there is a less significant effect of subjectivity condition, $\beta = .60$, $t(x) = 3.13$, $p = .002$, which suggests partial mediation. The effect of subjectivity on relative reliance on algorithmic advice decreased with perceived subjectivity.

Even though the results suggested a partial mediation, I used the Sobel test to test the mediation model.[20] I checked that subjectivity correlates with relative reliance on algorithmic advice, controlling for importance, $pr(276) = -.308$, $p < .001$. (Condition is coded as objective 0, subjective 1). The subjectivity correlates with perceived subjectivity when controlling for importance, $pr(276) = -.513$, $p < .001$. I then used the unstandardized coefficients and standard errors from subjectivity predicting perceived subjectivity, $\beta = -1.77$, $SE = .18$, and perceived subjectivity predicting relative reliance on algorithmic advice, $\beta = .26$, $SE = .05$ for the Sobel test. The Sobel test confirmed the significance of the mediation model ($z = -3.18$, $p = .001$).

Discussion

Experiment 6 provides evidence that subjectivity and expertise of the human advisor influence reliance on algorithmic advice relative to human advice. It complements the findings of Experiment 4 by providing evidence that people, themselves, are willing to rely more on algorithmic advice for more objective decisions. It also clarifies the findings from Experiment 5 and suggests that expertise of the human advisor moderates relative reliance on algorithms. Although Experiment 5 found conflicting patterns for usefulness and WOA, Experiment 6

suggests that participants rely more on expert than algorithmic advice, as participants had expected they would in Experiment 5.

Most importantly, this experiment helps reconcile the results of algorithm reliance with some results that suggest algorithm aversion. Experiment 6 shows that regardless of the decision's subjective nature, participants relied more on an expert than algorithm. This helps explain why participants in past work preferred a medical diagnosis from a doctor than algorithm (Promberger & Baron, 2006), even in a medical domain.

Experiments 1A, 1B, 2, 3, 4, 5, and 6 suggest that people are willing to rely more on advice from algorithms than non-experts in objective domains. One question that remains is whether they rely on algorithmic advice as much as they should. Experiment 7 examines whether people rely on algorithmic as much as they should.

CHAPTER 9

Experiment 7: Do people rely on algorithmic advice as much as they should?

Experiments 1A, 1B, 2, 3, and 5 control for the accuracy of advice because they provide the same advice across conditions. Experiment 7 allows the accuracy of human and algorithmic advice to vary, as it often does in the real-world. Describing how much information the advisor uses allows the participant to receive normative information. Providing normative information allows for a comparison of how much people rely on the advice relative to how much they should. Crowds are often wiser than individuals as the robust work on the wisdom of crowds demonstrates (Galton, 1907; Surowiecki, 2004), especially as the group grows (Einhorn et al. 1977).

The literature on wisdom of crowds and advice-taking suggests a normative benchmark for optimal reliance: participants should rely more on advice as the advice uses a greater number of estimates (Mannes, 2009; Soll & Larrick, 2009). Thus, even with limited information about the advice, people should average between their own estimate and the estimate from one other person, a WOA of .5. When advice comes from more than 300 people, they should essentially weight it 100%. Providing participants transparent information about the input of the advice (number of individual estimates used to inform the advice), allows participants to determine how much they ought to weight advice. Transparency of the input allows for a comparison of how much people rely on the advice to how much they *should* rely on it.

Method

Participants

The final sample included 671 participants (357 women; 314 men; *Mdn* age = 33). I determined a sample size of 620 a priori. A sample of 620 participants was necessary to detect an interaction with an effect size $d = .25$ ($f = .125$) at 80% power. The effect of reliance on algorithms was $d = .35$ from past experiments I had run with a weight guessing task at the time. I powered the analysis to detect a smaller effect in order to appropriately power a covariate (experience with a failed algorithm). [21] Participants were instructed that accurate responses would increase their chances of winning a $10 bonus.

Design

The experiment had a 2-cell (advisor: person vs. algorithm) between-subjects design. As in prior experiments, the experimental manipulation varied whether participants saw advice labeled as having come from an algorithm or from human advisors. This experiment used a similar paradigm to Experiment 5's non-expert condition with the same dependent variables, WOA (weighting of advice) and usefulness. Instead of guessing weight, participants guessed the age of a different person in a photograph, which should help increase the generalizability of the results. The experiment also provided more detail about how the advice was produced.

Procedure and Materials

Overview. The procedure was similar to Experiment 5 with one main change: participants guessed the age of a different man in a photograph. See Figure 13. Participants estimated the person's age, read that they were about to receive advice from an advisor and some details about the advice, and rated the expected usefulness of the advice. Then, they saw the advice and gave their final answer.

The advice in the human condition was a randomly chosen estimate from a group of 314 participants in the same past experiment as the weight estimates. The advice was 66 years old in the algorithm condition, the average age from the 314 past participants. The advice also happened to be an actual estimate from multiple individual past participants. The person in the photograph was actually 63 years old.

To view the online questionnaire as a participant, follow this link:

https://berkeley.qualtrics.com/jfe/preview/SV_3z1IdaSTRFiRU5D

Advisor manipulation prior to advice. Prior to seeing the advice, participants read that they were about to receive advice from a "person," or an "algorithm." In order to provide participants with a normative benchmark, they read the number of estimates used to produce the advice.

In the human condition, participants read that they would see advice from, "a randomly chosen participant from a pool of 314 participants who took a past study."

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."

Secondary dependent measure: Usefulness. As in Experiment 5, I measured how useful participants expected the advice would be, ranging from 1 = *Not at all useful* to 7 = *Extremely useful*.

Advisor manipulation. Participants then received different advice, depending on their condition.

In the human condition, participants read, "The estimate of another person, a randomly chosen participant from a pool of 314 participants who took a past study, was: X." Because another randomly selected participant is likely (on average) to have an estimate as accurate as one's own, it is normative to average one's own estimate with this advice (WOA = .5).

In the algorithm condition, participants read, "The estimate of an algorithm, based on estimates of 314 participants who took a past study, was: 66." An algorithm based on such a large number of useful inputs is likely to be quite accurate and optimal weighting of this advice (WOA) would be close to 1.

Main dependent measure: Weight of Advice (WOA). As in Experiment 1A, 1B, 3, and 5 the main dependent variable was 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), $F(1, 670) = 17.68$, $p < .001$, $d = .33$.[22] But did participants rely on the algorithmic advice as much as they *should* have, given that they saw the number of estimates on which each advisor's estimate was based? When reliance was 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 = .26, SD = .27) they did from the person (M = .66, SD = .34), $F(1, 670) = 275.08$, $p < .001$, $d = 1.30$. See Figure 14.

Usefulness. Similar to the WOA measure, people expected the algorithm to provide more useful advice (M = 5.09, SD = 1.20) than the other person (M = 4.08, SD = 1.72) before seeing advice, $F(1, 670) = 80.24$, $p < .001$, $d = 0.68$.[23]

Discussion

Experiment 7 replicated reliance on algorithms: participants relied more on an algorithm than a person in an absolute sense. When reliance was compared to a normative benchmark; however, participants underweighted the algorithmic advice more than they underweighted advice from a person. Although participants were not averse to algorithmic advice, there was still room for people to achieve higher accuracy by relying more on it. Nevertheless, it is possible that the provision of additional details about the source of the advice they were getting allowed participants to weight the advice more effectively.

CHAPTER 10

Discussion and Future Directions

Summary of Results

The results in these experiments suggest that people are willing to rely on algorithmic advice more than advice from other people, counter to the widespread conclusion of algorithm aversion. They highlight the contexts in which people are most likely to use advice generated by algorithms and thus increase their accuracy. Participants relied more on algorithmic than human advice for objective estimates and forecasts (Experiments 1A and 1B), for which they had limited experience. In Experiment 1A, WOA was positively correlated with numeracy in the algorithm condition.

Participants relied on an algorithm more than another person even when they chose directly between the two in a joint evaluation (Experiment 2). Experiments 3 and 4 found moderators of reliance: the option to choose one's own estimate and the subjectivity of the decision domain. Although the option to rely on one's own estimate attenuated reliance on algorithmic advice, people still relied more on the algorithm than their own estimate (Experiment 3). Participants expected greater reliance on algorithmic advice for more objective than subjective decisions but expected greater reliance on human advice for more subjective than objective decisions (Experiment 4).

Although Experiment 5 found that participants expected more useful advice from an expert than an algorithm, the reliance measure did not show the same pattern. Experiment 6 helped to clarify the results from Experiments 4 and 5. Participants relied more on advice from an algorithm than another person for objective compared with subjective decisions but relied more on an expert, regardless of subjectivity.

Experiment 7 introduced a normative benchmark and provided participants with more information about the advisor's inputs. Participants underweighted algorithmic advice more than they underweighted advice from another person. Although people are willing to listen to algorithmic wisdom, there is still room for people to glean greater accuracy from algorithmic advice.

Theoretical Implications

The 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). 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 4. Specifically, these results suggest that people will rely on algorithms more than another person in an objective domain. When the boundaries of those conditions change, so does reliance on algorithmic advice. People will less on algorithms when they can rely on themselves, rely on an expert, or when the domain is more subjective.

Although my results may appear to most directly contradict the work from Dietvorst et al. (2015), a closer examination of the control condition in that work shows that prior to seeing iterations of performance feedback on the algorithm, participants were willing to rely on the algorithm or were indifferent between their answer and the algorithm's. As discussed in the introduction, many consequential decisions are made regarding forecasts with long time horizons and 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 dissertation suggest that one simple way to increase adherence to advice in situations where it is heavily discounted is to provide advice from an algorithmic advisor. This is especially the case if people are already faced with the option to get advice from another non-expert or if they are making a decision in an objective domain.

Practical Implications

My dissertation has implications for understanding how people perceive lessons produced by big data. 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 deals with it (Chan, 1979). For instance, the benefits of algorithms to more accurately forecaste future political events (Baron et al., 2014) should have the U.S. Intelligence Communities and policy makers jumping to take advantage of algorithms for improved policy decisions. However, the extant literature predicts that even if algorithms produce more accurate information, it would fall on deaf ears. The results from this dissertation provide a more hopeful picture.

As organizations invest in the collection, analysis and exploitation of "big data," they use algorithms to sift through the information and produce algorithmic advice. As advances in technology allow organizations to better access and analyze "big data," they adjust how they manage it in order to collect more of it, collect a wider variety, all at a faster and faster rate (Laney, 2012). Google (as used in Reips & Matzat, 2014) and Twitter (as used in Reips & Garaizar, 2011) already share enormous quantities of data publicly.

And 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). The results in this dissertation shed light on how to best make use of efforts to collect data and produce more accurate information with algorithms by testing when people are most likely to rely on algorithmic advice.

Limitations

Experiments 1 through 3 compare reliance between a lay person's advice and an algorithm's advice (the purpose of which was to avoid potentially confounding expertise and human advice). However, a potential limitation to these studies is that people may sometimes find themselves in situations in which they compare expert and algorithmic advice. In fact, Experiment 5 showed that when expert advice is available, participants relied on it more than algorithmic advice.

Although there are situations in which expert and algorithmic advice are both available, say in the case of financial advice from a financial advisor or algorithm, expert advice is often costly and less accessible to the masses than algorithmic advice. For instance, the Common Cents Lab is creating a financial advisement application that provides algorithmic advice to lower income communities who cannot afford the time or money to meet with a human financial advisor (Berman, 2015).

Experts are not the only source of human advice and research should not overlook the comparison of algorithms to lay advice. Often, people ask friends and family members, who are clearly not experts, for advice. For instance, when things are broken, from our hearts to our home appliances, we call our friends, partners, or parents for potential solutions.

The comparison between an algorithm and another person, not an expert, is more relevant to many decisions, mostly because of cost and time constraints. We don't always call a therapist or a plumber right away, simply because access to both of those experts is painfully costly and often delayed. It sounds improbable that many would have an on-call therapist or plumber. An algorithm can replace an on-call expert because it is more affordable. The algorithm has the ability to replace the expert because it can often prove a more accessible form of advice.

Moreover, it's unclear how much participants *should* weight advice from an expert, as there does not seem to be an appropriate normative standard as there is for a random individual or advice based on multiple individuals. That is not to say that an expert comparison is unnecessary for future work, but there are good reasons to examine people's perceptions of algorithms relative to other lay people.

Future Directions

Theory of Machine. There are many exciting future directions for this research. Research on what I call "theory of machine" is needed to keep up with the rapid pace of technological advancement that injects algorithms into many aspects of our lives. 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 in other people (and even in other things, such as when people anthropomorphize inanimate objects) (Dennett, 1987).

Social psychology follows in that wake by examining 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 tests how anthropomorphizing machines, like a self-driving car, influences our trust in them (Waytz, Heafner, & Epley, 2014). Future work is needed to examine people's perceptions about algorithmic *input*, *process*, and *output*, a kind of theory of mind, that examines perceptions of how algorithmic and human judgment differ, rather than how we impart human judgment on algorithms.

This dissertation examines how people perceive advice, or output, from algorithmic judgment. There is much more work to do on this topic, including how aspects of the algorithm and human advisors could affect reliance on algorithms. First, I intend to test how expertise of

the decision maker influences his or her reliance on algorithmic advice. Second, the complexity of the algorithm's explicit procedure could influence reliance on algorithmic advice.

Another potential future direction is to examine algorithm reliance in domains that are more relevant to the decision maker's personal identity. This dissertation mostly focused on people's reactions to algorithmic advice about estimates and forecasts they made about the world around them. I plan to examine perceptions of algorithmic advice when people make predictions about their own performance, such as on a test. Negative feedback can be difficult to hear. However, the source of the feedback (human or algorithm) could influence 1) how much people listen to it and 2) the accuracy of their predictions about future performance.

Most importantly, my future work will examine different mechanisms for reliance on algorithmic advice by leveraging lay perceptions about the differences between algorithmic and human advice (i.e., theory of machine). While this dissertation focused on perceptions of algorithmic output, I categorize future directions that test mechanisms of reliance into perceptions about input and processing. First, I will test whether people view algorithmic judgment as constrained compared with human judgment. Second, I will test whether people expect algorithms to access a greater amount of informational input.

Expertise of Decision Maker. I am working to test whether decision makers in the U.S. government who belong to a leadership organization that focuses on improving national security rely on algorithmic advice more than other people. This experiment includes a comparison of decision makers within the government with a more lay population in order to test whether expertise moderates reliance on algorithms.

The Good Judgment Project (GJP), on which the author was a collaborator, was a federally funded forecasting tournament aimed at developing better forecasts of world events. The goal was to influence policy makers; the researchers on the project leveraged the wisdom of crowds through simple aggregation of forecasts from the tournament as well as more complex algorithms to improve the accuracy of forecasts (Mellers et al., 2014). One outstanding question was whether decision makers in the government and analysts in the CIA would listen to advice from algorithms, even if aggregating individual forecasts might improve predictive accuracy. Including forecasts taken from the tournament allows for a more specific test of whether people who work in the government are willing to listen to advice from algorithms relative to forecasts from the tournament.

As reliance was non-significant in Experiment 1B for the task with which participants had experience, future work will also *manipulate* the perceived expertise of participants. Participants will either receive practice forecasts to research on their own or receive no practice before a final bonus round of forecasts. This experiment will test the role of expertise in a domain as a moderator to reliance on algorithmic advice.

Complexity of an Explicit Procedure. 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 experiments will manipulate the opacity of an algorithm's process and whether people can understand it. 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 this to be the case. It shows that acceptance of climate change increases when participants are provided mechanistic information about how global working works (Ranney & Clark, 2016). Because the calculations of different algorithms may not be based on broader scientific consensus, a better understanding of the mechanism that drives an algorithm may encourage people to rely on it less if they are knowledgeable enough to question it and consider potential flaws.

It might also be the case that people may rely on algorithmic advice if they can access its calculations or procedures, even if they cannot comprehend them. Specifically, 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's. A future experiment will manipulate the complexity of algorithmic calculations or procedures using either simple or more complex notation. See the supplementary materials for an experiment on this topic.

Negative Feedback. These results suggest that providing advice from an algorithmic advisor helps to increase reliance on the provided advice. These implications are especially exciting when applied to situations in which people are particularly resistant to advice: negative feedback. For instance, people remain optimistic about their future performance, even when provided explicit negative performance feedback (Ferraro, 2010; Hacker et al., 2000; Helzer & Dunning, 2012) by questioning its accuracy or relevance (Sheldon, Dunning, Ames, 2014).

Future work will leverage algorithmic advice as an intervention to provide a context where people are less likely to psychologically "escape" negative feedback. First, it will test whether people update more to negative feedback on a test from a person or algorithm. Second, it will test whether people prefer negative feedback from an algorithm or person.

Negative feedback from an algorithm may help the poor performer avoid any shame associated with another person's knowledge of that poor performance. On the other hand, negative feedback from a person may provide the poor performer with the opportunity to receive feedback cloaked in positivity which could soften the blow of the negative content. Perceptions of algorithmic advice when negative feedback is likely and subsequent willingness to update to more realistic performance expectations can contribute to work on algorithms, performance, and overconfidence.

Perceived Differences between Algorithmic and Human Judgment. I intend to test perceived differences in the processing of information between algorithmic and human judgment: whether people think that algorithms are constrained in how holistically they process the same amount of information relative to human judgment. 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 algorithmic advice because they think it can process the information more effectively. However, when both advisors access the same five cues plus the patient's entire medical history, participants may rely more on the advice from the person because they expect the person will take a more holistic view.

Participants may view human judgment as more holistic and prefer a human advisor for environments that provide more 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 (Heden 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.

A second lay perception to test is the difference in information accessed by the advisors, or input to the advice. People may rely more on algorithmic advice because they assume that an algorithm often has access to more information than an individual does when constructing advice. One 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). Although Experiment 7 suggests that people are less sensitive to the amount of information behind advice than they normatively should be, assumptions about the amount of information algorithms access could drive reliance.

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.

Conclusion

Big data are changing the way organizations function and communicate, both within and external to the organization. Information technologies may change our exposure to big data but how do big data, and the algorithmic advice gleaned from them, change how we see the world? Organizations have an opportunity to learn from the ever-increasing amount of information they can access, and further, to communicate information from the data with others.

Without understanding how people incorporate information from algorithmic advice into their decisions about the world, organizations are spending precious time and effort to produce output that may not be utilized adequately by their own decision makers and employees, let alone their clients. Understanding where people focus their attention within the deluge of information that floods our email inboxes, Internet searches, and discussions is useful for any organization or decision maker. The results of my experiments shed light on how to best leverage algorithmic advice efficiently. Uncovering the mechanisms that encourage reliance on algorithmic over human advice may help organizations transition more effectively to the world of big data and algorithms.

REFERENCES

Axe, D. (2011, February 7). One in 50 troops in Afghanistan is a robot. *Wired*. Retrieved from http://www.wired.com/

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.

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

Berman, K. (2015, November 16) Metlife Foundation and Duke partner to invest in increase financial well-being of low to middle Americans [Web log post]. Retrieved from http://advanced-hindsight.com/metlife-foundation-and-duke-partner-to-invest-in-increase-financial-well-being-of-low-to-middle-americans/

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.

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

Conniff, Richard. (2011, March). What the Luddites Really Fought Against. *Smithsonian Magazine*. Retrieved from http://www.smithsonianmag.com/

Copeland, M. (2013, October 22). Where humans will always beat the robots. The Atlantic, Retrieved from http://www.theatlantic.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.

Ferraro, P. J. (2010). Know thyself: Competence and self-awareness. *Atlantic Economic Journal*, 38, 183–196. doi:10.1007/s11293-010-9226-2

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

Gross, T. (Producer). (2015, August 20). How Close Are We Really To A Robot-Run Society? [Audio Podcast]. Retrieved from http://www.npr.org/

Hacker, D. J., Bol, L., Horgan, D., & Rakow, E. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology*, *92*, 160–170. doi:10.1037/0022-0663.92.1.160

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

Helzer, E. G., & Dunning, D. (2012). Why and when peer prediction is superior to self-prediction: The weight given to future aspiration versus past achievement. *Journal of Personality and Social Psychology, 103*, 38–53. doi:10.1037/a0028124

Hodge, N. (2012, June 13). In the Afghan War, a little robot can be a soldier's best friend: Some are as adorable as WALL-E, and injured ones now to bot rehab. *The Wall Street Journal*. Retrieved from http://www.wsj.com/

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/

Kamenetz, A. (2013, July 13). The Four things people can still do better than computers. *Fast Company*. Retrieved from http://www.fastcompany.com/

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

Koba, M. (2013, January 24) High Frequency Trading. *CNBC.* Retrieved from http://www.cnbc.com/

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

Lewis, M. (2014). *Flash boys: a Wall Street revolt*. WW Norton & Company.

Lowenthal, D. (1993). *Preference reversals in candidate evaluation*. Working paper. Carnegie Mellon University, Pittsburgh, Pennsylvania.

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.

Markoff, J. (2014, August 11) 'Beep,' says the bellhop: Aloft Hotel to begin testing 'Botlr,' a Robotic Bellhop. *New York Times.* Retrieved from, http://www.nytimes.com/

McKinley, J. (2014, April 22). With farm robotics, the cows decide when it's milking time. *The New York Times.* Retrieved from http://www.nytimes.com/

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.

Mullen-Fortino, M., DiMartino, J., Entrikin, L., Mulliner, S., Hanson, C. W., & Kahn, J. M. (2012). Bedside nurses' perceptions of intensive care unit telemedicine. *American Journal of Critical Care, 21*(1), 24-32.

Murphy, M. (2015, August 25) Robots are using wikiHow to figure out how to cook us breakfast. *Quartz.* Retrieved from http://qz.com/

Nowlis, S. M., & Simonson, I. (1997). Attribute-task compatibility as a determinant of consumer preference reversals. *Journal of Marketing Research, 34*, 205-218.

O'Toole, J. (2014, August 7) Americans are warming to self-driving cars: Several states have passed laws allowing testing of self-driving cars on public roads. *CNN*. Retrieved from http://money.cnn.com/

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

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/

Richtel, M. (2013, April 27). How Big Data Is Playing Recruiter for Specialized Workers. The *New York Times.* Retrieved from http://www.nytimes.com/

Roadside bombs 'No. 1 threat' to troops in Afghanistan. (2009, July 9). *CNN.com*. Retrieved from http://www.cnn.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

Sheldon, O. J., Dunning, D., & Ames, D. R. (2014). Emotionally unskilled, unaware, and uninterested in learning more: Reactions to feedback about deficits in emotional intelligence. *Journal of Applied Psychology*, *99*(1), 125.

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.

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

Simonsohn, U. & Nelson, L. (2014, September 17) Thirty-somethings are shrinking and other U-shaped challenges [Web log post]. Retrieved from http://datacolada.org/

Simonsohn, U. (2014, March 12) No-way interactions [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

Tran, M. (2016, March 13). Go humans: Lee Sedol scores first victory against supercomputer. *The Guardian*. Retrieved from http://www.theguardian.com/

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.

Weinstein, M. (2013, December 15). Amazon drones: Orwellian mayhem? *The Huffington Post*, Retrieved from http://www.huffingtonpost.com/

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, Mullainathan, & Kleinberg (2014, April). *Recommendations vs. recommender systems: The rise of robots*. Presented at the Kellogg-Booth Student Symposium, Chicago, IL.

FOOTNOTES

[1] The first ten participants were paid $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.

[2] 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. The results remain significant when follow-up analyses exclude those participants. In all other studies, I pre-registered how I winsorized participants' WOA.

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 his or her WOA was less than -1. No one guessed the same weight as the advice (163) on his or her first estimate.

After winsorizing, sixty-eight participants (33.7%) did not change their estimates from Time 1 to Time 2 (WOA = 0), one hundred and thirty-one (64.9%) changed their estimate closer to the advice, and three (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.

[3] As pre-registered, for each task, participants were 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 was only used for the weight task). Participants' WOA were winsorized to 1 if they moved past the advice (WOA > 1) but still had a WOA less than 2 and were winsorized to 0 if they moved away from the advice (WOA < 0) but still had a WOA greater than -1.

Weight guessing: The 2 participants dropped from the weight guessing task 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.

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

Political Forecasts: The 3 participants dropped from the first political 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. The 1 participant dropped from the second political forecasting task had a WOA < -1.

[4]When the political 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: 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, as a few reported in the open-ended response at the end of the survey.

[5] Note, Cohen's d is double the effect size f found in G*power.

[6] I winsorized 8 participants' WOA to 1 because their WOA was greater than 1. No one 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.

[7] People do not frequently interact with algorithms and receive instantaneous performance feedback. Thus, it is worth examining reliance on algorithmic advice prior to performance feedback. For instance, when organizations use algorithms to make hiring decisions, they may not know how successful those hires are until a year or two later.

[8] I collected more participants than planned (447) in order to meet the goal of 400 for the final sample. 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 403.

[9] When participants had the choice of another person, significantly fewer participants (12.1%) chose the other participant than expected, $z = -3.2$, $p < .01$, but no more participants chose the algorithm (87.9%) than expected, $z = 1.7$, $p > .5$. When participants had the choice of themselves, significantly more participants (34.0%) chose themselves than expected, $z = 3.3$, $p < .01$, but no more chose the algorithm (66.0%) than expected, $z = -1.8$, $p > .5$.

[10] 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 (stating it was too vague a decision), the other coder's response of 1 or 3 was used (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).

[11] A total of 27 decisions were removed prior to analysis for: 11 decisions where one coder left the decision blank and the other rated it a 2, 15 decisions where both coders rated it a 2, and one decision where both left it blank.

[12] After pre-registering but prior to any analyses, I determined that the more appropriate tests of the hypothesis was the contrasts from a 4-cell ANOVA.

[13] Although I pre-registered a 2 X 2 ANOVA as the analysis of choice, prior to analyzing the data I determined that a one-way ANOVA with one four-level between-subjects factor was more appropriate. The hypothesis of interest is really the comparison of each condition to each other, specifically the expert and algorithm comparison. For the sake of a comprehensive reporting, I include the 2 X 2 ANOVA interactions for both WOA, and usefulness: the interaction was non-significant for WOA, $F(1, 465) = 2.77$, $p = .097$ and only marginally significant for usefulness, $F(1, 465) = 3.95$, $p = .047$.

[14] As pre-registered, I winsorized 34 participants' WOA to 1 who had WOA greater than 1 (moved towards and past advice) and winsorized two to 0 who had WOA less than 0 (moved away from advice). Unlike past studies, where no one guessed the same as the advice, 4 participants guess the same as the advice. Prior to analyzing the data, I excluded them.

Similarly, 65 participants guessed the actual weight, where virtually none had in past studies. These participants are included in the reported analysis. When I exclude them, the trends in the results are consistent and actually stronger than those reported.

[15] Exploratory analyses show that on average, participants in this study were 68% confident in the accuracy of their original estimate. Time 1 confidence does not correlate with WOA, $r(467) = .01$, $p = .887$. Oddly, confidence in original estimates (prior to exposure to advisor) differed by condition, $F(3,$

465) = 3.85, $p = .010$. Comparisons show that confidence differed between the algorithm condition and expert system condition, $p = .020$.

[16] This effect size is based on the average effect size of reliance on algorithms compared with people all studies run to-date that use WOA.

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

[18] In an analysis additional to the pre-registered ones, the correlation between subjectivity and reliance on algorithmic advice is stronger within the subjective condition, $z = 4.31$, $p < .001$, probably because subjectivity varied more acros those decision problems.

[19] An additional analysis with just the first subjectivity condition showed results which were redundant with the above ANCOVA. Controlling for importance, reliance on the algorithm was higher 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: $M = 3.15$, $SD = 1.23$), $F(1, 550) = .43$, $p = .513$.

[20] I had pre-registered to use the Sobel test if the mediation was a full mediation, but it seemed appropriate to use the Sobel test, even with the partial mediation.

[21] The first exploratory measure was 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 there were too few people to include experience with an inaccurate algorithm as a covariate to the analyses.

[22] I collected 891 survey responses 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 responses that were non-sense "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 WOAs were winsorized.

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

[23] A similar pattern of usefulness and WOA suggests that a follow up experiment is needed for Experiment 5 so that the two measures correlate.

Table 1. *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 2. *Mean WOA in algorithm and human conditions across weight guessing tasks.*

Experiment with weight estimate task	**Algorithm**	**Human**	**Effect Size**
1A	*M* = .45, *SD* = .37	*M* = .30, *SD* = .35	*d* = .39
1B	*M* = .46, *SD* = .37	*M* = .18, *SD* = .28	*d* = .85
2	*M* = .50, *SD* = .37	*M* = .35, *SD* = .36	*d* = .44
5	*M* = .38, *SD* = .32	*M* = .30, *SD* = .31	*d* = .25

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. *Moderators in current experiments can explain contradictory results in past work*

	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., 2015	***Rely on Expert*** Not 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: *Confounds algorithm and expertise by calling algorithm an "expert system"

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

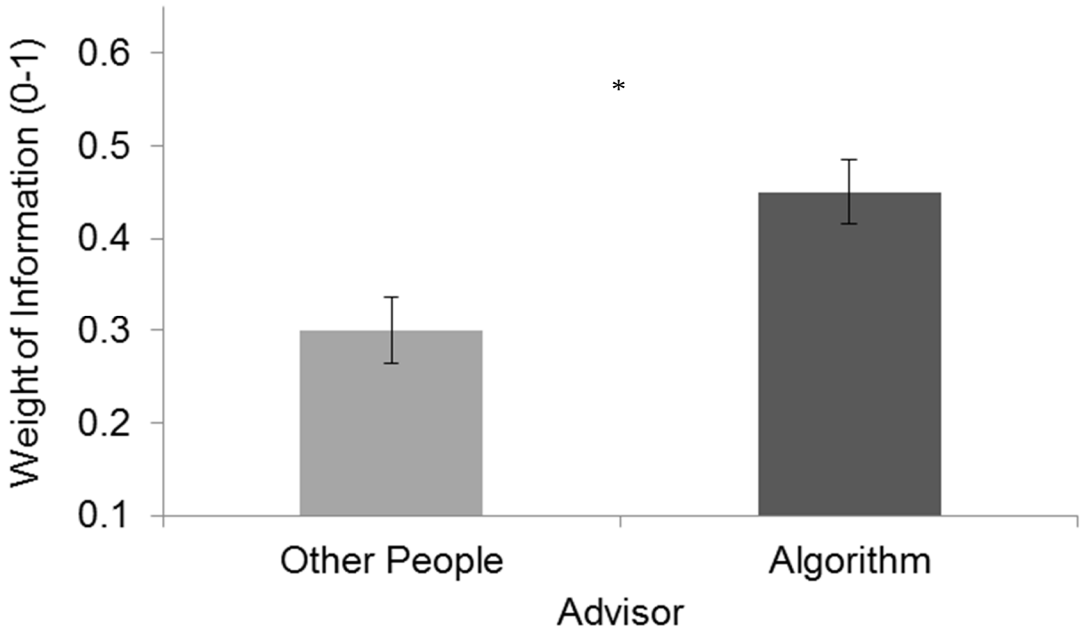

Figure 2. Weighting of Advice (WOA) as a function of experimental advisor (other people vs. algorithm), Experiment 1A. The higher the WOA, the more participants relied on the advice. Error bars indicate standard errors. Note: *$p < .05$.

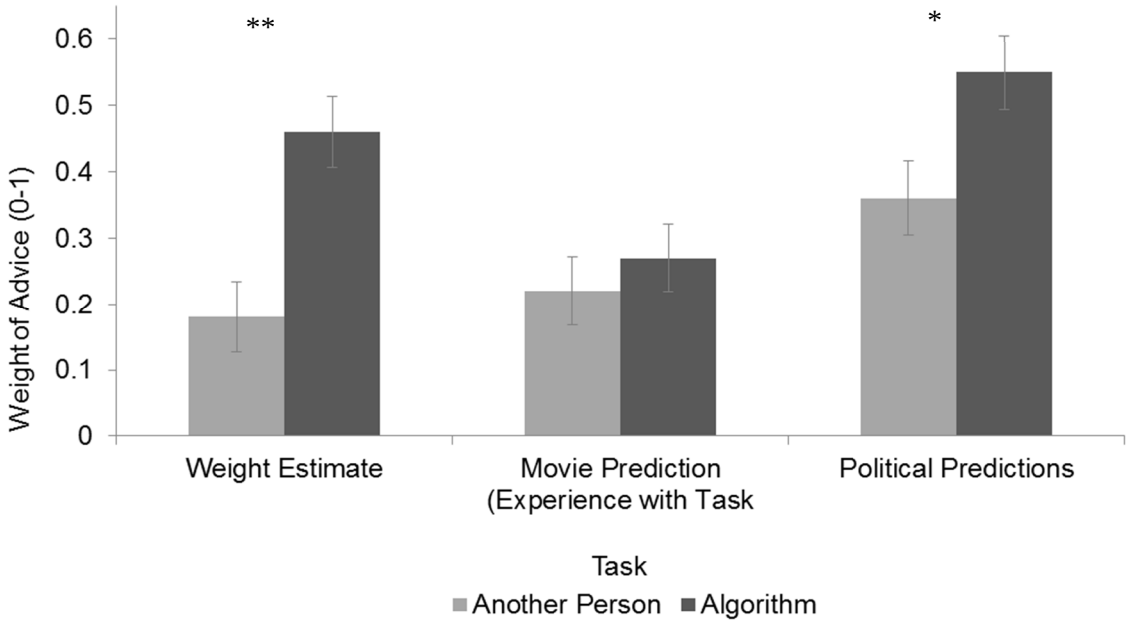

Figure 3. Weight of Advice (WOA) as a function of experimental advisor (another person vs. algorithm), Experiment 1B. Participants estimated someone's weight, forecasted the opening weekend gross for a movie, and forecasted two political world events. The higher the WOA, the greater the change in participants' Time 1 and Time 2 estimates. Note: $^{**}p < .01$, $^{*}p < .05$.

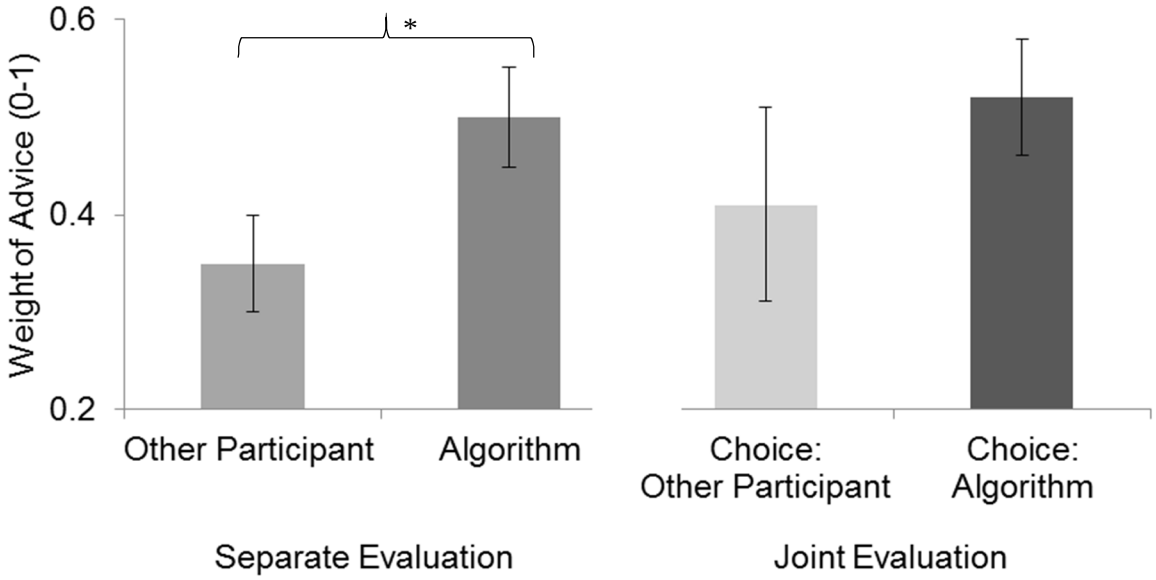

Figure 4. Weight of Advice (WOA) as a function of experimental advisor (person vs. algorithm) in the between- and within-subject designs, Experiment 2. Note: $*p < .05$.

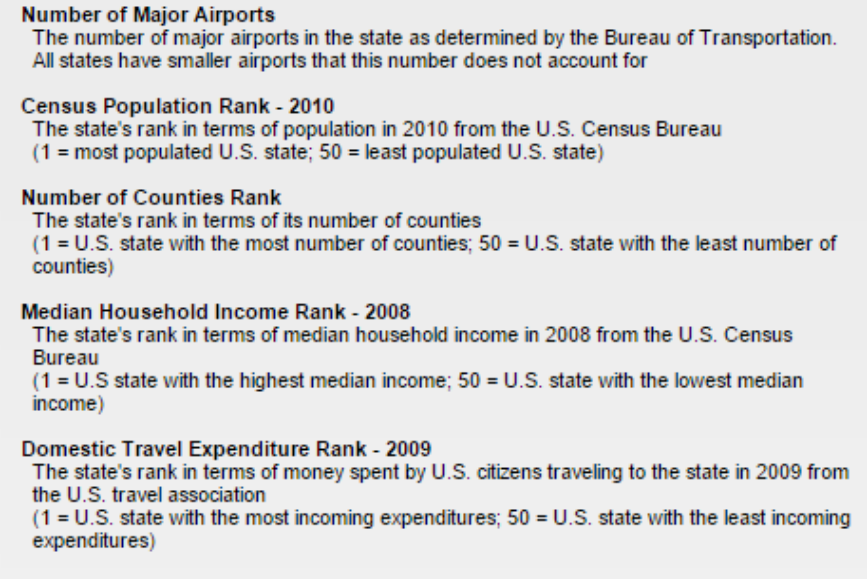

Figure 5. Screenshot of the information participants read they would receive to make their estimate, Experiment 3.

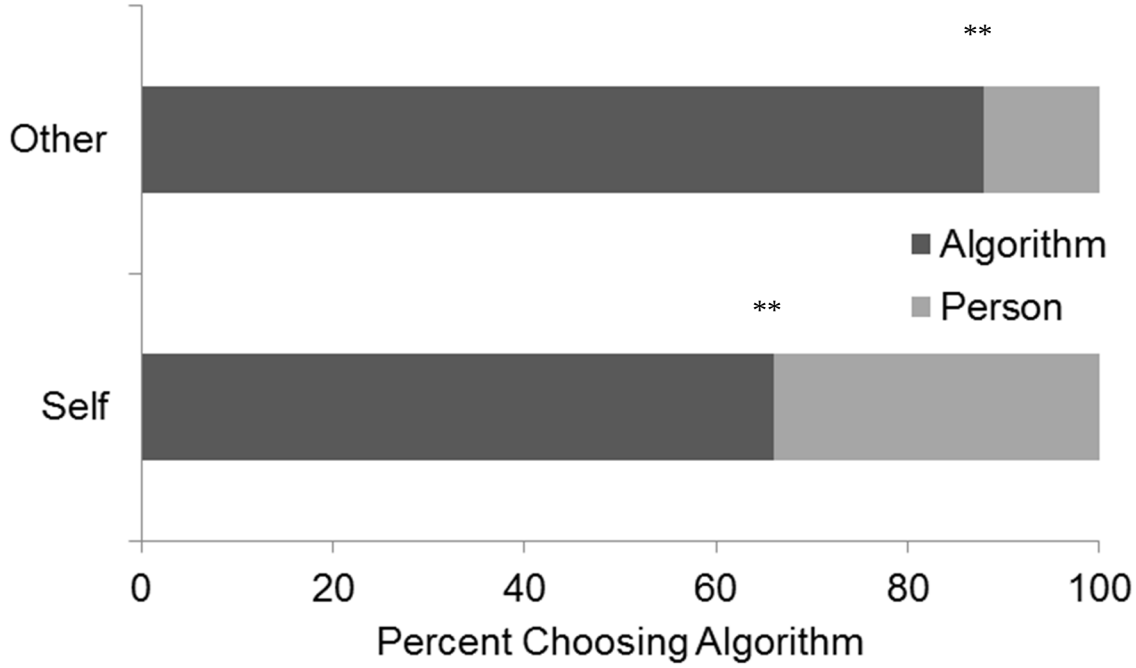

Figure 6. Percent of participants choosing the algorithm as a function of experimental advisor (person vs. algorithm) and self/other (self vs. other), Experiment 3. Note: $^{**}p < .001$. The graph is horizontal because the percentage for the other condition and the self condition each sum to 100%.

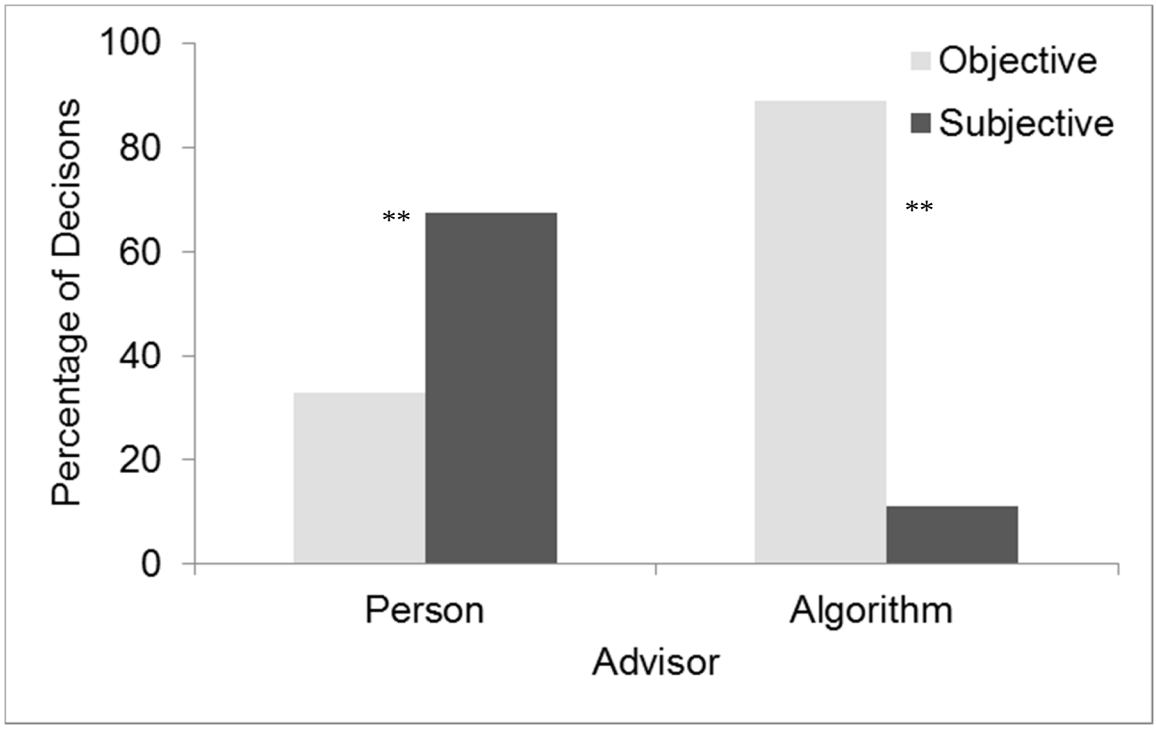

Figure 7. Percentage of decisions coded as objective or subjective as a function of experimental advisor (person vs. algorithm), Experiment 4. Note: $^{**}p < .001$.

Figure 8. The photograph used in Experiment 5.

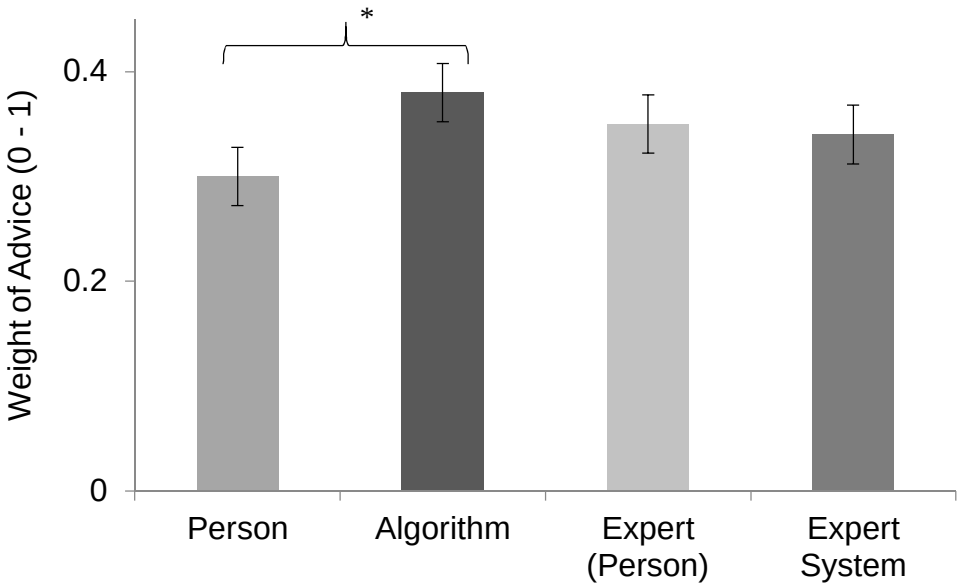

Figure 9. Weight of Advice (WOA) as a function of experimental advisor (person vs. algorithm vs. expert vs. expert system), Experiment 5. The higher the WOA, the more participants relied on the advice. Note: $^{*}p < .05$.

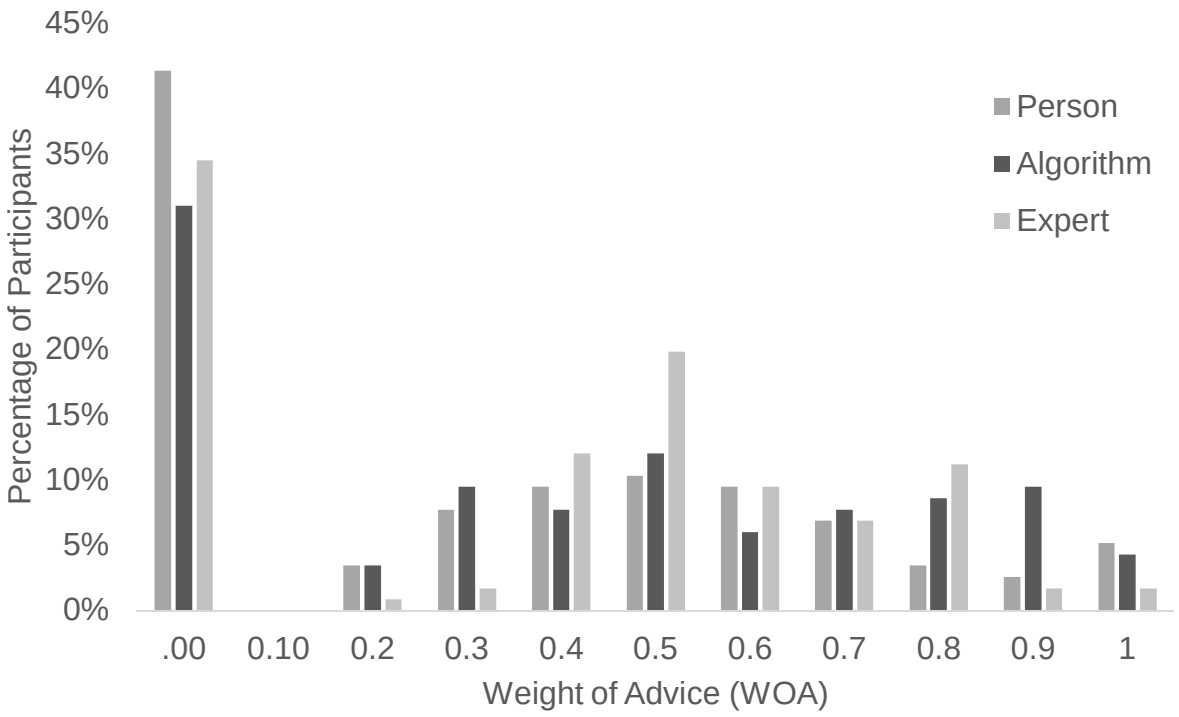

Figure 10. Percentage of participants whose weight of advice (WOA) falls within each range by experimental advisor (person vs. algorithm vs. expert), Experiment 6. Note: For ease of viewing, I removed the expert system condition.

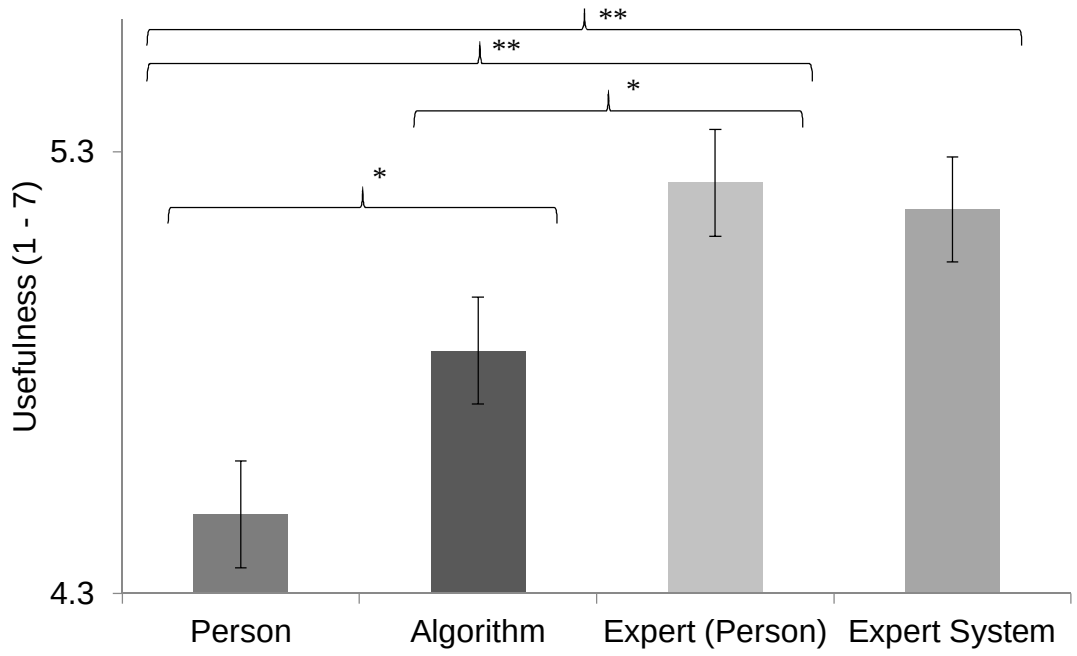

Figure 11. Expected usefulness of advice as a function of experimental advisor (person vs. algorithm vs. expert vs. expert system), prior to receiving advice, Experiment 5. Note: $^{**}p < .001$, $^{*}p < .05$.

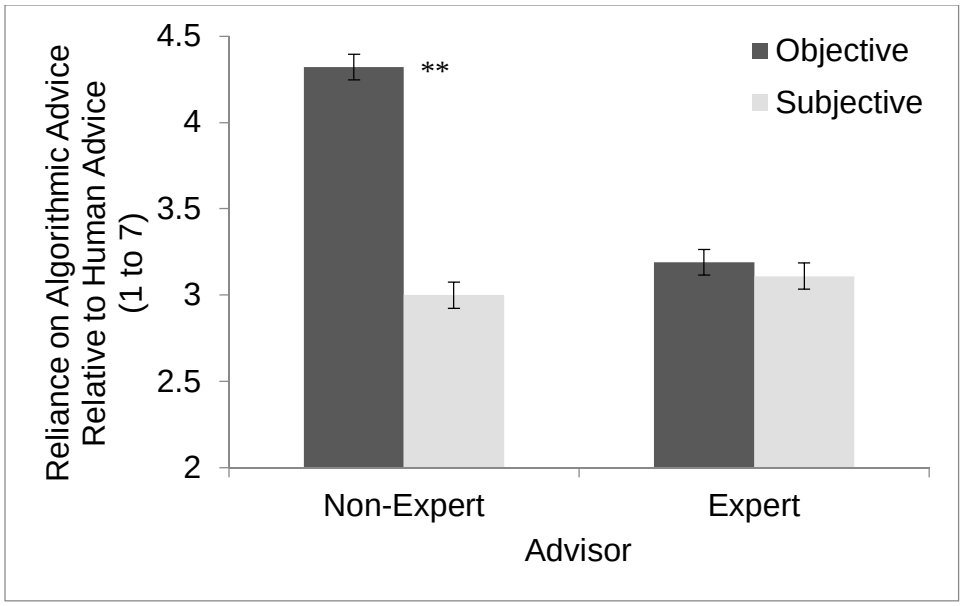

Figure 12. Reliance on algorithmic (versus human) advice as a function of experimental subjectivity (objective vs. subjective) and expertise of the human advisor (non-expert vs. expert), Experiment 6. Note: $^{**}p < .001$.

Figure 13. Photograph used in Experiment 7.

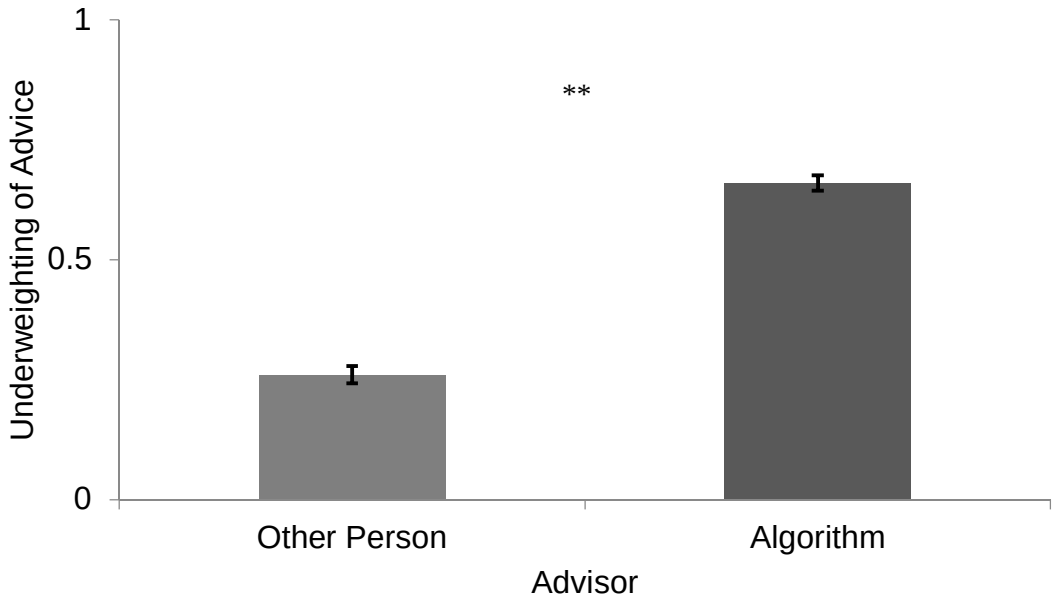

Figure 14. Magnitude of underweighting advice as a function of experimental advisor (person vs. algorithm), Experiment 7. The greater the underweighting, the more participants should have updated to the advice. Note: $^{**}p < .001$.

SUPPLEMENTARY MATERIALS

Pre-registrations, materials, and data for all experiments in this manuscript can be found in this direct link to the Open Science Frame Work: https://osf.io/b4mk5/. Also available are pre-registrations, materials, data, and results for dissertation experiments not included in this manuscript. All pre-registrations are available and all other noted documents will be uploaded shortly.