Essays on the Human Propensity to Rely on Algorithms as Tasks Become More Difficult

By Eric Bogert

(Under the Direction of Aaron Schecter and Richard T. Watson)

ABSTRACT

What makes people trust algorithms? We know that demonstrated accuracy, high interpretability, and prior familiarity with AI, among other factors, increase the likelihood that subjects comply with an algorithmic recommendation. However, most of the prior research investigates compliance with an algorithmic recommendation relative to one's belief, which is usually confounded by human overconfidence. We mitigate this confound by exposing subjects to identical advice labeled as either algorithmic or from a human crowd. Thus, we isolate the effect of algorithmic recommendations relative to the recommendations of a crowd without being confounded by natural human overconfidence. This three-experiment dissertation submits three research projects that investigate how people choose to respond to an algorithmic recommendation, moderated by the type and difficulty of task. The tasks are taken from three quadrants of McGrath's Circumplex Model of Group Tasks, to achieve task type diversity. Paper One investigates how humans weigh the estimates of a crowd compared with estimates of an algorithm for an objective, intellective task. Paper Two investigates how humans respond to recommendations from a crowd and an algorithm in the context of a creative task. Paper Three investigates how humans respond to recommendations from an algorithm when resolving conflicting interests.

INDEX WORDS: algorithms, wisdom of crowds, social influence, McGrath's Circumplex

Model, human-computer interaction

Essays on the Human Propensity to Rely on Algorithms as Tasks Become More Difficult

By

ERIC BOGERT

BBA, University of Georgia, 2013

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2021

Essays on the Human Propensity to Rely on Algorithms as Tasks Become More Difficult

By

ERIC BOGERT

Major Professors:	Aaron Schecter
	Richard T. Watson
Committee:	Amrit Tiwana
	Xia Zhao

Electronic Version Approved:
Ron Walcott
Dean of the Graduate School
The University of Georgia
August 2021

DEDICATION

This is dedicated to my parents, who have supported me in countless ways.

ACKNOWLEDGEMENTS

My committee chairs, Aaron and Rick, were more attentive and thoughtful than I deserved. Xia and Amrit provided valuable guidance to the framing, design, and future research stemming from this dissertation.

TABLE OF CONTENTS

Chapter 1

1. Problem Statement and Motivation

When humans make decisions, their minds apply logic, statistics, or heuristics (Gigerenzer and Gaissmaier 2011). Investigating when humans use each of these methods has been the subject of multiple Nobel prize-winning efforts. Herbert Simon, in his Nobel acceptance speech, framed his research around how humans reason, given they are boundedly rational (Simon 1979). Kahneman and Tversky treated heuristics as a crutch that removes humans from pure rationality, and Kahneman earned his Nobel prize for identifying biases that guide human behavior (Kahneman 2011; Kahneman and Tversky 1979). The research in this paper examines how human and non-human sources of information influence decision making.

When humans make decisions using any type of aid, such as an algorithm, they rarely act optimally. They might lean on its recommendations too heavily, *overutilization,* or not heavily enough, *underutilization.* Both overutilization and underutilization can result in suboptimal outcomes. For example, when humans overutilize an aid, they become less vigilant towards information gathering and processing, and demonstrate an *automation bias* that clouds their judgment (Mosier and Skitka 1996). Automation bias is not limited to input from machines – it is also a factor with non-automated input, such as that from other humans (Dzindolet et al. 2002).

There are two important sources of automation bias (Parasuraman and Dietrich 2010). First, humans are cognitive misers (Kahneman and Tversky 1972, 1974) – they tend to do that which is least taxing on their limited rationality. Thus, when humans receive advice, they typically expend less cognitive effort, as measured by physiological indicators such as heart rate, than they do

when they receive no advice (Alexander et al. 2018). Second, humans are predisposed to *social loafing*, employing less effort in teams than they would working alone (Karau and Williams 1993). Automation bias causes humans to over-prioritize external data. Of course, the goal of system designers should be to persuade individuals to trust the advice that is most likely to lead them in the correct direction, rather than to always trust themselves, other people, or AI.

It is likely that automation bias is affected by the type of task an individual is assigned. Humans may be naturally predisposed towards thinking some tasks are more worthy of their time or inclined to their unique skillset than others, which may affect the degree to which they lean on external advice. Depending on the task, an external source of advice might be tremendously helpful or horribly detrimental. For example, Steve Kerr might be an excellent basketball coach, and most humans would be happy to take his advice on basketball strategy. But those same people would likely be reluctant to take his advice on writing an MIS dissertation.

1.1.1 McGrath's Circumplex Model

To investigate how humans evaluate advice based on the juxtaposition of the task and source, we use McGrath's circumplex model of group tasks (McGrath 1984) (Figure 1.1). The circumplex model has been used as a foundation for the study of group behavior, particularly in the context of communication technologies and groups (Connolly et al. 1990; DeSanctis and Gallupe 1987; Straus 1999). Prior research has advocated using the circumplex model in the study of how task type affects human propensity to act on, be confident in, and expend effort on advice (Bonaccio and Dalal 2006).

The circumplex model of group tasks has four quadrants. Tasks that are closer together are more similar than those further apart.

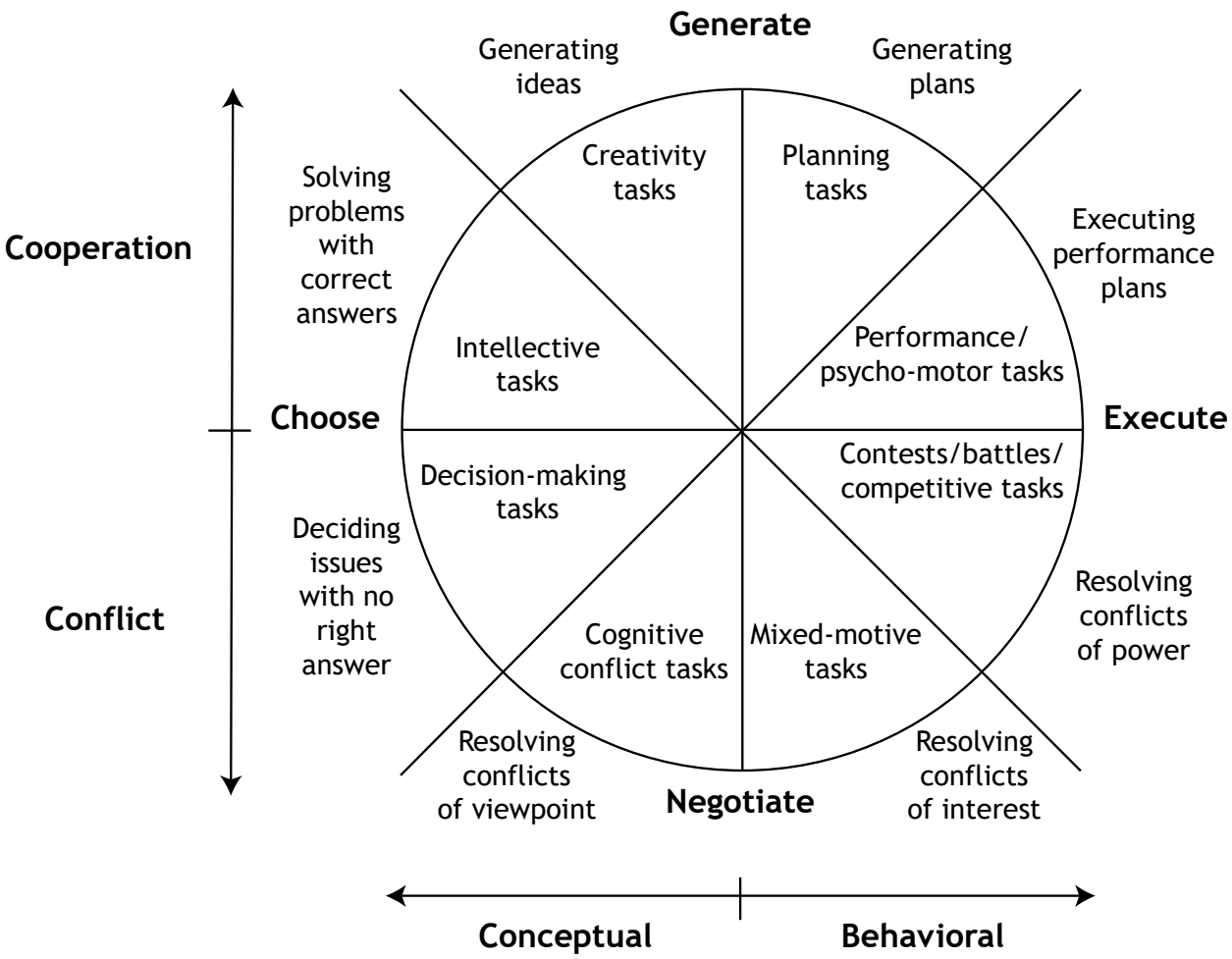


Figure 1.1: Circumplex Model of Group Tasks (McGrath 1984, p. 61)

The Choose quadrant contains both intellective and decision-making tasks. The original paper defining intellective tasks posited tasks on a continuum, with judgment tasks anchoring the opposite side of the continuum (Laughlin 2011). Most geometry, physics, chemistry, and other hard science problems are intellective problems (Laughlin 2011). This is in contrast to judgment tasks, such as which website design is more appealing (Laughlin 2011). Originally, intellective tasks were broken into three further types of tasks: first, tasks that create a "eureka" moment when the answer is explained; second, tasks with a correct answer that is difficult to demonstrate in an intuitively compelling way; third, tasks for which the consensus of experts *is* the correct answer. Decision-making tasks, on the other hand, are "based on peer consensus about what is

morally right or what is to be preferred" (McGrath 1984, p. 64). These tasks can involve social comparison between group members and answers can vary by culture. Decision-making tasks share some characteristics with intellective tasks, and with Negotiate tasks, which McGrath views as an extension of Choosing tasks (McGrath 1984, p. 64). In this paper, we use an intellective task in the first experiment.

The Generate quadrant contains planning tasks and creativity tasks. Tasks on the upper half of the circumplex model generally involve cooperation. The key goal for planning tasks is to create an "action-oriented plan" (McGrath 1984), and the key notions for creative tasks are brainstorming and creativity (McGrath 1984). In this dissertation, the task type in the Generate quadrant is a creative task, expanded upon in the second experiment.

In Negotiating tasks, the key is to *resolve*, not solve. Tasks on the lower portion of the circumplex model typically generate conflict within groups. Negotiating tasks pit two sets of conflicting interests against one another, such as labor and management. Our third experiment explores this type of task, in the context of a decision on whether to grant a suspect bail. Decisions on bail pit the interests of the public, who want to live in a safe society, against the interests of suspects who should be assumed innocent until proven guilty, and who face steep costs if they are not granted bail (Dobbie et al. 2018).

Executing tasks deal with physical behavior. Within the Execute quadrant there are two types: Contests and Performances. Contests pit two teams against one another, and crown one a clear winner and loser, such as in sports. Performances pit a team against what McGrath calls "nature" – examples include digging a hole or building a structure. These general purpose physical activities are the essence of execute tasks, which are defined as "those requiring physical movement, coordination, or dexterity, such as psychomotor tasks and athletic contests" (Straus

1999). This dissertation does not include a task from the Execute quadrant, because such tasks do not result in recommendations. Prior MIS literature has excluded the Execute quadrant for similar reasons (DeSanctis and Gallupe 1987).

Prior work on the effect of task types on majority influence has shown that the type of task has a direct effect on majority influence, and that the effect is stronger for preference tasks than intellective tasks (Tan et al. 1998). These effects are also not moderated by whether the groups are communicating through computers (Tan et al. 1998) – which is important because this dissertation focuses on people doing tasks while using a computer.

1.1.2 Overconfidence in Estimations

A thorough review of human preferences in decision making must begin with a simple observation: Most humans are overconfident (Harvey 1997; Moore and Healy 2008) for three reasons (Gino and Moore 2007). First, people are egocentric (Soll and Mannes 2011). They believe they are smarter and more skilled than average (Kruger and Dunning 1999). This is because people like to construe vaguely-defined, positive traits, such as intelligence, in a way that indicates they have high levels of that trait (Logg et al. 2018). This overconfidence is driven by confirmation bias when people observe that their predictions are correct (Nickerson 1998). Overconfidence prevents people from inculcating the predictions of other people as fully as they should. It is also a primary reason why people underutilize the advice of their peers. For example, in tasks of prediction in a dyad of peers, if each believed that the other person was equally capable, then they should simply average both predictions (Dawes and Corrigan 1974). Instead, humans tend to overweight their predictions and underweight those of their peers (Yaniv and Kleinberger 2000).

Second, humans understand their own calculus better than their advisor's. This leads them to trust their decision more (Yaniv and Kleinberger 2000). For example, people believe that their own predictions are more "objective" than those of their peers (Liberman et al. 2012).

Third, in studies in which advice is revealed only after an estimate is given, there is an anchoring effect on the original estimate (Kahneman and Tversky 1974). People take new advice only in the context of their original estimate, and thus discount it more than if they had received the advice prior to creating an estimate.

However, there are ways to elicit a greater dependence on other people. When humans pay for advice, they use it more (Gino 2008) – this is related to the sunk cost fallacy. When the experience of the advisor is emphasized, the advice is used more (Harvey and Fischer 1997). Believing advice comes from more individuals, rather than from fewer individuals, also makes people use it more (Mannes 2009; Minson and Mueller 2012).

Overconfidence is a confounding variable that plagues prior research. Much of the most significant research (Abeliuk et al. 2020; Dietvorst et al. 2015; Kawaguchi 2020; Logg et al. 2019) evaluating algorithmic advice compare a subject's weight on advice after being exposed to an algorithm, without a comparison of how subjects respond to advice from other people. This research setting makes sense in the context of the business world in the past decades, in which managers could access a recommendation system but not the opinion of a large crowd. However, the Internet makes it easy to access the opinions of crowds. Online reviews of movies, books, cars, and nearly every other consumer item are readily available (Dellarocas 2003). Most humans are exposed to a bevy of information – sometimes misinformation – from crowds on social media websites (Lazer et al. 2018). The human response to algorithmic versus crowd recommendations has gone from being an esoteric topic without a use case to an issue that has

dramatic implications for society. This has become relevant to social media platforms trying to optimize disinformation flags, for e-commerce platforms determining how to frame recommendations to potential consumers, and to a plethora of other Internet companies who can frame recommendations.

1.1.3 Preferences Toward Human Advisors

It has long been posited (Meehl 1954) that people prefer human to algorithmic advice. In a recent survey of academics who specialize in psychology and decision-making, the majority expected human subjects to respond more strongly to advice from humans than advice from algorithms (Logg et al. 2019). This preference toward human advice is generally labeled *algorithmic aversion* (Dietvorst et al. 2015), *which* is particularly strong after observing an algorithmic make a mistake (Dietvorst et al. 2015). However, what matters most after observing a mistaken algorithmic recommendation is whether a subject believes that the algorithm is *relatively better than their abilities* (Moray and Lee 1994). Algorithmic aversion is observed in preference tasks, such as joke recommendations (Yeomans et al. 2019); in subjective tasks, such as dating (Castelo et al. 2019); and in lab experiments comparing human doctors and algorithms giving advice (Promberger and Baron 2006).

A close reading of the seminal algorithmic aversion article (Dietvorst et al. 2015) reveals that before they observe an algorithm's mistake, subjects prefer the algorithm's advice (Logg et al. 2019). When people demonstrate a preference for recommendations from algorithms, they exhibit *algorithmic appreciation* (Logg et al. 2019).

Algorithmic aversion is not rational. A simple algorithm, such as weighting all variables equally, can outperform human prediction (Dawes 1979). More sophisticated strategies can perform even better. In a landmark meta-analysis of 136 studies, algorithms were 10% more accurate, on

average, than non-algorithmic judgment, and non-algorithmic judgment rarely outperformed algorithms (Grove et al. 2000). However, despite the evidence against human assessment, humans consistently prefer their judgment (Diab et al. 2011; Eastwood et al. 2011). This is not simply due to interpretability. When the process an algorithm uses is clear and easy to follow (clear box), people do not follow a machine learning algorithm's predictions more than when it is opaque and impossible to know the process (black box), and they are less likely to identify significant mistakes in predictions of clear-box algorithms (Poursabzi-Sangdeh et al. 2018). However, this is a rapidly changing field. New techniques such as Local Interpretable Model-Agnostic Explanation (LIME) are allowing even black-box algorithms to become more interpretable (Rai 2020).

Algorithms that are perceived as more similar to humans are preferred over algorithms that are perceived as less human (Castelo et al. 2019). Evolutionary psychology indicates that an innate disposition to trust has been key to homo sapiens' success as a species (Taylor 2014). Humans are more likely to trust other humans who look like them, ostensibly because they are more likely to be family or tribe members (DeBruine 2002). As AI becomes more human-like, it is possible that the trust calculus changes increasingly in favor of algorithms as AIs eventually become both human-like, more familiar, and seemingly super intelligent.

Unfortunately, humans are often exposed to algorithms that are flawed by design, likely tainting their overall trust toward algorithmic suggestions and forecasts. For example, weather forecasts from The Weather Channel are intentionally biased towards rain when the true probability is less than 30%, and The Weather Channel intentionally avoids forecasting a 50% chance of rain (Bickel and Kim 2008). When The Weather Channel predicts a 20% chance of rain, rain historically occurs about 5% of the time (ibid). This bias exists because people notice, and are

more annoyed by, unexpected rain than by unexpected sunshine (Silver 2012, p. 135). The Weather Channel is thus incentivized to lie about the true underlying probabilities to avoid irate consumers. In fact, one of the best ways to ensure one does not get caught by unexpected rain is to crowdsource the question – by looking out the window and seeing whether passerby are carrying umbrellas (Surowiecki 2004). Large samples of human guesses often yield estimates that are highly accurate.

1.1.4 Social Influence

When a decision is affected by other people, the force acting on the decision-maker is *social influence.* Social influence can be particularly strong when the subject is exposed to large numbers of people. An early test of the limits of social influence was to explore whether it could convince people to agree with the group assessment even when the group was obviously incorrect. In the 1950s, this was tested in very small human groups, some groups small enough that they barely fit the definition of a crowd, in a series of experiments by Solomon Asch. As the size of a crowd grows, the propensity to follow the crowd grows (Asch 1951; Milgram et al. 1969). In the seminal study on conformity in groups, Asch found that as the size of groups increase from 2-4 people, the propensity to agree with the group increases, even when the group consensus is clearly wrong (Asch 1951). However, this effect plateaus with larger majorities – majorities of 16 do not have more power to dissuade people than even majorities of 3. Later tests of social influence reveal that humans do not shift their beliefs as much as they should when they are exposed to the guess of a crowd (Mannes 2009; Soll and Mannes 2011).

When the number of people increases, given the right conditions, eventually the group develops a property where the average guess of the group becomes very accurate. This happens because a crowd combines its unique knowledge, although there is often an error in each person's estimate.

Under certain conditions, the errors of these estimates will cancel, resulting in highly accurate guesses. The ability of large human groups to accurately estimate an unknown quantity is *the wisdom of the crowd.* Crowds are wise when three conditions are true: 1) there is a diversity of opinion, 2) opinions are independent, 3) and opinions are decentralized (Surowiecki 2004). These three criteria were fulfilled in the seminal paper on the wisdom of the crowd, when it was discovered that the average guess of the weight of a dressed, butchered cow at a county fair was within 1% of the cow's true weight (Galton 1907).

Historically, it has been assumed that guesses need to be independent of one another to generate accurate crowd-based guesses. There are two commonly cited reasons for this. First, independence prevents guesses from becoming correlated with one another. Second, independence increases the likelihood that new information is utilized (Surowiecki 2004, p. 41). In practice, it is often difficult to get independent assessments. For instance, even though President Kennedy consulted with multiple people when considering the Bay of Pigs invasion, there was little cognitive diversity amongst them (Surowiecki 2004). This lack of diversity and drive to create consensus can create *groupthink* (Janis 1972; Surowiecki 2004, p. 37).

However, recent evidence has emerged that social influence can generate learning that increases the wisdom of the crowd over multiple iterations of guesses (Becker et al. 2017). Social influence can even improve overall network-level predictions of tasks containing highly motivated, politically contentious reasoning, even when people know the ideologically opposed motivations of their peers in the network (Guilbeault et al. 2018). However, people might be reluctant to listen to the wisdom of the crowd, if they believe they have above-average intelligence or expertise. They would be wise to abandon their initial guess and listen to the wisdom of the crowd, because crowds often outperform their wisest member, if given enough

time (Bauer et al. 2003). This is observed, for example, in financial markets. Although there are outliers such as Renaissance Technologies that have beaten the returns of the S&P 500 consistently (Zuckerman 2019), the vast majority of funds do worse than the S&P 500 over the long term (Fama 1970), as predicted by the efficient markets hypothesis (EMH). In other economic cases, such as in forecasts of macroeconomic variables such as real GDP or the CPI, consensus estimates outperform even the very best individual estimates over time (Bauer et al. 2003). In sports betting markets, the aggregate guess of prediction markets outperforms over 99% of all individual experts (Servan-Schreiber et al. 2004).

The wisdom of crowds is hugely important to the functioning of the United States' most important institutions. That the stock market, perhaps the most essential ingredient to sustaining American capitalism, cannot be consistently outperformed by managed funds is fundamentally a massive, real-time instantiation of the wisdom of crowds. One understudied facet of the wisdom of the crowds is how people weigh inputs from other humans compared to inputs from non-human sources, such as AI.

Humans are also likely to continue listening to what they perceive to be expert human judgment even when it underperforms aggregate human judgment. This is most strongly displayed by the propensity of humans to invest their money with expensive fund managers who, on average, dramatically underperform benchmark passive funds after fees (Malkiel 1973). It is likely that this is due to a mistaken belief that a fund manager will improve in the subsequent years. This is evidence that humans are likely to forgive inaccurate human guesses in certain contexts. However, it is also possible that the mystique of an AI makes humans believe that it fundamentally knows something that they do not, particularly when the AI makes a significantly different prediction than its human counterparts. This could be because it has been trained on

large datasets, or because it can do more calculations per second than humans, or some other factor.

Surprisingly, even small groups can outperform not only the average member of the group, but also the smartest member (Blinder and Morgan 2005). The Asch experiments, for example, were designed so that the non-confederate member of the group was asked their opinion only after each confederate had revealed their erroneous guess. This reveals how early decisions can affect later decisions. In their most extreme form, these early decisions can create an *information cascade*, which occur when "it is optimal for an individual, having observed the actions of those ahead of him, to follow the behavior of the preceding individual without regard to his own information" (Bikhchandani et al. 1992). Early decisions are reinforced by later decision makers, who ignore their private information (Bikhchandani et al. 1992; Surowiecki 2004). In a situation where people are close to the borderline between two decisions, such as what restaurant to go to or what clothes to buy, a small informational shock can cause significant downstream effects (Bikhchandani et al. 1992).

1.1.5 Task Difficulty

One reason why we might observe under-reliance on advice (Soll and Mannes 2011; Yaniv and Kleinberger 2000) in some tasks and over-reliance on advice on other tasks (Malkiel 1973) is that lab experiments demonstrating under-reliance might be easier than, for example, predicting the stock market (Gino and Moore 2007). Thus, manipulating task difficulty is important to understanding of how humans use algorithmic advice.

Prior IS research indicates that humans use technology differently depending on the task difficulty. In group decision support systems (GDSS), for example, teams facing harder tasks coalesce on better decisions when they are supported by technology, even though the technology

reduces confidence in their answer (Gallupe et al. 1988). Task difficulty is a known predictor of system use. Surprisingly, the easier the task, the more IS success (Petter et al. 2013). This is likely because of higher satisfaction with the supporting MIS (Gelderman 2002).

Task difficulty has important interaction effects with the human psyche. Overconfidence, for example, remains high despite task difficulty. Overconfidence does not vary with task difficulty (Klayman et al. 1999), however people are more overconfident for difficult tasks (meaning they create overly narrow confidence intervals) while incorrectly believing they are worse than others at those tasks (Moore and Healy 2008). Crucially, people also tend to use advice more for difficult tasks and less on easy tasks (Gino and Moore 2007).

1.1.6 Other Relevant MIS Literature

1.1.6.1 Recommendation Agents

The literature on recommendation agents (RAs) is relevant to any discussion of how humans respond to input from algorithms. The vast majority of research on recommendation agents is set in an e-commerce context (e.g. Komiak and Benbasat 2008; Qiu and Benbasat 2009; Xiao and Benbasat 2007; Xu et al. 2014). This setting is fundamentally a decision-making task – there is no objectively correct answer. In an e-commerce context, each decision to purchase (or not to purchase) a product has an associated utility payoff, which can be heterogeneous among individuals. The recommendation agent literature is thus distinct from the decision-making literature. We now briefly review the important findings of the recommendation agent stream of research.

Many recommendation agent papers focus on trust in the RA, which is a process, and trust and distrust in RAs are different constructs (Komiak and Benbasat 2008; Wang and Benbasat 2008). Unlike the algorithms we use, which provide the same recommendation to all individuals, RAs

are designed to elicit the preferences of consumers (Xiao and Benbasat 2007) and make recommendations that would maximize the utility of the recommendation agent owner (usually an e-commerce corporation).

RAs that use voice-based communication and other forms of human embodiment are considered to have greater social presence, which makes people trust, enjoy, and intend to use them more (Qiu and Benbasat 2009). Consumers prefer RAs that explain their decision processes – explanations make consumers view an RA as more competent and benevolent (Wang and Benbasat 2007). We can build on the edifice created by the recommendation agent literature in an intelligent way. For example, we know that trust is a process that changes based on use with a recommendation agent. We statistically control for this in our experimental designs, by randomizing the order of survey questions, so that later questions with recommendations from algorithms are not systematically more trusted. Similarly, we know that social presence predicts intention to use an RA – we keep the level of social presence static throughout experiments and within experiments, so that we can more concretely identify the effect of our other manipulations. Lastly, we keep the explanation underlying our algorithm identical between and within experiments.

Because recommendation agents help consumers in decision-making tasks, we expect there to be different effects in how people respond to algorithms. However, the Group Decision Support Systems Literature (GDSS) is a branch of research that also studies how humans respond to recommendations.

1.1.6.2 Group Decision Support Systems

A Group Decision Support System (GDSS) "combines communication, computing, and decision support technologies to facilitate formulation and solution of unstructured problems by a group

of people" (DeSanctis and Gallupe 1987). GDSS research looks at two dependent variables that are similar to what we evaluate in our study: time to arrive at a solution, and solution quality (Benbasat and Lim 1993). A third dependent variable GDSS research often includes is satisfaction, which is similar to confidence, the variable we measure.

The GDSS literature also uses task types from McGrath's circumplex model. Although we do not consider our research to be directly contributing to the GDSS literature, for reasons we will later discuss, it can still be informed in part by this paradigm. In the context of the DeSanctis and Gallupe's model (Figure 1.2) we keep group sizes identical within an experiment, and we keep participants dispersed. Thus, we precisely focus on what the effect of moving between tasks is.

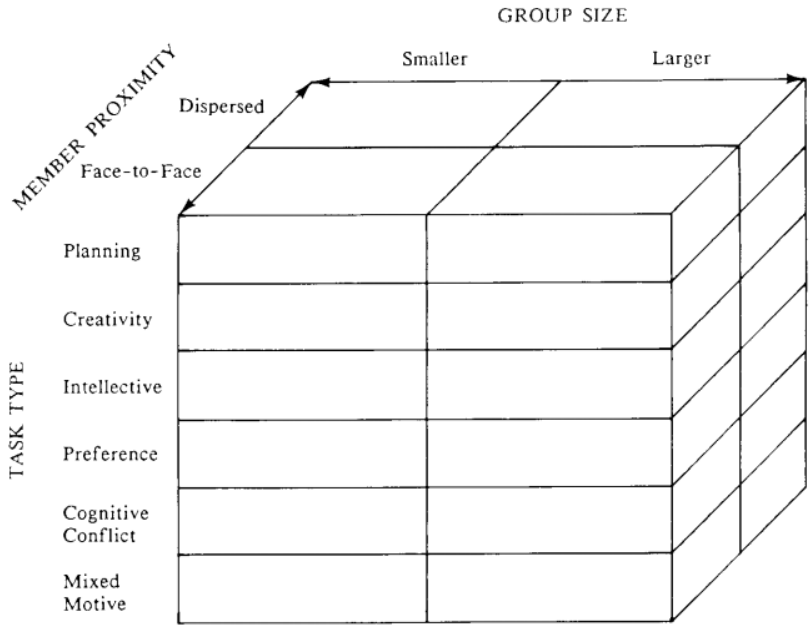


Figure 1.2: Desanctis and Gallupe's (1987) GDSS Summary

Task and technology have to fit in order for a GDSS to be effective (Zigurs and Buckland 1998). A GDSS increases the time to reach a decision (DeSanctis and Gallupe 1987), GDSSs result in more information being shared (Dennis 1996), and GDSS support more participation from lower-ranking group members (Dennis and Garfield 2003; George et al. 1990).

There are several reasons why we believe our work is distinct from the GDSS research. First, a GDSS is about group-level outcomes. Our research is interested in how *individuals* respond to different types of recommendations. Thus, the unit of analysis is different. Second, the goals of GDSS, such as directing discussion, removing communication barriers, and structuring decision analysis (DeSanctis and Gallupe 1987), are not our goals. Third, GDSS research is about designing technologies that improve decision outcomes. This can mean varying the anonymity of the group (George et al. 1990), the size of the group (DeSanctis and Gallupe 1987), whether there is a leader in the group (George et al. 1990) or other factors. Our research keeps the technology, group size, leadership presence, anonymity, and other factors static, and evaluates how people decide to incorporate recommendations based on differences in the information source. We believe there is no GDSS research that changes the information source between algorithmic and human sources.

1.1.7 Definitions

The following table of definitions define the key terms used in this paper. Note that we use the words *advice* and *recommendation* interchangeably.

Table 1.1: Definitions of Key Terms

Term	Definition	Paper	Authors
Algorithm	A mathematical, step-by-step procedure or formula for computation	*A systematic review of algorithm aversion in augmented decision making*	Burton et al. (2020)
Algorithmic Aversion	The reluctance of human decision makers to use superior but imperfect algorithms	*A systematic review of algorithm aversion in augmented decision making*	Burton et al. (2020)

Anchoring Bias	Different starting points yield different estimates, which are biased toward the initial values.	*Judgment Under Uncertainty: Heuristics and Biases*	Tversky and Kahneman (1974)
Automation Bias	Errors made when decision makers rely on automated cues as a heuristic replacement for vigilant information seeking and processing	*Human Decision Makers and Automated Decision Aids: Made for Each Other?*	Mosier and Skitka (1996)
Availability Bias	Estimating frequency or probability by the ease with which instances or associations could be brought to mind.	*Availability: A heuristic for judging frequency and probability*	Tversky and Kahneman (1973)
Bounded Rationality	Organisms adapt well enough to 'satisfice'; they do not, in general, 'optimize'	*Rational Choice and the structure of the environment*	Simon (1956)
Cognitive Effort	the engaged proportion of limited-capacity central processing	*Cognitive Effort and Memory*	Tyler et al. (1979)
Creative tasks	Generating ideas (e.g., brainstorming)	*Groups, Interaction, and Performance*	McGrath (1984)
Intellective Tasks	Tasks for which there is a demonstrable right answer	*Groups, Interaction, and Performance*	McGrath (1984)
Mixed-motive tasks	Tasks that resolve conflicting interests	*Groups, Interaction, and Performance*	McGrath (1984)
Overconfidence	Believing you are better than you are in reality	*The Evolution of Overconfidence*	Johnson and Fowler (2011)
Wisdom of the crowd	The ability of human groups to accurately estimate an unknown quantity	*The Wisdom Of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*	Suroweicki (2004)

In order to maximize clarity and brevity, please assume that each research project inherits the discussion in sections 1.1.1 through section 1.1.5.

1.2 Research Designs:

We propose a conceptual model of decision making in which humans respond to decision tasks based on two factors: information source and task characteristics (Figure 1.3). We acknowledge there are likely other factors influencing a response, but a review of the prior research indicates information source and task characteristics are central factors. We posit that an information source has a direct effect on an individual's response to the task. Information sources are perceived differently. Humans think non-algorithmic advice can more effectively incorporate qualitative data and believe algorithms can be dehumanizing (Grove and Meehl 1996). Humans are also prone to recollect instances when algorithms were outperformed by human judgement, due to the availability bias (Dawes 1979; Kahneman and Tversky 1974) These effects are moderated by characteristics of the task. For example, in highly important tasks humans prefer non-algorithmic advice because it is considered more ethical (Dawes 1979). There is also a direct effect of task characteristics on an individual's response. In easier tasks humans will usually expend less cognitive effort, be more confident, and be less willing to incorporate advice.

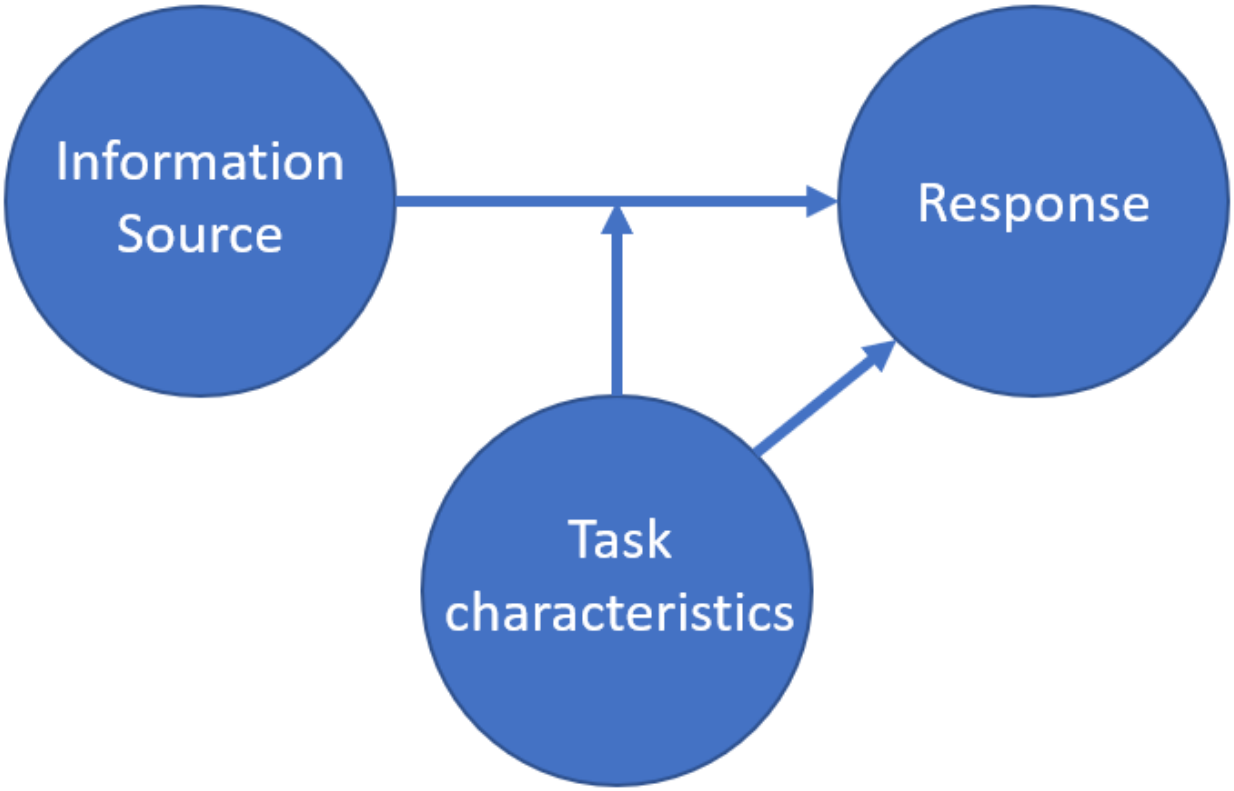


Figure 1.3: Research Model

We study the response to two types of information sources, the wisdom of crowds and algorithms. We also study two facets of task characteristics. First, we vary the task type by using three types of tasks drawn from McGrath's circumplex model. Second, we study difficult and easy tasks. This leads to the central proposition of the research.

Proposition 1: *An individual's response to decision advice is determined by characteristics of the task and the information source*.

We believe that the type of task will determine how people respond to new information. There are several reasons for this. First, it's speculated that humans are likely to be more accepting of algorithmic input if they are in a context in which algorithms are commonly used, such as weather forecasting (Castelo et al. 2019; Logg et al. 2019). This has its roots in the *mere exposure effect* of human psychology – exposure alone creates a preference towards something

(Zajonc 2001). Algorithms are used at a far greater rate in intellective tasks than creative or negotiating tasks, although we are not aware of research that quantifies this difference. Thus, if the public is more cognizant of algorithms in intellective tasks, they are more likely to be willing to use algorithmic advice for intellective tasks than in other tasks.

Second, humans who are familiar with the current landscape of AI know that algorithms are currently incapable of human-quality creative output. Thus, people may be less likely to use algorithms for creative tasks.

Third, many intellective tasks are numerical, more so than creative tasks such as writing a novel. It is easier to write or train algorithms for numerically driven tasks, and thus we expect that humans will show preference toward algorithms for intellective tasks.

Fourth, people might be less inclined to inculcate algorithmic advice in *important* tasks, because they are afraid of the consequences of the algorithm failing (Castelo et al. 2019). It is possible that intellective tasks are more or less important to people than creative or mixed-motive tasks.

Fifth, informing people of the superiority of algorithmic recommendations makes them more likely to use algorithms (Castelo et al. 2019). This indicates that if some people are already aware of algorithmic superiority on most intellective tasks, they will display a slight preference, on average, for algorithmic recommendations for intellective tasks.

Lastly, humans are prone to trust algorithmic recommendations more than human recommendations in situations that humans perceive as *objective* (Castelo et al. 2019). It is possible that people will perceive an intellective task as more objective than a creative or negotiating tasks.

Task difficulty is an important, understudied concept in the literature on decision preferences between algorithms and human decision assistance. Prior research on the effects of task difficulty does not compare algorithmic to human advice. Rather, the seminal papers on task difficulty simply look at the effect of learning the estimate of a peer for easy or difficult tasks (Gino and Moore 2007). We build on this in two ways. First, by evaluating the interaction of algorithmic and human advice across task difficulty. Second, we vary the degree to which effort predicts success across tasks – prior research uses tasks such as guessing weight, where effort has a weaker relationship to an accurate answer. In a task such as guessing weight, skill matters more than effort.

In tasks where effort is a strong predictor of accuracy, algorithms should strongly outperform humans. Humans are cognitive misers and prefer to shirk work (Simon 1956). Algorithms, in contrast, are tireless, and have no preference towards "working" or not. Furthermore, most humans likely have an innate understanding that humans prefer to be lazy, despite probably never hearing the phrase "cognitive miser". Thus, humans might have conscious or sub-conscious biases toward algorithms depending on the difficulty of the question, because they might *assume that their peers' laziness for hard questions* subsumes any benefits of another humans' cognition.

We now discuss the dependent concepts we measure, followed by the corollaries to our proposition. Rather than repeat the suggested effects of overconfidence, preferences toward human advice, wisdom of the crowds, and the effects of task difficulty, we very briefly review these effects and state how we propose they should affect each dependent concept.

1.2.1 Dependent Variables

We use three dependent variables: Cognitive Effort, Confidence, and Belief Change.

1.2.1.1 Cognitive Effort

Cognitive effort is an increasingly important topic in the MIS community. The objectives of all decision makers are to minimize effort and increase accuracy (Payne et al. 1993). The preponderance of research indicates that the effect of effort is stronger than the effect of accuracy (Johnson et al. 1988; Johnson and Payne 1985). Originally, cognitive effort was brought to the forefront in IS research by the Decision Support Systems (DSS) literature. The theory was that decision strategies would be improved if a DSS could make better strategies cognitively easier than worse strategies, then people would adopt those better strategies, thus improving the efficacy of the decision (Todd and Benbasat 1999).

In the DSS literature, perceived effort expenditure dictates the type of decision strategy people use, and they determine performance. We control for this by keeping effort expenditure the same between algorithmic and human crowd conditions. We differentiate from the DSS literature by taking a true measure of effort (discussed in the operationalization section), rather than a self-reported measure of effort. Furthermore, we build on the edifice of the DSS literature by treating effort as a dependent concept (Tiwana and Kim 2019).

There are two possible effects that algorithmic or crowd recommendations can have on cognitive effort. First, the more humans are disposed toward recommendations from an information source, the more they will rely on it without spending effort vetting its advice. Thus, if we observe algorithmic appreciation, we should see *decreased* cognitive effort when receiving recommendations from algorithms. Second, cognitive effort can *increase* if a person is unfamiliar with the recommendation source in that context. These two effects can happen simultaneously. Imagine a machine learning researcher who is naturally disposed to believe AI can outperform humans is writing a novel with an AI recommendation system. If that person sees

the AI recommends a beautiful, topical, crisp ending to the novel, they may be inclined to use it because they naturally trust AI or confused because they are aware AI is not capable of writing beautiful prose.

1.2.1.2 Confidence

Humans are naturally overconfident in their predictions. We control for this overconfidence because subjects are not choosing between their estimation and that of an AI, rather they are choosing how to incorporate information depending on the type of information source. In most intellective tasks, people are more confident when receiving advice from algorithms relative to advice from humans (Logg et al. 2019). Confidence is likely tied to our other concepts – if people expend significant cognitive effort, that likely affects their confidence. For example, confidence usually increases with effort (Paese and Sniezek 1998). Similarly, if someone is not confident in their answer, they may be more likely to respond strongly to new advice.

1.2.1.3 Belief Change

Lastly, we measure how much individuals shift their beliefs in response to advice. This is the most often studied dependent concept in the algorithmic appreciation literature (Castelo et al. 2019; Gino 2008; Gino and Moore 2007; Logg et al. 2019), and oftentimes it is the only dependent variable in an experiment. This is currently considered the best measure of whether people prefer the advice of algorithms relative to the advice of humans. Ultimately, this research seeks to understand how people rely differently on advice depending on the information source and task difficulty, and thus belief change is the key dependent variable.

1.2.2 Proposition Development: Intellective Tasks

There is substantial heterogeneity in algorithmic appreciation between tasks (Castelo et al. 2019). Perceived task objectivity predicts algorithmic appreciation – the more objective the more algorithmic appreciation. In subjective tasks such as recommending jokes, people prefer

receiving recommendations from humans, and report that human recommendations are easier to understand (Yeomans et al. 2019). However, people also demonstrate algorithmic appreciation in seemingly subjective tasks, including predicting how popular music will be, or predicting how romantically compatible two people will be (Logg et al. 2019).

Although perceived objectivity influences algorithmic appreciation, there are other features that make algorithmic appreciation stronger. In intellective tasks that have answers that are directly observable, but which require significant mental calculations (e.g., arithmetic on large sets of digits) it is likely that humans display significant algorithmic appreciation, because humans intuitively understand that algorithms do not tire, whilst humans do. Many humans also have experience using algorithms for numerical problems. Spreadsheets, calculators, and depictions in pop culture of algorithms being highly numerical all should drive humans to believe that algorithms are accurate for well-defined numerical tasks. Overconfidence magnifies this effect – humans are likely to believe their own estimates more than a crowd's estimate.

Thus, we propose the following supplements to Proposition 1:

Corollary 1: *For intellective tasks, the effect of algorithmic advice on belief change will be stronger than the advice of a crowd.*

Although humans are likely to place more weight on the advice of algorithms than the advice of a human crowd, this is likely to decrease as tasks get easier. An oversimplified example can intuitively explain this. Humans are likely to think an algorithm simply made a mistake if an algorithm states that there are 5 people in an image that very clearly shows three people and two dogs. However, if an image is of a crowded park, with 2,000 people and 2,000 dogs, and the algorithm states there are 5,000 people, humans are far more likely to believe the algorithm. We

expect that task difficulty moderates the effect of algorithmic advice on belief change. Thus, we propose:

Corollary 2: *For intellective tasks, the effect of algorithmic advice on belief change will be stronger for a more difficult task.*

We believe these factors will have a similar impact on cognitive effort. People will respond more strongly to an algorithm's estimates, because answering intellective decision tasks correctly is a function of effort. When humans trust their advisors and understand how they arrive at a conclusion, they will use less cognitive effort. Furthermore, and most importantly, they are likely to assume that their peers are unlikely to put in as much effort as an algorithm. Thus, we propose:

Corollary 3: *For intellective tasks, algorithmic advice will result in less cognitive effort compared to the advice of crowds.*

We also propose that there are competing effects on how algorithmic advice will affect cognitive effort based on difficulty. In more difficult tasks, people may be more inclined to rely on algorithms, because they suspect that their human peers are unlikely to expend the necessary cognitive effort for an accurate answer. When executing easy tasks, on the other hand, humans may believe that human abilities outweigh an algorithm's tirelessness. This is because humans are less likely to think their peers are unwilling to put in the requisite effort during easy tasks compared to difficult tasks. In easier tasks, the work humans *are* adept at is also magnified. If a person believes that they, or other humans, are better than algorithms at some subset of tasks, and easier tasks make it more obvious which of those subset of tasks are embedded in the problem, then people will likely trust a crowd's answer more in easier tasks. Thus, the effect of

algorithmic advice should be weighted more heavily in hard intellective tasks than in easy intellective tasks.

Corollary *4: For intellective tasks, the effect of algorithmic advice on cognitive effort will be stronger for a more difficult task.*

There are two overriding effects that point to algorithmic advice resulting in greater confidence for difficult tasks. First, people are more confident of their decisions than the decisions of other humans, because of egocentrism and by perceiving themselves as more objective than other humans. Second, in prior research on intellective tasks, people exposed to algorithmic estimates reported greater confidence than people exposed to the estimate of a crowd (Logg et al. 2019). Confidence alone did not predict how accurate someone's guess was. Furthermore, confidence in algorithmic guesses was higher than confidence in the guess *of another individual*, but lower than confidence in one's own guess (Logg et al. 2019). This is in general agreement with the overconfidence literature. In intellective decision tasks, it has previously been established that confidence increases with algorithmic advice relative to the advice of a crowd of humans. We expect this effect to be even larger in our setting, in which the answer is more strongly a function of effort. Thus, we hypothesize:

Corollary *5: For intellective tasks, algorithmic advice makes humans more confident in their decisions than the advice of crowds.*

Lastly, we expect that for difficult intellective decision tasks, algorithmic advice will have a stronger effect than it will for easy tasks. This is because humans know algorithms are tireless, and they know fellow humans are likely to expend minimal effort when possible. Thus, in easier tasks, when humans must spend relatively little effort, the average human effort expended should

be relatively closer to the effort an algorithm expends. In harder tasks, when the gap between human effort and algorithmic effort widens, then humans will respond accordingly and prefer the advice of algorithms even more strongly. Thus, we propose:

Corollary *6: For intellective tasks, the effect of algorithmic advice on decision confidence will be stronger for a more difficult task.*

1.2.3 Proposition Development: Creative Tasks

When humans are given a recommendation in a decision task, they usually expend less effort on that decision. This is automation bias. Prior research on how humans exhibit automation bias is focused on intellective tasks (Dzindolet et al. 2002; Mosier and Skitka 1996). Our research takes this further by focusing on a creative decision task, in which effort alone does not predict success. This is a significant difference from intellective tasks. In creative tasks there is a strong relationship between long-term effort and success (Gladwell 2008). However, creating a high-quality output for a difficult creative task is a function innate ability and hours of practicing the creative act, rather than hours spent on the task. For example, extremely few could produce a painting as good as one by Van Gogh, even if they were given orders of magnitude more time. Thus, the relationship between effort and quality is weaker in creative tasks. We believe that this will cause humans to rely more on a human recommendation in a creative task than they would an algorithmic recommendation.

However, we expect this will be moderated by task difficulty. In an easy creative task, such as labeling a landscape, it is unlikely that subjects judge algorithms' recommendations to be significantly worse than those of a human crowd. If humans believe there are many good answers, then they will *likely believe even an algorithm can provide a satisfactory answer*. Thus, in an easy creative task, humans will show no preference between human and algorithmic

recommendations because they believe the optimal answer is not significantly better than many other answers. We conclude that for easy tasks, there is likely no effect of algorithmic recommendations compared with human recommendations. However, we expect the differences to manifest in hard tasks, where there is no obvious optimal solution.

In creative tasks, human effort still reigns supreme over algorithmic effort. No algorithm has won a contest in a creative domain against human experts, and it is not close. For example, algorithms have begun to pass the Turing test in extremely limited scenarios, such as when the algorithm claims that English is its second language, and that it is a young child (McCormick 2014). Given that algorithms can rarely pass the Turing Test, it is not surprising that they have yet to win a major award in a creative domain.

Lastly, most creative tasks are judged by human crowds, not algorithms. People care about reading books on the New York Times Bestseller List. Citations are a mark of influence in academia. Nobel prizes are awarded based on the vote of a committee. We thus expect the influence of crowds expect to be stronger and lead people to change more toward the advice of a crowd in a difficult, creative task.

Corollary 7: *For creative tasks, the effect of algorithmic advice on belief change will be stronger for a more difficult task.*

Prior research has looked at the *outcomes* of algorithmic and human advice, but not the *process*. Our investigation into **cognitive effort** bridges that gap. We expect that humans perform an evaluation, perhaps even a subconscious one, of input from any source and that this evaluation includes an assessment of how that recommendation occurred. If humans are confused about how their advisor arrived at a recommendation, then resolving this confusion could require

intense cognitive effort. The type of source likely has an effect on cognitive effort when humans are not familiar with the recommender's expertise at a given task, for example a computer's ability to be humorous.

We expect that for easy creative tasks, algorithmic recommendations will not cause different levels of cognitive effort compared to crowd recommendations. For easy creative tasks with an obvious suitable answer, people are not likely to be concerned with the source of the advice. However, in hard tasks, this changes. We believe that humans will spend more cognitive effort when they receive algorithmic suggestions in hard creative tasks, because they will be puzzled that the algorithm was able to come up with such a sophisticated solution.

It is possible that people are taken aback by what appears to be an algorithm that can make witticisms that surpass the subject's abilities. This may lead to the subject thinking about the nature of the algorithm (or even about the nature of their intelligence). When people spend time to puzzle about how an algorithm could possibly be so creative, they are likely to be expending more cognitive effort than they would if they were told that the same suggestion came from a crowd of humans. This is because people are familiar with suggestions from crowds for creative tasks. We believe these factors will lead to more cognitive effort, and thus we propose:

Corollary 8: *For creative tasks, the effect of algorithmic advice on cognitive effort will be stronger than the effect of crowd advice for a more difficult task.*

These effects should also manifest in **confidence.** Confidence is also likely to be negatively affected by algorithmic advice for difficult creative tasks. This is because we expect that some people are aware that algorithms are not adept at creative tasks, which will cause confusion and possibly a diminished assessment of the self. If an individual receives a recommendation from a

crowd, by contrast, they may have hope that there was a luminary in the given creative task in that crowd, and they are thus receiving particularly effective advice. Thus, we propose the following:

Corollary 9: *For creative tasks, the effect of algorithmic advice on confidence will be weaker for a more difficult task.*

1.2.4 Proposition Development: Mixed Motive Tasks

Algorithmic recommendations have recently become the subject of controversy in tasks such as hiring, autonomous cars, and determining bail. What these tasks have in common is that they involve a choice weighing the interests of multiple stakeholders. Algorithmic recommendations were initially heralded as a lifeline to companies – no longer would they need to rely on human judgment, with the messy implicit biases that humans subconsciously incorporate in their decisions. However, more recent research has indicated that these algorithms often incorporate racist or sexist criteria in their decision, even when race or sex is not specifically incorporated as a variable (Noble 2018).

In mixed motive tasks, people weigh conflicting interests. There is an innate human component to this – most of the time when we consider conflicting interests, there are human interests on both sides. Of course, it can be possible to have non-human interests in a mixed motive task, such as weighing the interests of animals versus humans in a debate about vegetarianism.

The algorithmic appreciation literature is not clear what humans would respond more strongly to in mixed-motive tasks. We believe people view this task as not entirely subjective but also not entirely objective. There is not a correct answer that a mathematical proof can deliver – it is not entirely objective. But there are degrees of how correct a reasonable person can be. It is not as subjective as arguing whether Brahms or Bach was the better composer.

The degree to which people change their beliefs mixed-motive tasks is likely to incorporate large numbers of unobserved characteristics. These could include how much a person already knows about algorithmic biases, how much they know about existing algorithms used mixed-motive tasks, and their beliefs about the biases of an average crowd.

There are strong reasons to argue for either algorithmic appreciation or aversion influencing mixed motive tasks. While writing the dissertation proposal, we argued there would be algorithmic aversion in mixed motive tasks, and that difficulty would not moderate the effect of advice source on reliance. However, after observing the effects for the intellective and creative tasks, we changed our position and proposed that there would be algorithmic appreciation and that it would be stronger in difficult tasks. We have updated the corollaries below.

Corollary 10: *For mixed motive tasks, the effect of algorithmic advice on belief change will be greater than the effect of the advice of a crowd on belief change.*

Corollary 11: *For mixed motive tasks, the effect of algorithmic advice will be moderated by task difficulty.*

1.2.5 Operationalization of Concepts

Cognitive effort

Cognitive effort is measured by the time required to complete a task. This operationalization has been used before in MIS (Moravec et al. 2019). We acknowledge that on the Mechanical Turk platform we may not observe different levels of time used on questions. Our review of tasks on Mechanical Turk indicated that many tasks do not reward hard work. Thus, even though in some tasks we reward effort with bonuses, we may observe significant changes across experiments and experimental conditions because the Turkers may not have viewed working hard on the assigned task as worth the effort.

Confidence

Confidence is measured using a four-point Likert scale in each experiment, by asking a question about the respondent's confidence that is appropriate to the task. For the intellective task, we ask how confident the participant is that their answer is within 10% of the true value. This is similar to the operationalization used in Logg et al. (2019). For the creative task, we ask whether the participant believes their answer is within the top 10% of submissions. For the mixed-motive task, we ask whether a subject is confident that their decision was fair.

We use a single-item measure to assess decision confidence. This is common in the algorithm appreciation literature. See Table 1.2.

Table 1.2: Prior Operationalizations of Confidence

Authors	**Question**	**Scale**	**Publication**
Logg et al. 2019	"How likely is it that your estimate is within 10 lb of the person's actual weight?	0 (no chance) to 100 (absolutely certain)	*Organizational Behavior and Human Decision Processes*
Dietvorst et al. 2016	"How much confidence do you have in your estimates?"	(None, Little, Some, A Fair Amount, a Lot)	*Management Science*

Belief Change

For the intellective and mixed-motive tasks, where belief change is relative to a numerical benchmark, we use **weight on advice** (WOA) as the dependent variable (Dawes and Corrigan 1974). WOA is the absolute value of the difference between the initial and revised judgment divided by the absolute value of the difference between the initial judgment and the advice (Logg et al. 2019).

$$WOA = \frac{|final\ estimate - initial\ estimate|}{|recommendation - initial\ estimate|}$$

For the creative task, we measure weight on advice by using the final answer provided by the subject. The four answer choices were as follows, with the first two choices having text boxes allowing for input from the subjects. More details are available in Chapter 3.

1. I want to create a new caption to improve my chances of receiving a bonus
2. I want to tweak the recommended caption to improve my chances of receiving a bonus
3. Use the recommended caption
4. Keep my first caption

We synthesize how we operationalize our dependent variables in Table 1.3.

Table 1.3 Dependent Variables Across Experiments

	Confidence	Belief Change	Cognitive Effort
Intellective	Likert Scale	WOA	Time
Creative	Likert Scale	Custom Measure of WOA	Time
Mixed-motive	Likert Scale	WOA	Time

1.3 Dissertation Structure

In Chapter One we use theory to inform the Proposition and Corollaries. Chapters 2, 3, and 4 focus on how each experiment was conducted. All three chapters use a similar strategy for measuring the dependent variables and for manipulating the independent variables.

1.3.1 Intellective task experiment

The **intellective task** experiment explores how humans respond to human and algorithmic input, by asking subjects to estimate the size of a crowd in an image. This experiment brings together machine learning and automation bias in the context of large, diverse, connected crowds. Crowd counting is a task with an objective, correct answer, and thus is an intellective task, in the choose

quadrant of the task circumplex model. The experimental subjects were exposed to the guesses of an apparent crowd and an AI. In order to ensure that we were isolating the effect of receiving recommendations from an algorithm versus from a crowd, each subject will be told that the advice they receive is the product of either an algorithm or crowd, when in reality all recommendations will identical, and all will be the correct answer. Subjects will have the option to revise their answers, and the manipulation will allow us to observe whether humans are more influenced by humans or algorithms depending on task difficulty.

The scholarly implications are important – knowing the theoretical mechanisms behind when people shift their behavior in the perceived presence of fellow humans or algorithms is a novel, important contribution.

The practical implications are also significant – counting crowds is important for government services, and unbiased estimates from crowd-counting algorithms can be useful in situations such as estimating the size of the medical or policing resources needed. Furthermore, it is important to determine the general level of confidence that the public has in these machine generated counts.

1.3.2 Creative task experiment

The **creative task** experiment will explore how humans differentially rely on AI suggestions compared to human suggestions for a creative task. Each subject will write a caption for an image. It is likely that humans will perceive AI-provided advice differently from human-provided, because creative tasks such as writing do not have a correct answer, though they can be of differing quality.

Intellective tasks with correct answers have been deeply investigated (Alexander et al. 2018; Grove et al. 2000; Parasuraman and Riley 1997), as have decision-making tasks with no correct

answers, such as moral dilemmas (Awad et al. 2018; Bonnefon et al. 2016). However, how humans interact with AI in creative tasks is underexplored, because it is harder to measure outcome quality.

For the creative task, knowing how people respond to algorithmic help could inform a huge swathe of tasks algorithms are currently employed in. Recommending songs or genres on apps such as Spotify, recommending dates on apps such as OkCupid, or recommending phrasing using apps such as Grammarly, are all non-intellective tasks with creative components that might be informed by this research.

1.3.3 Mixed-motive task experiment

The **Mixed-motive** task experiment will investigate how human suggestions are perceived differently from algorithmic suggestions in mixed-motive tasks resolving conflicts of interests, which belong to the Negotiate quadrant of the circumplex model. Subjects will be asked to assess how much bail an accused individual should pay.

The closest research to our experiment is the research on moral situations, which focuses on what the algorithm should do and who should receive blame (e.g. Ames and Fiske 2015; Awad et al. 2018). We are aware of no research that investigates the cognitive effort of subjects, nor their confidence, nor their change in beliefs depending on the type of information source in mixed motive (or morality-related) tasks.

The mixed-motive task also has significant social value. There is a push towards more algorithmic assessment in bail decisions, and learning how the public perceives this decisions is an important step in the acceptance or rejection of these technologies.

1.4 Experimental Procedures Common to all Research Projects

All experiments will use the Judge Advisor System, in which subjects submit their answer to a prompt, and then are exposed to new information (Sniezek and Buckley 1995). All will compare subjects' responses to the advice of an algorithm or to a crowd. Each will keep the quality of advice identical. This will let us isolate the effect of the type of task, the difficulty of the task, and the type of advice.

A graphical summary of the experimental design for each survey is shown below in Figure 1.4.

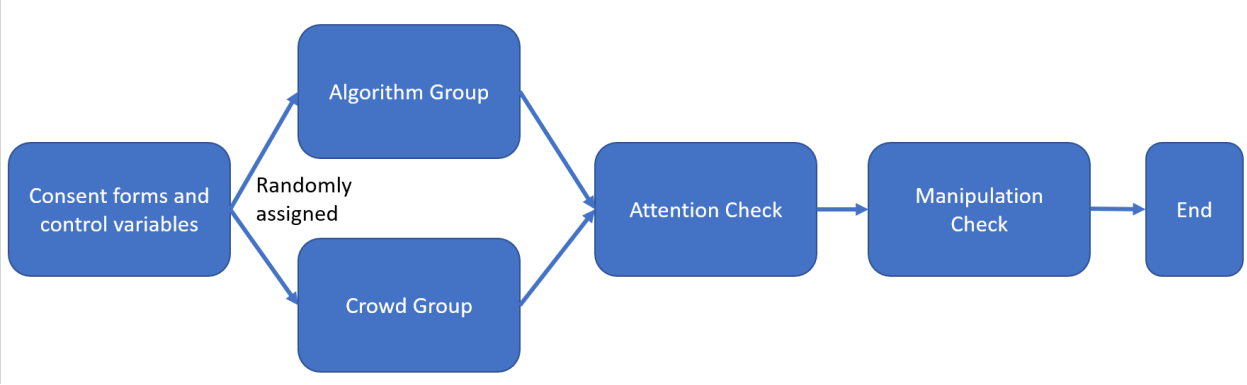


Figure 1.4: Survey Design

1.4.1 Incentives

Because incentives cause higher levels of effort (Camerer and Hogarth 1999), we provide incentives where applicable. For the intellective and creative tasks, we tell each participant, during their first guess and their revised estimate, that the closer they are to the true answer the more money they will earn. We do not provide incentives for the mixed-motive task, because it is possible that providing an incentive would cause the subject to view it as an intellective task, thus confounding our results. Providing incentives through bonuses is an extremely widespread practice in studies on Amazon Mechanical Turk. In addition to bonuses, we pay all subjects a

baseline of $10.00 per hour. The widespread consensus is that paying $0.10 per minute ($6.00 per hour) is the minimum ethical standard for payment.

1.4.2 Task timing

For each question, subjects submit both their answer and how confident they are in it. We automatically collect how much time the subject spends on the question. We also conduct a robustness check to see whether excluding people based on time changes the results meaningfully.

1.4.3 Disclosure of answers

We never disclose the correct answers to our questions to the participants. This prevents the possibility that people inform later participants of the answers. We award bonuses only after we have collected all responses.

1.4.4 Preregistration

We preregistered experiments at the Open Science Framework (osf.io). The intellective task (https://osf.io/ym3ug) and creative task (https://osf.io/mjvre), are publicly viewable. The mixed-motive task (https://osf.io/7v4p3) is currently under embargo and thus cannot be viewed at the time of the publication of this dissertation.

1.4.5 Experimental consistency

We keep the underlying advice quality the same to control for the effect on belief change, and we keep the incentive structure the same. However, in a supplemental experiment for the intellective task we introduce bad advice. We have not conducted the same type of supplemental experiments for the creative and mixed-motive tasks, although this would be an area for future experiments.

1.4.6 Control Variables:

We randomly assign the subjects into one of the two groups in each experiment. This random assignment makes it more likely that any differences we observe between groups are likely the result of the experimental conditions (Van de Ven 2007). Control variables that are likely to covary with both the independent and dependent variables should be included in an experiment. However, it is possible to use unnecessary control variables, and a strong principle for control variables is "When in doubt, leave it out" (Carlson and Wu 2012). Prior research in algorithmic appreciation has observed that age, and gender do not effect algorithmic appreciation (Logg et al. 2019), and there is no theoretical reason to assume demographic characteristics such as gender or race will predict algorithmic appreciation. Thus, we do not collect data on those characteristics.

Systematic differences between the sample and the population are important when researchers want to make claims about point estimates of populations. However, for research that simply wants to observe the relationship between measured variables, then systematic differences become less important. In our case, even though workers on AMT have systematic variation from the population, it is very unlikely that AMT workers will display different behavior in relation to the manipulated variables. Thus, even if the true population effect of, for example algorithmic aversion, is different from what we observe, the difference in effects between easy and hard tasks is very likely to approximate the true value, and the difference between types of tasks is likely to approximate the true value. Prior research has noted that if researchers are not interested in population-level estimates then AMT is likely an excellent source of subjects for experimental studies (Chandler and Shapiro 2016).

We gather experiment-specific control variables for the creative and mixed-motive experiments. For the creative task we collect data on general word-play creativity through the

Remote Associates Test (RAT) (Mednick 1968). Because the experiment focuses on the ability to generate a caption, we control for innate word-play ability, and are able to test whether people who are more skilled at wordplay are more likely to rely on algorithms.

For the mixed-motive task we gather data on how much a subject identifies with criminal others (ICO) (Simourd 1997) and an individuals beliefs about cutting-edge technologies, through the Insecurity component of the Technology Readiness Index 2.0 (TRI) (Parasuraman and Colby 2015). We collect these two measures to control for any biases towards criminals and general inclinations towards technology.

1.4.7 Subject Selection Plan

We recruit human subjects from Amazon Mechanical Turk (AMT), an online platform which connects people with Human Intelligence Tasks (HITs). Workers on Mechanical Turk are often called Turkers, a phrasing we adopt in this thesis. Common HITs on AMT include labeling images, transcription, and filling out surveys. Samples from Amazon Mechanical Turk are transforming academic studies that rely on human participants, because studies on Mechanical Turk are generally cheaper and more convenient (Dance 2015). Samples from Mechanical Turk are usually more diverse than other online samples (Buhrmester et al. 2011; Paolacci and Chandler 2014), and the quality of data produced by individuals from Mechanical Turk can rival engineering graduate students on complex tasks, provided there is a sufficient tutorial (Staffelbach et al. 2015). In comparison to student samples, Turkers are more representative of the US population than college students and people in college towns (Berinsky et al. 2013). Turkers are also generally younger and better educated than the US population (Paolacci and Chandler 2014). Turkers are also usually honest and consistent with non-identifying

demographic data, such gender or age. For example, personality characteristics measured three weeks apart using the Big Five Inventory averaged $r = .85$ (Buhrmester et al. 2011).

There has been some criticism of Mechanical Turk samples. For example, Mechanical Turk workers are often asked to take IQ tests – after enough IQ tests, individual IQ test scores increase significantly, despite their true intelligence remaining unchanged. An interesting instantiation of this is that there is a high correlation between number of HITs a Turker has completed and that Turker's Cognitive Reflection Task score (Chandler et al. 2014).

We exclude responses from subjects who do not complete the survey, fail the attention check (see Appendix 3), or fail the manipulation check (see Appendix 4). This is in accordance with the practices used in other studies of algorithmic and human recommendations (Yeomans et al. 2019). We exclude responses of people whose initial guess is equal to the advice they receive, in accordance with the best practices of using the Weight on Average measure (Gino and Moore 2007). We also exclude people who always either reject or accept the advice, due to the high probability that they are not paying attention.

1.4.8 Manipulation check

We check whether subjects register the source of the advice. Thus, at the end of each survey we ask subjects whether they were given advice from a human crowd or from an algorithm. If a subject does not recall the source of advice, or recalls incorrectly, they are excluded from further analysis.

1.5 Summary of the Dissertation:

As a summary of the dissertation, we present our model, proposition, and corollaries in Figure 1.5.

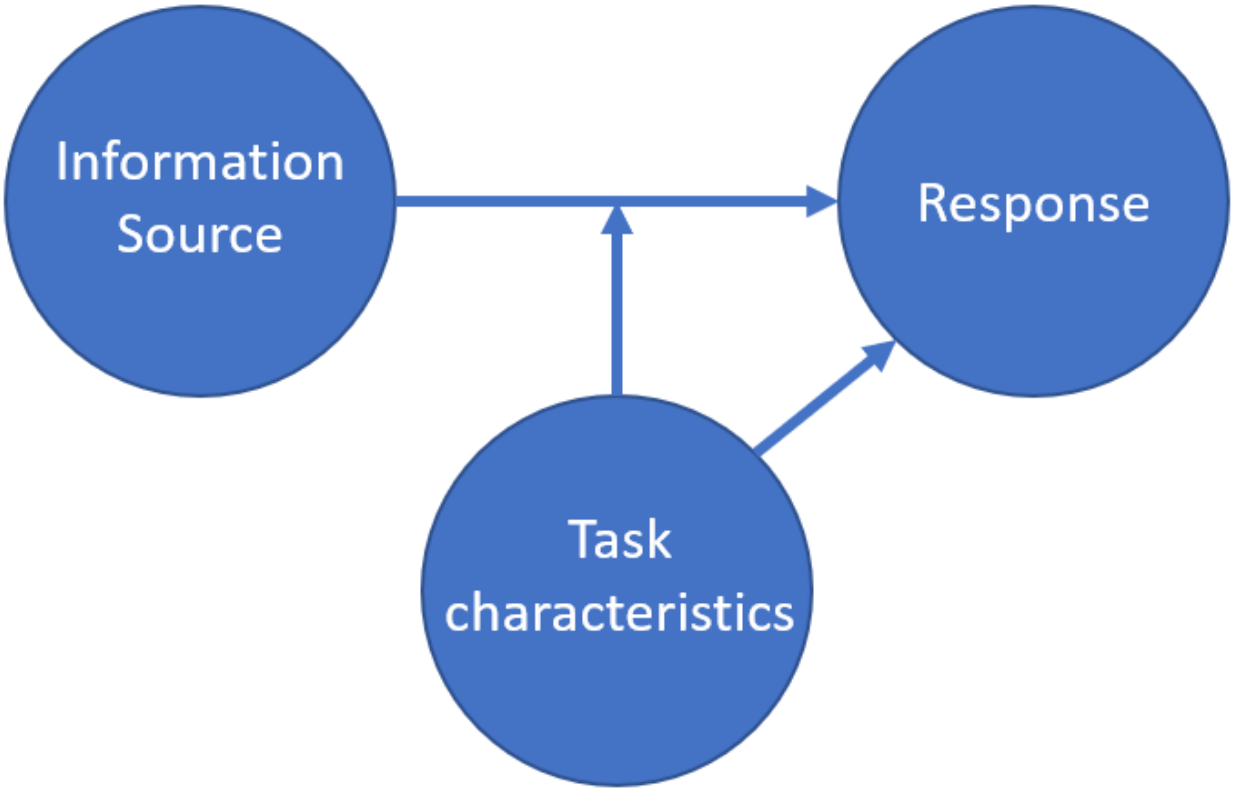


Figure 1.5: Research Model

The proposition and their corollaries are as follows:

Proposition 1: *An individual's response to decision advice is determined by characteristics of the task and the information source.*

Corollary 1: *For intellective tasks, the effect of algorithmic advice on belief change will be stronger than the advice of a crowd.*

Corollary 2: *For intellective tasks, the effect of algorithmic advice on belief change will be stronger for a more difficult task.*

Corollary 3: *For intellective tasks, algorithmic advice will result in less cognitive effort compared to the advice of crowds.*

Corollary *4: For intellective tasks, the effect of algorithmic advice on cognitive effort will be stronger for a more difficult task.*

Corollary *5: For intellective tasks, algorithmic advice makes humans more confident in their decisions than the advice of crowds.*

Corollary *6: For intellective tasks, the effect of algorithmic advice on decision confidence will be stronger for a more difficult task.*

Corollary 7: *For creative tasks, the effect of algorithmic advice on belief change will be stronger for a more difficult task.*

Corollary 8: *For creative tasks, the effect of algorithmic advice on cognitive effort will be stronger than the effect of crowd advice for a more difficult task.*

Corollary 9: *For creative tasks, the effect of algorithmic advice on confidence will be weaker for a more difficult task.*

Corollary 10: *For mixed-motive tasks, the effect of algorithmic advice on belief change will be greater than the effect of the advice of a crowd on belief change.*

Corollary 11: *For mixed-motive tasks, the effect of algorithmic advice will be moderated by task difficulty.*

Chapter 2

Humans Rely More on Algorithms than Social Influence as a Task Becomes More Difficult[1]

[1] Bogert, E., Schecter, A. & Watson, R.T. Humans rely more on algorithms than social influence as a task becomes more difficult. *Sci Rep* **11,** 8028 (2021).
Reprinted here with permission of the publisher.

2.1 Abstract

Algorithms have begun to encroach on tasks traditionally reserved for human judgment and are increasingly capable of performing well in novel, difficult tasks. At the same time, social influence, through social media, online reviews, or personal networks, is one of the most potent forces affecting individual decision-making. In three preregistered online experiments, we found that people rely more on algorithmic advice relative to social influence as tasks become more difficult. All three experiments focused on an intellective task with a correct answer and found that subjects relied more on algorithmic advice as difficulty increased. This effect persisted even after controlling for the quality of the advice, the numeracy and accuracy of the subjects, and whether subjects were exposed to only one source of advice, or both sources. Subjects also tended to more strongly disregard inaccurate advice labeled as algorithmic compared to equally inaccurate advice labeled as coming from a crowd of peers.

Algorithms have mastered checkers (Schaeffer et al. 2007), chess (Dockrill 2017; Silver et al. 2017), poker (Brown and Sandholm 2019), and tasks with fewer boundaries such as information search (Brin and Page 1998). This expertise has led humans to rely heavily on algorithms. For example, people rely so heavily on Google that they treat it as an external memory source, resulting in them being less able to remember searchable information (Sparrow et al. 2011). As big data has flourished, people have become so comfortable with algorithms that drivers will sleep in their self-driving cars (Baker 2019), go on dates with algorithmically-recommended matches (Hickey 2019), and allow algorithms to run their retirement accounts (Chafkin and Verhage 2018). However, there are some tasks for which humans prefer to take advice from other humans, such as in medical advice (Promberger and Baron 2006) or predicting how funny a joke will be (Castelo et al. 2019).

Humans often demonstrate greater reliance on advice from algorithms compared to non-algorithmic advice, exhibiting *algorithmic appreciation* (Logg et al. 2019). Relying upon algorithms for analytical tasks is typically advantageous. Even simple algorithms, such as weighting all variables equally, can outperform human prediction (Dawes 1979). In a meta-analysis of 136 studies, algorithms were 10% more accurate, on average, than non-algorithmic (human) judgment (Grove et al. 2000). Consequently, for analytical tasks, we would expect a rational human to demonstrate algorithmic appreciation.

Of course, much of human behavior is not strictly rational (Kahneman 2011). People tend to discount or disregard advice, even when it is not logical to do so (Yaniv and Kleinberger 2000). Often, the source of advice dictates how much it is discounted. When people discount advice from other people less than they discount advice from algorithms, particularly after observing an algorithm make a mistake, they demonstrate *algorithmic aversion* – the opposite of algorithmic

appreciation. There is evidence for both algorithmic aversion (Yeomans et al. 2019) and appreciation (Abeliuk et al. 2020; Kawaguchi 2020; Logg et al. 2019), and it is task dependent (Castelo et al. 2019). Prior research has also shown that people rely on advice more heavily when tasks become more difficult (Gino and Moore 2007). However, this effect may not be uniform across sources of advice.

Given these empirical observations, we question whether task difficulty is an important explanatory variable in determining whether people demonstrate algorithmic appreciation or aversion. In our studies of reliance on algorithmic advice, we consider two critical factors: the source of advice and task difficulty. We conducted three preregistered experiments with N = 1,500 participants to test the influence of algorithmic advice, compared to social influence, on human decision making. Broadly speaking, social influence encapsulates the myriad ways that humans change their behavior based on the actions of other people. Prior experiments show that when humans are exposed to social influence, the wisdom of the crowd can be reduced (Lorenz et al. 2011), and that the structure of the social network dictates how social influence affects decision-making (Becker et al. 2017). Based on subject responses across multiple tasks and under different manipulation conditions, we find that people rely more on algorithmic relative to social advice, measured using Weight on Advice (WOA) (Dawes and Corrigan 1974). Further, we establish that advice acceptance varies as tasks increase in objective difficulty and as advice varies in quality.

2.2 Results

In each experiment, subjects were asked how many people were in a photograph and provided advice that was purported to be from either “an algorithm trained on 5,000 images” or “the average guess of 5,000 other people.” There was no other introduction to the algorithm or a

description of what types of people made the estimates. An equal number of subjects were in each group. We used a large group of peers as a reference group because large groups often makes guesses that are accurate, on average (Galton 1907; Lorenz et al. 2011; Surowiecki 2004), and people respond more strongly to advice from large numbers of people compared to advice from a single person(Mannes 2009). We chose to design the experiment such that the only difference between the two sources of advice was the label, so that we could isolate the effect of advice source. This is a common method of judging reliance on algorithmic advice (Castelo et al. 2019; Logg et al. 2019).

We use the Judge Advisor System (JAS) in every experiment. The JAS is an experimental method in which subjects answer a question, are provided advice related to that question, and then asked to answer the question again (Bonaccio and Dalal 2006; Liberman et al. 2012; Sniezek and Buckley 1995; Sniezek and Van Swol 2001). In experiments using the JAS, a common dependent variable is Weight on Advice (WOA). WOA calculates the degree to which an individual changes their answer towards the advice, and thus is a useful measure for describing the extent of algorithmic appreciation or aversion.

All tests described below are two-tailed at the alpha 0.05 significance level and are t-tests of coefficient values from a regression. Summary statistics of the data can be found in Table S1.

2.2.1 Experiment 1: Advice as Between Subjects Treatment

In the initial experiment, subjects were asked to determine how many people were in a picture and received advice that was labeled as either algorithmic or the average of human guesses, and they never received advice from the other source. All advice was the true answer, which was determined by the publisher of the dataset (Idrees, Saleemi, et al. 2013).

2.2.2 Task Validation and Randomization Check

We first assessed whether subjects responded to more difficult problems by taking more time, being less confident, and being less accurate. When comparing within-person easy to hard questions, individuals are significantly more accurate ($t = 2.745$; $P = 0.006$; 95% confidence interval (CI) = 0.281 to 1.684), more confident ($t = 24.291$; $P < 0.001$; 95% CI = 0.536 to 0.630) and take less time ($t = 4.179$; $P < 0.001$; 95% CI = 0.041 to 0.113) for easy problems. In all three models we observed that subjects relied more on advice in difficult questions. Subjects placed more weight on advice for hard questions in our baseline model ($B = 0.150$; $P < 0.001$; 95% CI = 0.134 to 0.167), in the model including hypothesized interactions ($B = 0.132$; $P < 0.001$; 95% CI = 0.108 to 0.155), and in the model including all interactions and control variables ($B = 0.081$; $P < 0.001$; 95% CI = 0.057 to 0.105). Thus, we conclude that subjects perceived the relative difficulty of the questions as designed.

We compared the average initial accuracy, initial confidence, and initial time taken across treatments using a two-sample t-test. For individuals exposed to algorithmic advice, there was not a statistically significant difference in initial accuracy ($t = -0.767$; $P = 0.443$; 95% CI = -1.000 to 0.438). Individuals receiving algorithmic advice reported higher initial confidence ($t = 3.93$; $P < 0.001$, 95% CI = 0.149 to 0.050) and spent less time on a problem ($t = 2.00$; $P = 0.045$; 95% CI = 0.00076 to 0.07293) when we analyzed all questions. However, if we compare confidence for only the first question subjects saw (before they received any advice), the difference in initial confidence is not significant ($t = -0.20$; $P = 0.403$; 95% CI = -0.203 to 0.082). The difference in time spent on a problem is also not significant when looking at only the first question ($t = 0.054$; $P = 0.586$; 95% CI = -0.084 to 0.149). These results indicate that subjects were effectively equivalent in both conditions, as expected from random assignment. In

the aggregate, when they received algorithmic advice, subjects became more confident in their initial guesses and spent less time on a problem in later questions.

2.2.3 Main Analyses

To test the preregistered hypotheses, we fit a series of mixed effects linear regressions with random slopes for each subject. The regression results are given in Table S2.

Effect sizes and confidence intervals are shown for the effect of algorithmic advice and difficulty in Figure 2.1 below. All figures were made using ggplot2 (Wickham 2016), version 3.3.

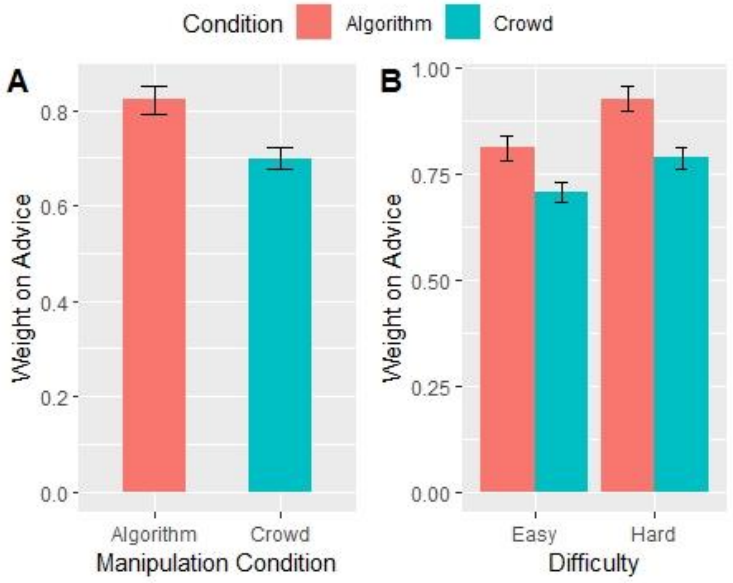


Figure 2.1 Source of Advice Affects Subject Weight on Advice (Experiment 1). Each bar chart depicts results of the mixed effects regression model on N = 5,083 observations. All models include accuracy as a control. Error bars correspond to the standard error of the estimated effect. (A) shows the main effect of advice source on WOA; the difference across conditions is significant ($p < 0.001$). (B) shows the effect of advice source on WOA across levels of difficulty; all differences are significant ($p < 0.05$). Panel A shows the effects using Model 1 from Table S2, Panel B shows the effects using Model 2 from Table S2.

There is a positive and significant main effect of algorithmic advice on WOA ($B = 0.108$; $P < 0.001$; 95% CI = 0.058 to 0.158). Similarly, we find a positive and significant interaction effect of algorithmic advice and difficulty on WOA ($B = 0.036$; $P = 0.029$; 95% CI = 0.004 to 0.068)). That is, subjects who receive advice from algorithms on easy problems will revise their responses 11% more than subjects receiving advice from the crowd. Further, if a problem is

difficult, subjects revised their answers by an additional 3.6% more when they receive advice from an algorithm, indicating that subjects rely even more on algorithms than they do on the advice of a crowd when the task is difficult.

Finally, we checked whether highly accurate subjects were disproportionately relying on algorithmic advice and found there was no significant difference (B = -0.007, P = 0.81, 95% CI = -0.067 to 0.053). We did not hypothesize this in our preregistration for the first experiment, although we investigated this further in experiments two and three.

2.2.4 Experiment 2: Advice Source as Within-Subjects Treatment

In the second experiment we again show subjects pictures of human crowds and ask them to guess how many people are in the picture. However, in experiment two we make advice source a within-subjects condition. We do so because within-subject designs better control for any differences among subjects (Gueorguiva and Krystal 2004). Subjects received five questions for which they received advice that was labeled as the average of 5,000 human guesses, and five questions for which they received advice that was labeled as being from an algorithm trained on 5,000 pictures. We also introduced numeracy as a new control variable in this experiment (Schwartz et al. 1997). The second experiment includes 514 people, after following the same exclusion procedures for the first experiment, with the exception of the manipulation check, which we did not use because the advice condition was within-subjects.

Results from the second experiment reinforced the results from the first experiment. In the baseline model without interactions, subjects relied more strongly on advice when it was labeled as algorithmic (B = 0.069; P < 0.001; 95% CI = 0.052 to .086). When interactions are analyzed however, the main effect of algorithmic advice becomes non-significant (B = 0.027; P = 0.18;

95% CI = -0.013 to 0.0670). We also found that participants relied more on algorithmic than crowd advice for difficult questions (*B* = 0.038; *P* = 0.037; 95% CI = 0.002 to 0.074). The results indicate that there is a net effect of algorithmic appreciation, but that a positive impact is driven entirely by a reliance on algorithms for hard problems. The effects and associated standard errors can be seen in Figure 2.2.

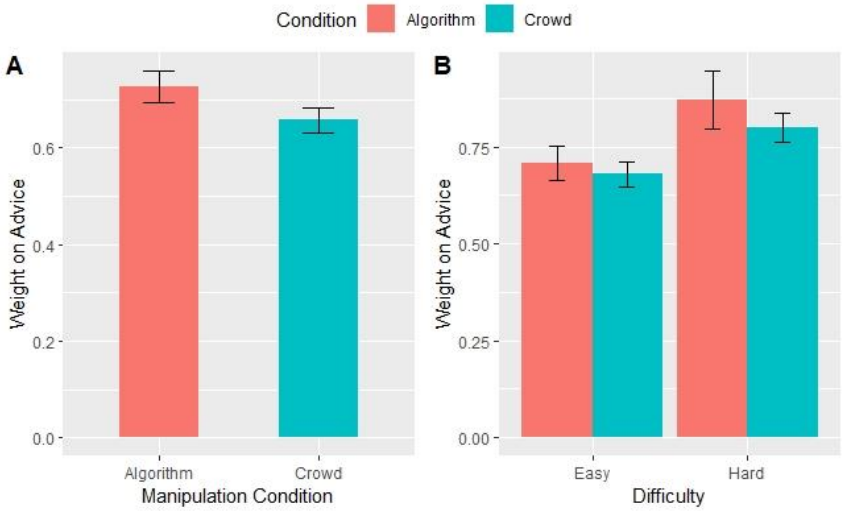


Figure 2.2. Source of Advice Affects Weight on Advice (Experiment 2). Each bar chart depicts results of the mixed effects regression model on N = 4,905 observations. All models include accuracy as a control. Error bars correspond to the standard error of the estimated effect. (A) shows the main effect of advice source on WOA; the difference across conditions is significant (p < 0.001). (B) shows the effect of advice source on WOA across levels of difficulty; all differences are significant (P < 0.05). Panel A is for Model 1 in Table S3, Panel B is for Model 2.

Finally, more accurate subjects relied on algorithmic advice to the same degree as less accurate subjects (*B* = 0.045; *P* = 0.15, 95% CI = -0.017 to 0.107).

2.2.5 Experiment 3: Incorporating Low-Quality Advice

In the third experiment, we relax a significant assumption made in the other two experiments, in which the advice provided was always the correct answer, and thus was strictly high-quality advice. In the third experiment, we introduce low quality advice, to test whether the findings relied on providing subjects with high quality advice. Low quality advice was a within-subjects condition such that all participants saw the correct answer as advice for half of the questions, and

advice that was 100 percent too high for the other half of the questions. The choice of 100 percent too high was based on pilot that tested advice that was too high by 50 percent, 100 percent, and 150 percent. Experiment three reinforced the results from the first two experiments. We show the effects of quality and advice source below in Fig. 3 – these effects are taken from Model 3 in Table S4.

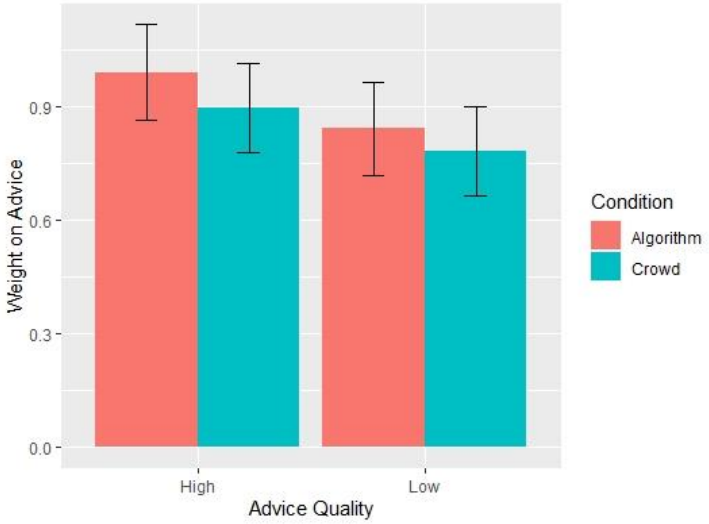


Figure 2.3. Source of Advice Affects Weight on Advice (Experiment 3). Each bar chart depicts results of the mixed effects regression model on N = 4,365 observations. The exact model used is Model 3 in Table S4. Error bars correspond to the standard error of the estimated effect.

Subjects relied more strongly on algorithmic advice (B = 0.059; P < 0.048; 95% CI = 0.0004 to 0.1180), and this effect was magnified for difficult tasks (B = 0.037; P = 0.028; 95% CI = 0.004 to 0.071). Subjects who were more accurate initially did not rely more on algorithmic advice than the advice of a crowd (B = 0.064; P = 0.052). Subjects relied more strongly on good advice than on bad advice (B = 0.11; P < 0.001, 95% CI = 0.084 to 0.144), and this effect was greater when the source was an algorithm (B = 0.035; P = 0.043; 95% CI = 0.001 to 0.068). Another way to interpret this finding is that subjects penalized algorithms more for providing bad advice. When a crowd of peers provided low quality advice compared to high quality advice, the baseline from experiments one and two, subjects exhibited a WOA of 11% lower, while bad

advice from an algorithm reduced WOA by more than 14%. Lastly, the effect of bad advice was moderated by the difficulty of the question (B = -0.146; P < 0.001; 95% CI = -0.181 to -0.111). What this means in light of the research question is more nuanced. Our research question is whether people rely more on algorithmic advice than social advice when intellective tasks become harder, and whether advice quality moderates that effect. The interaction of advice quality and question difficulty may not specifically answer that question, but what it does tell us is that subjects are sensitive to both difficulty and quality in tandem, even after accounting for other factors. Further, we find that our primary treatment – advice source – has a significant effect on WOA even after including this interaction. This result suggests that source has a robust effect across combinations of conditions, lending additional support to one of our main claims.

2.2.6 Additional Analyses and Robustness Checks

It is possible that the findings are due to some unobservable individual skill or quality not eliminated by random assignment. Consequently, we conducted an analysis of covariance(Keppel 1991) to predict WOA and change in confidence using initial accuracy, initial confidence, and initial time spent on the task across each level of advice source and difficulty. Thus, we are able to determine the effect of advice source and difficulty on WOA after controlling for differences in individuals' skill (accuracy), perceived skill (confidence), and effort (time).

Across all levels of accuracy, initial confidence, and initial time, subjects consistently exhibited higher WOA when receiving advice from an algorithm, when comparing hard questions to hard questions and easy questions to easy questions, see Figure 2.4 below. This combination of algorithmic advice and problem difficulty creates the most significant change in subject estimates, with virtually no overlap of the 95% confidence intervals.

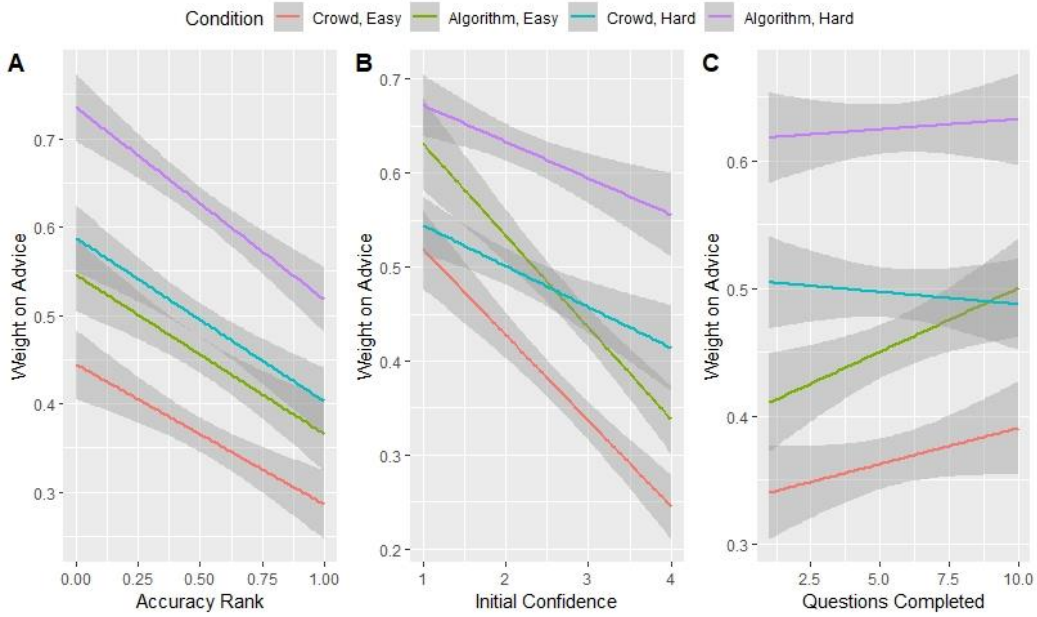


Figure 2.4 Effects of Accuracy, Confidence, and Initial Time. Each plot depicts a linear regression using a control variable to explain WOA delineated by advice condition and difficulty (N = 1,249). The shaded areas depict 95% confidence intervals. WOA is regressed on (Panel A) initial accuracy, (Panel B) initial confidence, and (Panel C) the number of questions a subject has completed thus far.

Finally, we conducted robustness checks on the main models (Fig. S2). We removed subsets of the data to ensure extreme values were not adversely impacting the findings. We excluded the top and bottom 2.5% responses for confidence, time per question, and overall time spent. Across all alternative regressions the findings are consistent. To check for multicollinearity, we removed control variables stepwise. Removing accuracy, initial confidence, and both accuracy and initial confidence did not change the results.

2.2.7 Summary of Experimental Results

When comparing effects across all three experiments, there is remarkable consistency in the most important finding, namely, people rely more on algorithmic advice than crowd advice as tasks become more difficult. When using the baseline model outlined in the Analytical Approach in the following section, we find no significant differences in the interaction between algorithmic

advice and question difficulty. For all three experiments, the effect is between 0.035 and 0.038, indicating that people rely substantially more on algorithmic advice for difficult questions than for easy questions, even after accounting for numeracy, accuracy, confidence, advice quality, and the number of prior questions answered. A summary of the results in each experiment in Table 1.

2.3 Discussion

These three experiments contribute to the burgeoning literature on social influence, the wisdom of the crowds, (Becker et al. 2017; Guilbeault et al. 2018; Lorenz et al. 2011) and the role of algorithms in decision making. We provide large-sample experimental evidence that for intellective tasks, humans are more accepting of algorithmic advice relative to the consensus estimates of a crowd, echoing the results of prior literature (Logg et al. 2019). Most importantly, we found that subjects exhibit greater algorithmic appreciation as intellective tasks became more difficult. With difficult intellective tasks, there is a robust and practically significant impact of algorithmic appreciation.

Other findings using experiments and the Judge Advisor System have found that the difficulty of a task had no effect on algorithmic appreciation (Logg et al. 2019), or that as tasks became more difficult humans would rely less on algorithms (Abeliuk et al. 2020). Our paper finds the opposite, while more strongly and precisely manipulating difficulty, in an environment with incentives to do well, while controlling for the accuracy of subjects, whether the advice was within or between subjects, the quality of the advice, the numeracy of the subjects, and the confidence of the subject.

Table 2.1. Summary of Experimental Findings

Claim	Experimental support		
	1	2	3
Subjects will rely more on algorithmic advice than equally good human advice.	Yes	Partial*	Yes
Subjects receiving algorithmic advice will rely more on advice in difficult questions.	Yes	Yes	Yes
Highly accurate subjects will rely more on algorithmic advice than inaccurate subjects**	No	No	No
Bad advice from an algorithm more strongly reduces weight on advice than bad advice from a crowd	N/A	N/A	Yes

*supported in baseline model without interactions and controls

**This hypothesis was preregistered only for Experiment 3. When using an alternative measure of accuracy that allowed for significant outliers, we observed a positive and significant effect of the interaction between accuracy and algorithmic advice in both Experiment 1 and 2, so we preregistered a hypothesis about this effect for Experiment 3. We then observed that the observed post-hoc effect in Experiment 1 and 2 were due to outliers. We thus changed the accuracy measure to be the percentile rank of accuracy on a question, to prevent outliers from strongly influencing results.

Humans may show a preference toward algorithmic advice depending on how close their initial guess is to the advice they receive (Abeliuk et al. 2020). Those with a history of recent accuracy may strongly discount the advice of others, while incorporating the advice of an algorithm, because if people perceive they are good at intellective tasks then they likely question why they should accept the advice of the less skilled crowd (Abeliuk et al. 2020). Thus, we expected that highly accurate individuals would demonstrate algorithmic appreciation more than the less accurate; however, our experiments did not support this claim.

Humans can discriminate between good and bad advice, and rely less on low-quality advice than they do on high-quality advice (Harvey et al. 2000). However, the interaction between advice quality and whether the advice comes from an algorithm or group of other humans is largely ignored – humans might respond differently to algorithmic mistakes compared to mistakes from a wise crowd when a question is easy or hard. We build on prior research that examines algorithmic advice-taking (Abeliuk et al. 2020; Dietvorst et al. 2015) and advice quality by

introducing a reference group, the advice of a crowd, with equally good (bad) advice. We tested whether low-quality advice from algorithms creates a stronger negative effect than low-quality advice from humans. Our experimental results suggest that when advice quality deteriorates (i.e., goes from high to low), algorithms will be penalized to a greater degree than a crowd of advisors.

An important feature of our experiment is the choice of reference group relative to algorithmic advice. Large, dispersed human crowds have both historically made accurate guesses(Galton 1907; Surowiecki 2004) and people strongly respond to the wisdom of the crowd(Mannes 2009). Indeed, we observed that subjects who received advice from the crowd significantly revised their answers. However, the recommendation of an algorithm still had a stronger effect, across multiple specifications and experimental conditions. Thus, we argue that simply labeling advice as "algorithmic" or derived from machine learning can cause a meaningful shift in human behavior. We used a relatively weak manipulation –simply changing the label of the advice as either algorithmic or the average of a crowd. The consistent, statistically robust differences observed by changing only a few words demonstrate that these effects are strong.

The study has some limitations. The subjects recruited might have been more comfortable with technology and thus had a higher propensity towards algorithmic advice than the larger public. However, even if the subjects demonstrate more algorithmic appreciation than the public overall, we expect that the shift towards algorithmic advice for difficult, intellective tasks is a universal effect. Further, as experiment two demonstrates, there is equal appreciation for crowd and algorithmic advice when completing easy tasks. It is also possible that this task, which is relatively mundane and tedious, may have unique characteristics that cause people to lean disproportionately on algorithmic advice as difficulty increases. Specifically, for intellective

tasks, people may be more likely to rely on algorithmic recommendations, whereas for tasks that have significant negotiation or generative components, which require subjective judgments, people may feel less comfortable relying on algorithms entirely. However, we leave these alternative task types to future research.

Governments and corporations have a strong interest in leveraging AI. This can be at the expense of consumers and citizens, who may not know that their data are harvested, stored, and analyzed. People whose data are used to calibrate algorithms could be affected by them, positively or negatively, by social or corporate policies based on AI. The public seeks interventions that solve important societal problems, such as income inequality, medical research, or systemic biases in institutions. Because interventions can be harmful, carefully managed research, followed by trials, is necessary to minimize unintended effects. If governments wish to spend citizens' taxes wisely, we need them to take an evidence-based approach to social policy, with AI as a potential research methodology. Citizens need to be engaged by freely sharing data that might address private matters, such as spending patterns when evaluating the potential outcomes of universal basic income. There is an inherent trade-off in evidence-based public decision making in that some proportion of the population need to take a health, privacy, or other risk to support societal goals. Further research should investigate how improving predictive capabilities can be responsibly leveraged across government and private enterprises.

As tasks become more complex and data intensive algorithms will continue to be leveraged for decision making. Already, algorithms are used for difficult tasks such as medical diagnoses (Gruber 2019), bail decisions (Arnold et al. 2018), stock picking (Zuckerman 2019), and determining the veracity of content on social media(Field and Lapowsky 2020). The findings reveal a reliance on algorithms for difficult tasks and it is important for decision-makers to be

vigilant in how they incorporate algorithmic advice, particularly because they are likely predisposed towards leaning on it for difficult, thorny problems. While algorithms can generally be very accurate, there are instances of algorithms quietly making sexist hiring decisions in one of the largest companies in the United States (Dastin 2018), initiating plane crashes (MacGillis 2019), or causing racist bail decisions (Stevenson 2017). Consequently, individuals and organizations leveraging big data to make decisions must be cognizant of the potential biases that accompany algorithmic recommendations, particularly for difficult problems. Decision makers should remember that they are likely to rely more on algorithms for harder questions, which may lead to flawed, biased, or inaccurate results. Accordingly, extreme algorithmic appreciation can lead to not only complacency, but also ineffective policies, poor business decisions, or propagation of biases.

2.4 Methods

This study was approved by the University of Georgia Institutional Review Board, project 00001012. Subjects gave written informed consent both before and after participation in the study. All methods were carried out in accordance with relevant guidelines and regulations. We conducted three preregistered experiments to test the conditions under which humans accept advice. Following a Judge Advisor System approach (Sniezek and Buckley 1995), subjects were asked to answer a question, then were exposed to advice, and then asked to submit a second answer. The preregistrations can be found for experiment 1, experiment 2, and experiment 3 at the Open Science Foundation.

2.4.1 Subjects

For experiment 1, we conducted a power analysis that indicated we needed 235 subjects per group. With two groups that is 470 subjects. We used a t-test for evaluating the difference

between two independent means using the statistical software G Power (Faul et al. 2007). We conducted a two tailed test, with an effect size of 0.3, an error probability of 0.05, power of 0.90, and an allocation ratio of 1. Subjects were recruited from Amazon Mechanical Turk (AMT). We started with 611 respondents recruited from AMT. Of those, 16 were duplicate IP addresses, 27 failed the attention check, 3 did not consent to their data being used, and 21 failed the manipulation check. Lastly, we excluded 14 subjects who had no deviation in their weight on advice, i.e. subjects who always either took the advice perfectly or who always completely ignored the advice. The analysis is based on the 530 remaining subjects, compared with the preregistered plan of 470 subjects. We oversampled because we did not know a priori how many subjects would be excluded. As part of our robustness checks we removed subjects based on time spent on a problem and confidence. The findings did not meaningfully change. Each subject was paid USD 1.50 to complete the experiment, and an additional bonus of USD 0.50 was given to subjects in the top 20% of accuracy in their final answers. Subjects were aware that a bonus was available for the most accurate respondents, but were not told the exact amount of the bonus, following prior usage of bonuses in online experiments (Guilbeault et al. 2018).

For experiments two and three we followed a similar approach, again recruiting subjects from Amazon Mechanical Turk. Subjects who participated in one of the experiments were not allowed to participate in a subsequent experiment, because we wanted to obtain as large a cross-section of the population as possible, and because we informed subjects of the experimental manipulation after the experiment was completed. We review the details of how we excluded subjects for experiments two and three in the Supplementary Information.

2.4.2 Task

All subjects saw ten images with crowds of between 15 (for the easiest question) and 5,000 (for the hardest question) humans. Easier questions were either the bottom left or bottom right quadrant of a harder image and were zoomed in so that each picture was the same size. The pictures were from an annotated dataset with professional assessments of the number of people in a picture (Idrees, Tayyab, et al. 2013). For each picture, a subject submitted an initial guess, along with their confidence. Subjects were then given advice and asked to resubmit an estimate along with their new level of confidence. Each subject saw ten pictures, five easy and five hard, which vary by the number of people pictured. The difficulty manipulation was within-subjects – all respondents saw the same questions. The type of advice was between subjects. Each subject was placed in one of two groups – one received advice described as "an algorithm trained on 5,000 images" and one received advice described as "the average of 5,000 other people". To control for advice quality, which is known to affect advice discounting (Yaniv and Kleinberger 2000), the advice was always the correct number of people in an image, as reported in the image database. We later manipulate advice quality in experiment three.

Subjects were reminded of their prior answer when answering the question the second time. Subjects answered how confident they were in both the initial and subsequent guess. Easier questions are subsets of harder questions – for each picture the easier version of the question was always the bottom left or bottom right quadrant of the harder picture. We bolded the source of the advice, which was described as either "an algorithm trained on 5,000 images similar to this one" or "the average guess 5,000 other people". In experiment 1 the source of the advice was between subjects, and thus never changed for a subject. In experiment two we relaxed this assumption and showed advice as within-subjects. Question order was randomized, so that

subjects could see easy or hard questions in any order, but subjects always saw the Post question directly after the Initial Question.

2.4.3 Analytical Approach

We used multilevel mixed-effects linear regression with random intercepts – fit using the lme4 package, version 1.1.23, in the R computing environment (Bates et al. 2014) – to analyze the effects of the advice type and task difficulty on weight on advice, time, and confidence. WOA as dependent variables, we control for both the initial confidence in an estimate prior to seeing advice, and for accuracy prior to advice. Our main model is:

$$y_{ik} = \beta_{0i} + \beta_1 \text{AlgoCondition}_i + \beta_2 \text{Difficulty}_k + \beta_3 \text{AlgoCondition}_i \times \text{Difficulty}_k + \beta X_{ik} + \varepsilon_{ik}$$

Here, y_{ik} is one of the dependent variables for participant i and problem k; β_{0i} is the slope for participant i; AlgoCondition_i and Difficulty_k are categorical variables indicating the advice condition and problem difficulty respectively; and X_{ik} is a vector of control variables. For experiment three we added categorical variables for quality of advice and the interaction between quality of advice and algorithmic advice. We also added variables for accuracy of an estimate prior to advice being given, and the interaction of accuracy with algorithmic advice. We tested for the appropriateness of using a linear mixed effects model by plotting the standardized residuals against the standard normal distribution, see Fig. S3 in the Supplemental Information.

2.4.4 Dependent Variable

Weight on Advice (WOA): The formula for WOA is $WOA_{ik} = \frac{|final\ estimate_{ik} - initial\ estimate_{ik}|}{|recommendation_k - initial\ estimate_{ik}|}$ .A WOA of one means an individual changed their answer to equal the advice given. A WOA of zero means an individual did not change their answer at all after receiving advice, and a WOA of

0.5 means an individual took the average of the advice given and their initial answer. According to recommended practices, we drop observations where the initial estimate is equal to the recommendation. We excluded observations where WOA was greater than two and less than negative one (Gino and Moore 2007).

2.4.5 Independent Variables

Difficulty: Categorical variable representing whether an image was easy or hard. Hard images coded as 1.

Algorithmic Advice: Categorical variable representing whether a subject received algorithmic advice for that question. Algorithmic advice coded as 1.

Accuracy: To control for skill in estimating crowd size, we calculate a subject's relative question-level accuracy as follows:

$$Error_{ik} = |Initial\ Answer_{ik} - Correct\ Answer_k| / Correct\ Answer_k.$$

To improve interpretability, we take the inverse of a subject's error: $Accuracy_{ik} = Error_{ik}^{-1}$. Thus, subjects who are more accurate had lower error estimates. To control for outliers, we then transform each subject's accuracy into the percentile rank for that question.

Advice Quality: Dummy variable representing whether advice was accurate or inaccurate. Inaccurate advice was 100% too high. Accurate advice coded as 1.

2.4.6 Control Variables

Initial Confidence: a subject's response to the question "How confident are you that your answer is within 10% of the true answer?" prior to receiving advice. 1 = Not at all confident, 2 = Not very confident. 3 = Somewhat confident. 4 = Extremely confident.

*Round Number***:** Because questions were in a random order, this variable described how many questions a subject had worked on thus far. Ranges from one to ten.

Numeracy: A measure to determine how well a subject understands fractions, decimals, and other numbers, previously used to establish numeracy in assessments of algorithmic advice taking (Logg et al. 2019) and medical decisions (Schwartz et al. 1997). Ranges from one to eleven.

Data availability

The datasets generated and analyzed during the current study are available in the Open Science Foundation repository: experiment 1, experiment 2, and experiment 3.

ACKNOWLEDGMENTS

This work was supported in part by Army Research Office grant W911NF1910427.

Author contributions

E.B., A.S., R.W., designed the research. E.B. and A.S. analyzed the data. E.B., A.S., and R.W. prepared the manuscript and reviewed and edited the manuscript.

Competing Interests

The authors declare no competing interests.

2.5 Supplementary Information

Humans Rely More on Algorithms than Social Influence as a Task Becomes More Difficult

Eric Bogert[1,*], Aaron Schecter[1], Rick Watson[1]

[1]Management Information Systems Department, University of Georgia, Athens, GA, USA

***Correspondence to etbogert@uga.edu**

Table S1. Summary Statistics of Experiment 1

Variable	Experiment 1		Experiment 2		Experiment 3	
	Mean	SD	Mean	SD	Mean	SD
WOA	0.487	0.370	0.525	0.367	0.451	0.381
Change in Confidence	0.178	0.440	0.192	0.453	0.116	0.380
Change in Time	-0.146	1.212	0.068	2.720	0.063	10.04
Initial Accuracy	0.500	0.287	0.501	0.288	0.500	0.287
Initial Confidence	2.438	0.904	2.637	0.823	2.696	0.826

Table S2. Model Results predicting Weight on Advice (Experiment 1: Algorithmic Advice Between-Subjects).

Variable	Model 1	Model 2	Model 3
Intercept	0.455 ***	0.466 ***	0.683 ***
	(0.016)	(0.018)	(0.026)
Algorithmic Advice	0.114 ***	0.094 ***	0.108 ***
	(0.018)	(0.026)	(0.025)
Difficulty	0.150 ***	0.132 ***	0.081 ***
	(0.008)	(0.012)	(0.012)
Initial Accuracy	-0.203 ***	-0.205 ***	-0.202 ***
	(0.016)	(0.022)	(0.022)
Algorithmic Advice * Difficulty		0.037 *	0.036 *
		(0.017)	(0.017)
Algorithmic Advice * Initial Accuracy		0.002	-0.007
		(0.031)	(0.031)
Initial Confidence			0.003 *
			(0.001)
Round Number			-0.088 ***
			(0.006)
Observations	5083	5083	5083
AIC	3043.487	3054.017	2892.011

The dependent variable is Weight on Advice. Initial Accuracy is the percentile rank score (0.00 to 1.0) of how accurate a subject's first guess was, relative to other subjects for that question. Initial confidence is on a Likert scale from 1-4. Algorithmic Advice is a categorical variable (algorithmic condition = 1, human advice = 0) and Difficulty is an ordinal variable (hard questions = 1, easy questions = 0). Standard errors are in parentheses. There were N = 530 subjects in Experiment 1.

**** p < 0.001; **p<0.01; *p<0.05.*

Table S3. Model Results Predicting Weight on Advice (Experiment 2: Algorithmic Advice Within-Subjects).

Variable	Model 1	Model 2	Model 3
Intercept	0.444 ***	0.465 ***	0.730 ***
	(0.014)	(0.017)	(0.054)
Algorithmic Advice	0.069 ***	0.027	0.027
	(0.009)	(0.020)	(0.020)
Difficulty	0.139 ***	0.120 ***	0.089 ***
	(0.009)	(0.013)	(0.014)
Initial Accuracy	-0.050 **	-0.072 **	-0.069 **
	(0.016)	(0.023)	(0.023)
Algorithmic Advice * Difficulty		0.039 *	0.038 *
		(0.018)	(0.018)
Algorithmic Advice * Initial Accuracy		0.045	0.045
		(0.032)	(0.031)
Initial Confidence			-0.043 ***
			(0.008)
Numeracy			-0.014 **
			(0.005)
Round Number			-0.003
			(0.002)
Observations	4905	4905	4905
AIC	3110.667	3119.503	3111.134

The dependent variable is Weight on Advice. Initial Accuracy is the percentile rank score (0.00 to 1.0) of how accurate a subject's first guess was, relative to other subjects for that question. Initial confidence is on a Likert scale from 1-4. Algorithmic Advice is a categorical variable (algorithmic condition = 1, human advice = 0) and Difficulty is a categorical variable (hard questions = 1, easy questions = 0). Standard errors are in parentheses. There were N = 514 subjects in Experiment 2.

**** $P < 0.001$; ** $P < 0.01$; * $P < 0.05$*

Table S4. Model Results predicting Weight on Advice (Experiment 3: Including Low-Quality Advice).

Variable	Model 1	Model 2	Model 3
Intercept	0.329 ***	0.609 ***	0.865 ***
	(0.019)	(0.034)	(0.059)
Algorithmic Advice	0.135 ***	0.092 **	0.059 *
	(0.023)	(0.030)	(0.030)
Difficulty	0.160 ***	0.098 ***	0.167 ***
	(0.009)	(0.013)	(0.016)
Initial Accuracy	-0.052 **	-0.080 ***	-0.080 ***
	(0.017)	(0.023)	(0.023)
Initial Confidence		-0.071 ***	-0.074 ***
		(0.008)	(0.008)
Round Number		-0.008 ***	-0.008 ***
		(0.002)	(0.001)
Algorithmic Advice * Difficulty		0.035 *	0.037 *
		(0.017)	(0.017)
Algorithmic Advice * Initial Accuracy		0.058	0.063
		(0.033)	(0.033)
Quality			0.114 ***
			(0.015)
Numeracy			-0.036 ***
			(0.005)
Algorithmic Advice * Quality			0.035 *
			(0.017)
Difficulty * Advice Quality			-0.146 ***
			(0.018)
Observations	4365	4365	4365
AIC	2459.132	2390.082	2272.17

The dependent variable is Weight on Advice. Initial Accuracy is the percentile rank score (0.00 to 1.0) of how accurate a subject's first guess was, relative to other subjects for that question. Initial confidence is on a Likert scale from 1-4. Algorithmic Advice is a categorical variable (algorithmic condition = 1, human advice = 0) and Difficulty is a categorical variable (hard questions = 1, easy questions = 0). Standard errors are in parentheses. There were N = 456 subjects in Experiment 3.

*** $P < 0.001$; ** $P < 0.01$; * $P < 0.05$

2.5.1 Preregistered Hypotheses and Analyses with Alternate Dependent Variables

In addition to studying how people responded behaviorally by their weight on advice, we preregistered hypotheses related to how people responded cognitively, both through the time they spent on a question and their self-reported confidence in their answers. This allowed us to study both behavioral (weight on advice) and cognitive (confidence and time spent on a problem) manifestations of reliance on machine intelligence relative to social influence. We chose to report these results on alternate dependent variables in the supplementary information to increase the readability and decrease the length of the main paper.

We hypothesized that subjects would demonstrate greater reliance on machine intelligence than social influence, and most importantly, that effect would be stronger in more difficult tasks. For time-related measurements this would manifest through subjects spending less time when they are advised by an algorithm, and that this effect would be stronger in more difficult tasks. Using time as a measure of cognitive dependence is a previously established measure of cognitive effort (Alexander et al. 2018). For confidence-related measurements this would result in subjects becoming more confident, based on a measure using a Likert scale, when receiving algorithmic advice, and this effect being stronger in more difficult tasks.

For experiment 1 and 2, our hypotheses were as follows.

Table S5. Preregistered Hypotheses for Experiment 1 and Experiment 2

#	Hypothesis
H1a	The effect of algorithmic advice on weight on advice will be greater than the advice of a crowd.
H1b	The effect of algorithmic advice on weight on advice will be stronger for a more difficult problem.
H2a	Algorithmic advice will result in less time spent determining an answer relative to advice from a crowd.
H2b	The effect of algorithmic advice on time, relative to advice from a crowd, will be stronger for a more difficult problem.
H3a	Algorithmic advice makes humans more confident in decisions than the advice of crowds.
H3b	The effect of algorithmic advice on decision confidence will be stronger for a more difficult problem.

In experiment three we introduced low-quality advice, and added several hypotheses related to those effects. Those additional hypotheses are listed below. Because we preregistered these hypotheses we include them here, because it is a best practice of preregistered research to include all hypotheses from a preregistration (Simmons et al. 2011). We believe the most interesting hypothesized effects are for Hypothesis 4 and Hypothesis 5b because those effects are related to algorithmic advice.

Table S6. Additional Preregistered Hypotheses for Experiment 3

#	Hypothesis
H4	Subjects who are more skilled at a task will rely more strongly on algorithmic advice than advice from a crowd.
H5a	Low quality advice reduces future reliance on the advice source
H5b	Low quality advice will more strongly reduce reliance on algorithmic advice than reliance on crowd advice
H5c	Low quality advice will more strongly reduce reliance on advice for easy questions than hard questions.

These hypotheses directly measured effects related to machine intelligence and social influence, whereas Hypothesis 5a and Hypothesis 5c did not hypothesize about differences between machine intelligence and social influence. Thus, Hypothesis 5a and 5c are more applicable to general behavior and decision-making literature, whereas Hypothesis 4 and 5b are directly applicable to the burgeoning literature on algorithmic appreciation.

2.5.2 Results

2.5.2.1 Experiment 1:

Increase in confidence is higher when receiving advice from an algorithm and increases with task difficulty

We found that algorithmic advice increased confidence (B = 0.043; P = 0.046; 95% Confidence Interval (CI) = 0.002 to 0.083), supporting H3a and in line with the effect we observed on weight on advice. Furthermore, the interaction between algorithmic advice and difficulty was significant and positive (B = 0.053; P = 0.003; CI = 0.022 to 0.093), supporting H3b and in line with the

effect we observed on weight on advice. When individuals receive advice from an algorithm rather than a crowd for a hard task, they are 5.3% more confident in their final answer compared to their initial guess. Overall, the results suggest that when subjects receive advice from an algorithm rather than a purported crowd of other humans, they become more confident in their answers, and this effect is stronger as tasks become more difficult. Contrary to our hypotheses, these results were not observed when we studied time as a dependent variable, indicating no support for H2a and H2b. Results are in Table S7 below.

Table S7: Experiment 1 Analyses on Alternative DVs

	Change in Confidence DV			**Change in Time DV**		
Variable	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
Intercept	-0.035	0.001	0.777 ***	-0.097 *	-0.097	0.023
	(0.018)	(0.021)	(0.029)	(0.043)	(0.055)	(0.085)
Algorithmic Advice	0.043 *	-0.031	0.016	0.048	0.049	0.047
	(0.021)	(0.030)	(0.027)	(0.038)	(0.078)	(0.078)
Difficulty	0.092 ***	0.060 ***	-0.103 ***	0.263 ***	0.238 ***	0.236 ***
	(0.011)	(0.015)	(0.014)	(0.033)	(0.047)	(0.048)
Initial Accuracy	0.287 ***	0.246 ***	0.253 ***	-0.413 ***	-0.387 ***	-0.389 ***
	(0.020)	(0.028)	(0.024)	(0.059)	(0.083)	(0.083)
Algorithmic Advice * Difficulty		0.064 **	0.058 **		0.049	0.047
		(0.021)	(0.018)		(0.066)	(0.066)
Algorithmic Advice * Initial Accuracy		0.081 *	0.053		-0.053	-0.047
		(0.039)	(0.034)		(0.117)	(0.117)
Round Number			-0.002			-0.018 **
			(0.002)			(0.006)
Initial Confidence			-0.286 ***			-0.007
			(0.007)			(0.021)
Observations	5083	5083	5083	5083	5083	5083
AIC	5255.830	5257.079	3880.841	16279.534	16288.811	16296.853

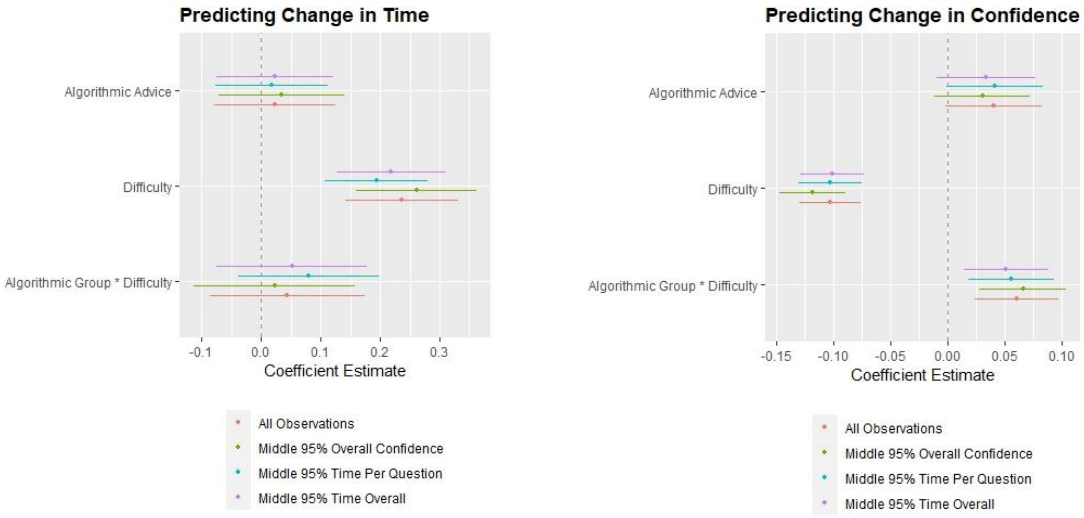
*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

The dependent variables are the percentile change in confidence (model 1, 2, and 3) and the percentile change in time spent on a problem (model 4, 5, and 6). Initial Accuracy is the percentile rank score (0.00 to 1.0) of how accurate a subject's first guess was, relative to other subjects for that question. Initial confidence is on a Likert scale from 1-4. Algorithmic Advice is a dummy variable (algorithmic condition = 1, human advice = 0) and Difficulty is a dummy variable (hard questions = 1, easy questions = 0). Standard errors are in parentheses. There were N = 530 subjects in Experiment 1.

2.5.2.2 Experiment 1: Robustness Checks

Finally, we conducted several robustness checks on our main models. See Figure S1A for change in time spent on a problem and S1B for change in confidence. We removed subsets of our data to ensure extreme values were not adversely impacting our findings. The subgroups exclude the

following people: the top 5% and bottom 5% and both top and bottom 2.5% in accuracy, and initial time spent on a question, and the top and bottom 2.5% in confidence on initial questions. Across all alternative regressions our findings were consistent. To check for multicollinearity, we removed control variables step-wise. Removing accuracy, initial confidence, and both accuracy and initial confidence did not change our results – all supported hypotheses remained supported at the 0.05 level, and all unsupported hypotheses remained unsupported.



(A) (B)

Figure S1. Robustness checks for (A) change in time and (B) change in confidence across a variety of subsets.

We ran identical exclusions for WOA, our dependent variable from the main text. We removed subsets of our subjects to ensure our results were robust to outliers. Figure S2 shows that our effects are robust to removing these subsets.

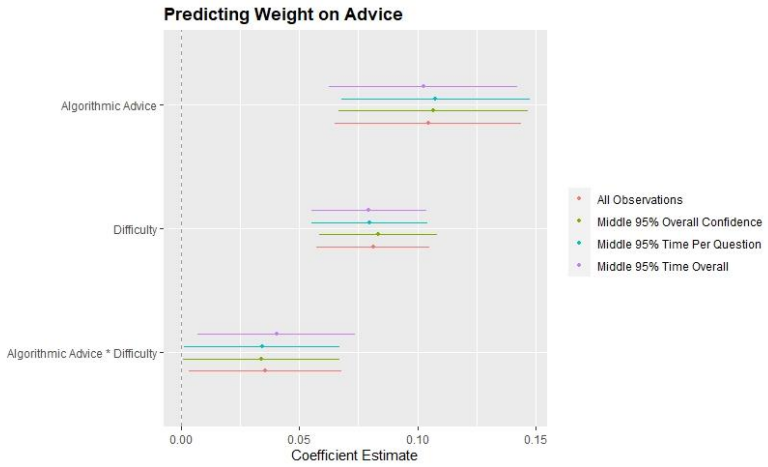


Figure S2. Robustness checks for Weight on Advice across subsets

2.5.2.3 Experiment 2

In experiment two, we followed the same procedure, but this time making it so that subjects received advice as a within-subjects condition, resulting in each subject receiving advice labeled as algorithmic five times and advice labeled as the average of other humans five times. We used the same hypotheses from experiment one, and observed similar effects, shown in Table S8 below. Again, we observe that the effect of algorithmic advice on a change in confidence is positive and significant ($B = 0.055$; $P < 0.001$; 95% CI = 0.034 to 0.077), but that effect disappears after incorporating the interaction between algorithmic advice and difficulty, indicating partial support for H3a and support for H3b. We also observe consistent effects related to time – subjects did not statistically significantly change the amount of time spent on a problem, regardless of advice source and whether we include interactions or other controls, indicating no support for H2a and H2b.

Table S8: Experiment 2 Analyses on Alternative DVs

	Change in Confidence DV			**Change in Time DV**		
Variable	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
Intercept	-0.071 ***	-0.053 **	1.084 ***	0.363 ***	0.374 **	0.717 ***
	(0.017)	(0.020)	(0.030)	(0.096)	(0.123)	(0.217)

Algorithmic Advice	0.055 ***	0.018	0.021	-0.114	-0.135	-0.137
	(0.011)	(0.026)	(0.021)	(0.077)	(0.175)	(0.175)
Difficulty	0.170 ***	0.135 ***	-0.132 ***	0.136	0.032	0.020
	(0.011)	(0.016)	(0.014)	(0.077)	(0.109)	(0.115)
Initial Accuracy	0.294 ***	0.294 ***	0.282 ***	-0.615 ***	-0.528 **	-0.515 **
	(0.021)	(0.029)	(0.024)	(0.135)	(0.191)	(0.191)
Algorithmic Advice * Difficulty		0.070 **	0.074 ***		0.207	0.202
		(0.023)	(0.019)		(0.155)	(0.155)
Algorithmic Advice * Initial Accuracy		0.001	0.003		-0.172	-0.166
		(0.040)	(0.032)		(0.269)	(0.269)
Round Number			0.000			-0.059 ***
			(0.002)			(0.013)
Initial Confidence			-0.378 ***			-0.007
			(0.008)			(0.054)
Observations	4905	4905	4905	4905	4905	4905
AIC	5247.113	5252.188	3385.244	23726.639	23731.135	23726.567

*** $p < 0.001$; **$p<0.01$; *$p<0.05$.

The dependent variable is the percentile change in confidence (model 1, 2, and 3) and the percentile change in time spent on a problem (model 4, 5, and 6). Initial Accuracy is the percentile rank score (0.00 to 1.0) of how accurate a subject's first guess was, relative to other subjects for that question. Initial confidence is on a Likert scale from 1-4. Algorithmic Advice is a categorical variable (algorithmic condition = 1, human advice = 0) and Difficulty is a categorical variable (hard questions = 1, easy questions = 0). Standard errors are in parentheses. There were N = 514 subjects in Experiment 2.

2.5.2.4 Experiment 3

In experiment three we introduced low-quality advice and returned advice source to a between-subjects condition. We found that the effects related to confidence reinforced the results related to weight on advice. Subjects became more confident after receiving algorithmic advice relative to crowd advice, although this effect lost significance after controlling for advice quality (B = 0.020; P = 0.718; 95% CI = -0.038 to 0.069), demonstrating partial support for H1. Most importantly, subjects became more confident in difficult questions when receiving algorithmic advice, relative to advice of equal quality from a crowd, demonstrating support for hypothesis 2 (B = 0.067; P < .001; 95% CI = 0.032 to 0.102). Highly accurate subjects did not become statistically significantly more confident when receiving algorithmic advice relative to human advice (B = 0.066; P = 0.050; 95% CI = -0.00005 to 0.13185), although this effect is significant at a p value of 0.1. This indicates that we cannot reject the null hypothesis for hypothesis 4. Subjects became more confident when receiving high quality advice (B = 0.065; p < 0.001; 95% CI = 0.034 to 0.097), supporting hypothesis 5a. Unlike our results for weight on advice, subjects

did not become significantly less confident when receiving low-quality algorithmic advice relative to low-quality crowd advice, as hypothesized in h5b (B = 0.014; p = 0.410; 95% CI = -0.020 to 0.050). Subjects also became significantly less confident when receiving low-quality advice for hard questions relative to easy questions, supporting hypothesis 5c (B = 0.049; P = 0.008; 95% CI = -0.085 to -0.013), demonstrating support for hypothesis 5c.

No hypotheses related to changes in time spent on a question were supported, regardless of whether we include only main effects used in our prior experiments (model 1), main effects plus the interaction between difficulty and algorithmic advice (model 2), or all interactions related to hypotheses plus main effects and controls (model 3).

Table S9: Experiment 3 Analyses on Alternative DVs

	Change in Confidence DV			**Change in Time DV**		
Variable	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
Intercept	-0.028	-0.012	1.019 ***	0.006	0.244	1.353
	(0.017)	(0.018)	(0.051)	(0.378)	-0.415	(1.171)
Algorithmic Advice	0.077 ***	0.048 *	0.010	0.418	-0.020	0.224
	(0.017)	(0.020)	(0.027)	(0.304)	-0.436	(0.750)
Difficulty	0.113 ***	0.084 ***	-0.073 ***	0.633 *	0.202	-0.202
	(0.010)	(0.015)	(0.017)	(0.304)	-0.433	(0.545)
Accuracy Rank	0.092 ***	0.090 ***	0.067 **	-0.961	-0.994	0.063
	(0.019)	(0.019)	(0.024)	(0.529)	-0.53	(0.752)
Algorithmic Advice * Difficulty		0.057 **	0.067 ***		0.850	0.829
		(0.021)	(0.018)		-0.609	(0.608)
Advice Quality			0.065 ***			-0.517
			(0.016)			(0.533)
Initial Confidence			-0.288 ***			0.130
			(0.008)			(0.204)
Numeracy			-0.021 ***			-0.230 **
			(0.004)			(0.079)
Round Number			-0.001			0.025
			(0.002)			(0.053)
Algorithmic Advice * Accuracy Rank			0.066			-1.462
			(0.034)			(1.059)
Algorithmic Advice * Advice Quality			0.015			0.877
			(0.018)			(0.607)
Difficulty * Quality		-	0.049 **			1.023
			(0.019)			(0.608)
N	4365	4365	4365	4365	4365	4365
AIC	3541.062	3541.574	2406.201	32532.609	32531.812	32532.145

*** $p < 0.001$; **$p<0.01$; *$p<0.05$.

The dependent variable is the percentile change in confidence (model 1, 2, and 3) and the percentile change in time spent on a problem (model 4, 5, and 6). Initial Accuracy is the percentile rank score (0.00 to 1.0) of how accurate a subject's first guess was, relative to other subjects for that question. Initial confidence is on a Likert scale from 1-4. Algorithmic Advice is a categorical variable (algorithmic condition = 1, human advice = 0) and Difficulty is a categorical variable (hard questions = 1, easy questions = 0). Standard errors are in parentheses. There were N = 456 subjects in Experiment 3.

2.5.3 Synthesized Results

In Table S10 we summarize results across each experiment and dependent variable. We combine H1A, H2A, and H3A into a single hypothesis, listed first in Table S10 below and combine H1B, H2B, and H3B into a single hypothesis, listed second in Table S10 below. We say a hypothesis was supported if, across all three experiments and all specifications of the experiments, we observed the relevant effect. We indicate partial support if, in at least one of the experiments, there was support for the hypothesis.

Table S10: Analyses on All Dependent Variables Across All Experiments

Hypothesis	Support (WOA)	Support (Time)	Support (Confidence)
Algorithmic advice will result in greater reliance on the advice than advice from a crowd	Yes	No	Partial
Algorithmic advice will be relied on more than advice from a crowd as task difficulty increases	Yes	No	Yes
Subjects who are more skilled at a task will rely more strongly on algorithmic advice than advice from a crowd.	No	No	No
Low quality advice reduces future reliance on the advice source*	Yes	No	Yes
Low quality advice will more strongly reduce reliance on algorithmic advice than reliance on crowd advice*	Yes	No	No
Low quality advice will more strongly reduce reliance on advice for easy questions than hard questions*	Yes	No	Yes

*Tested on experiment three exclusively

Overall, we observed that subjects became more confident when receiving algorithmic advice than when receiving crowd advice for difficult questions. This result provides further evidence that 1) the manipulation was salient, and 2) that participants relied meaningfully on the algorithmic advice to improve their answers. Clearly, people not only followed algorithmic

advice, but felt better about themselves when they did so. Our null results related to time could be accounted for through several explanations. The most plausible is that time spent on a problem is not a manifestation of cognitive effort in the same way as weight on advice or scales related to confidence. Alternatively, this might be attributed to the fact that we recruited subjects from Amazon Mechanical Turk, resulting in our subjects being highly incentivized to work quickly, because they are paid per task rather than per unit of time.

2.5.4 Methods

2.5.4.1 Subjects:

We reviewed how we excluded subjects from experiment one in the main text of this article. In experiment two we used the same procedure for recruiting and excluding participants as we did in experiment one, with the additional criterion that we did not allow subjects from experiment one to participate. We started with data from 593 subjects. We oversampled slightly, relative to our stated sample size in the preregistration, because we did not know in advance how many subjects would fail our exclusion criteria. All subjects consented to their data being used. Nine failed the attention check. We removed three subjects because all of their weights on advice were either above two or below negative two, following prior literature and our preregistration (Logg et al. 2019). We removed 67 subjects because they exclusively put either no weight on advice or exactly the advice for every question. Multiple subjects emailed us explaining that they thought the instructions, which said to "note the source of the advice", meant that they were supposed to simply write the advice itself rather than their best estimate. We did not use a manipulation check for this experiment, because subjects were exposed to both sources of advice. We ran our models on the remaining 514 subjects.

In experiment three we used the same procedure for recruiting and excluding participants as we did in experiment one and two. We did not allow subjects who completed either experiment one

or experiment two to participate in experiment three. We started with 673 responses. 100 of those respondents claimed they were in the army, and thus we did not allow them to complete the experiment (due to limitations imposed by our funding). Of those 573, four did not consent to our use of their data. 73 subjects failed the attention check. 39 subjects either exclusively took the advice without consideration for their initial estimate (WOA always equal to one) or exclusively disregarded the advice completely (WOA always equal to 0). We ran our models on the remaining 461 subjects.

2.5.4.2 Analytical Approach

As with our analyses on weight on advice, we again used multilevel mixed-effects linear regression with random intercepts – fit using the lme4 package in the R computing environment (Bates et al. 2014) – to analyze the effects of the advice type and task difficulty on weight on advice, time, and confidence. For the models with the change in time spent as the dependent variables, we control for both the initial confidence in an estimate prior to seeing advice, and for accuracy prior to advice. For the model with the change in confidence as the dependent variable, we do not control for initial confidence so as to not include initial confidence on both sides of our structural equation. Our main model:

$$y_{ik} = \beta_{0i} + \beta_1 \text{AlgoCondition}_i + \beta_2 \text{Difficulty}_k + \beta_3 \text{AlgoCondition}_i \times \text{Difficulty}_k + \beta X_{ik} + \varepsilon_{ik}$$

Here, y_{ik} is one of the dependent variables for participant i and problem k; β_{0i} is the slope for participant i; AlgoCondition_i and Difficulty_k are dummy variables indicating the advice condition and problem difficulty respectively; and X_{ik} is a vector of control variables. As with the main analysis, we also included terms for advice quality and the corresponding interactions for experiment three.

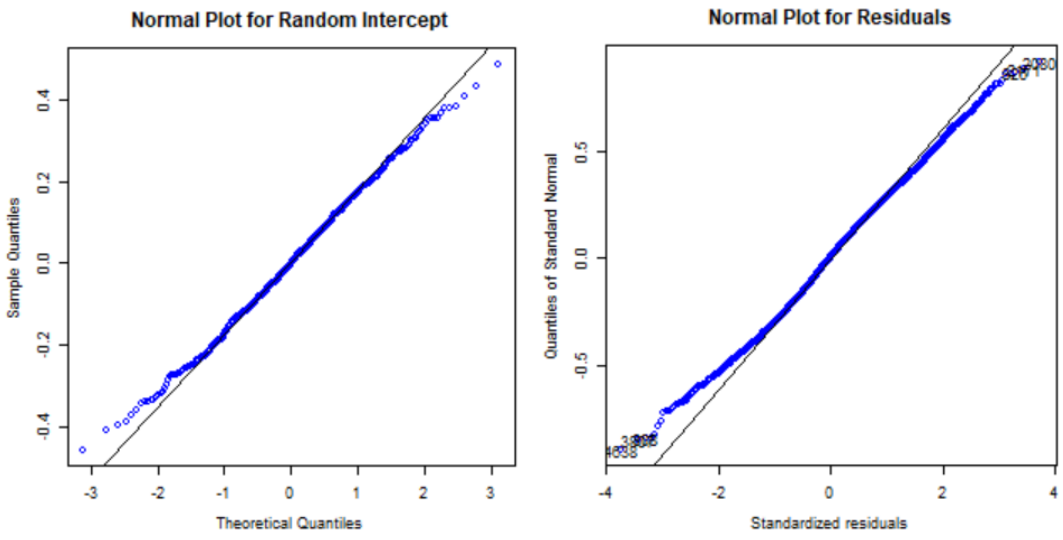
2.5.4.3 *Variables Unique to the Supplemental Analysis:*

Time Change: The percentage change in the amount of time taken to submit the page, between the first answer and second answer. It includes both the time taken to answer the question about crowd size and time taken to submit confidence. A negative value indicates the subject spent less time on their answer after receiving advice than on their answer prior to receiving advice.

Confidence Change: the percentage change in confidence between and initial answer and an answer after advice was received. A negative value indicates a subject became less confident after receiving advice.

2.5.4.4 *Normality Analysis*

We assessed the appropriateness of using linear mixed-effects models on our Weight on Advice variable by plotting the theoretical quantiles of the random intercept against the sample quantiles (Figure S3a below) and the standardized residuals against the quantiles of the standard normal (Figure S3b below). These plots indicate that a linear mixed effects model is an appropriate tool to analyze our data.



(A) *(B)*

Figure S3. Plotting theoretical quantiles against sample quantiles (A) and standardized residuals against the normal distribution (B)

2.5.5 Consent and Debrief Materials

2.5.5.1 Consent Form

Dear Participant,

My name is Aaron Schecter and I am a faculty member in the Management Information Systems Department at the University of Georgia. I am inviting you to take part in a research study. I am doing research on how individuals work in teams to complete problem-solving tasks. Specifically, I am investigating how different technologies such as artificial intelligence or virtual communication can affect teamwork. Your responses may help us understand how teams can function effectively while using these advanced technologies.

I am looking for individuals age 18 and older who are residents of the United States and not current military personnel.

If you agree to take part in this study, you will be asked to view an image and assess how many elements are contained in that image. As part of the decision-making process, you will be shown an image twice. The first time you see it you will not have any advice. Your guesses may be visible to future participants. In order to make this study a valid one, some information about the study will be withheld until completion of the task. You will also be asked to complete a set of survey questions regarding your experience. This task should take approximately ten minutes, including time to answer survey questions.

Participation is voluntary. You can refuse to take part or stop at any time without penalty. Your decision to participate will have no impact in your participation in future studies. You may find the task difficult or frustrating. There is no penalty for submitting an incorrect answer or not completing the task accurately. If any survey questions make you uncomfortable, you can skip these questions if you do not wish to answer them.

Research records will be labeled with study IDs only. There will be no record of your name or other identifiable information. Because all responses are anonymous, the information may be used in future research studies or shared with other researchers without additional consent. This research involves the transmission of data over the Internet. Every reasonable effort has been taken to ensure the effective use of available technology; however, confidentiality during online communication cannot be guaranteed.

As compensation for your participation, you will receive money in your worker account. For completing the task, you will receive $1.50. **We will give more money, in the form of a bonus, to workers who are more accurate.**

This research is supported by the Army Research Office. Representatives of the Department of Defense are authorized to review research records.

If you are interested in participating or have questions about this research, please feel free to contact me at aschecter@uga.edu. If you have any complaints or questions about your rights as a research volunteer, contact the IRB at 706-542-3199 or by email at IRB@uga.edu.

Please keep this letter for your records.

Sincerely,

Aaron Schecter

Debrief Form

Thank you for your participation in this research study. For this study, it was important that we withhold information about some aspects of the study. Now that your participation is completed, we will describe the withheld information to you, why it was important, answer any of your questions, and provide you with the opportunity to make a decision on whether you would like to have your data included in this study.

What you should know about this study

During the study, you were able to view the guesses of other participants or the guess of an algorithm and factor them into your decisions. In reality, the only difference was the label, the advice was always the same. All participants, including yourself, viewed the same images, and no participant had an advantage over another. Not explaining that the advice was identical is an important component of this study. Humans typically interact with other people differently than they interact with technology, and they have different expectations about what each are capable of. By keeping the advice identical but labelling it differently, we can determine if your actions and perceptions are dependent on the label “algorithm". This research helps us better understand how beliefs regarding AI technology affects problem solving and teamwork.

Right to withdraw data

You may choose to withdraw the data you provided prior to debriefing, without penalty or loss of benefits to which you are otherwise entitled. Please check the box below if you do, or do not, give permission to have your data included in the study:

_____I give permission for the data collected from or about me to be included in the study.

_____I DO NOT give permission for the data collected from or about me to be included in the study.

Whether you agree or do not agree to have your data used for this study, you will still receive the $1.50 in your worker account.

Disclosure

Please do not disclose research procedures and/or purpose to anyone who might participate in this study in the future as this could affect the results of the study. This includes online forums, message boards, or social media sites.

If you have questions

The main researcher conducting this study is Aaron Schecter. If you have questions, you may contact the IRB at 706-542-3199. If you have any questions or concerns regarding your rights as a research participant in this study, you may contact the Institutional Review Board (IRB) Chairperson at irb@uga.edu.

Chapter 3
Algorithmic Recommendations for Creative Tasks

3.1 Introduction

In creative tasks, we expect algorithmic appreciation to disappear, and algorithmic aversion to manifest, because the nature of creative tasks is to come up with new ideas in tasks with an infinite number of possibilities. Algorithms excel when the possibilities are finite – even in complex games such as chess or poker there are limited moves, but in creative tasks such as brainstorming ideas or writing poetry there are practically infinite options at every turn, with no clear mechanism to evaluate those options computationally. Highly creative tasks, such as writing a coherent novel, are thus far out of reach for algorithms. At different levels of creative writing, such as the novel, chapter, and paragraph, there is no objective metric that dictates success. Even the rules of grammar and English are not immutable or untouchable. Sentences, a highly granular unit of storytelling, do not *have* to follow prescribed grammatical structures. The most highly acclaimed grammar books recommend following grammatical rules in almost all cases, but concede there are exceptions (Strunk and White 2005, p. 32). For example, in *Infinite Jest*, considered the magnum opus by David Foster Wallace, Wallace strings together multiple conjunctions to start sentences (e.g. "And, but, so") (Wallace 1996). This would be considered flawed by an algorithm but was considered brilliant by human reviewers.

The best AI-generated text uses a long-short-term-memory that allows it to "remember" prior phrases in a given corpus and thus avoid repetition (Sutskever et al. 2011). Despite this ability to remember at a sentence level, AI is currently unable to author a multi-page coherent plot. However, pop-culture projects using recurrent neural networks (RNNs) have created imitation

Kanye West songs (Alexander 2017) and additional chapters of the *Game of Thrones* series (Tewari 2019). These chapters follow grammatical rules but do not tell an intelligible story. See Appendix 1 for examples of both.

Early efforts at creativity, such as digital portraits, produced grotesque results (Kageki 2012). Recently, picture generation has improved dramatically – NVIDIA has released software that makes realistic human faces (Kerras et al. 2019; "Thispersondoesnotexist.Com" 2019). Despite this, no AI-generated paintings, novels, or poetry have won competitions against human artists. This is likely because algorithmic paintings, novels, and poetry are necessarily derivative, and creative art competitions reward ingenuity.

Text creation by AI is improving. Online reviews generated by RNNs have fooled humans into believing the review's author is human (Yao et al. 2017), but this is only possible because reviews are short and extremely similar to one another. Image labeling is also improving rapidly. Google and Microsoft have both recently debuted image labeling software (Kamps 2016; Torbet 2019). Google claims 93.9% accuracy in labeling images. Of course, these labels are hardly artful prose. In a picture of several cows grazing, Microsoft's AI outputs "I think it's a herd of cattle standing on top of a grass covered field." See Figure 3.1.

Figure 3.1: I think it's a herd of cattle standing on top of a grass covered field

3.2 Operationalization of Corollaries as Hypotheses

We operationalize the creative task as an image captioning problem. In each case, a hypothesis is an operationalization of its correspondingly numbered corollary. We list the hypotheses for the convenience of the reader.

Hypothesis 7: *For a writing problem {creative task}, task difficulty will negatively moderate the effect of algorithmic advice on belief change.*

Hypothesis 8: *For a writing problem {creative task}, task difficulty will positively moderate the effect of algorithmic advice on believe change.*

Hypothesis 9*: For a writing problem {creative task}, task difficulty will negatively moderate the effect of algorithmic advice on confidence.*

3.3 Experimental Design

We use a two-by-two between-and-within-subjects design (Table 3.1). The first condition is information source: algorithm or a crowd of humans. The second condition is difficulty. All subjects are told there is a prize for authoring better captions. Although this may cause the task to

lean more towards being an intellective task, subjects are told that there is a prize for better captions to increase the likelihood that subjects expend real effort while captioning.

Table 3.1: Experiment Two Design

Task difficulty	**Algorithm Description**	**Human Description**
Descriptive vs humorous	An algorithm, trained on 5,000 pictures similar to this one, recommends X.	A crowd, comprised of 5,000 people, recommends X.

It is likely that some subjects are better at captioning cartoons than others, because it requires a mastery of English. To measure each subject's knowledge of English and capability with wordplay, we use the Remote Associates Test (RAT) (Mednick 1968). The Test gives three words that are tied together by some unspecified word, which the test-taker is required to provide. For example, the three words "cottage", "swiss", and "cake" are tied together with the word "cheese". Each subject received 6 questions from a repository of the Remote Associates Test. Using the difficulty scores provided from an external website, https://www.remote-associates-test.com/, we give each subject the same six questions. One very easy question, one hard question, two easy questions, and two medium questions. We collect these data so that we can examine whether people who were skilled at wordplay rely differently on advice sources, although we did not make a formal hypothesis about this effect.

We also collect data on a subject's affect toward technology, through the Insecurity components of the Technology Readiness Index (TRI) 2.0 (Parasuraman and Colby 2015). The TRI measures feelings about technology, and capturing this data enables us to see whether self-reported affect toward technology affects reliance on technology.

3.3.1 Cartoon Choice

We used a publicly accessible dataset of 177 New Yorker cartoons that were used in the weekly New Yorker Cartoon Caption Contest. In that contest, people submit a caption to a cartoon and then vote on whether they believe other captions are funny. The top captions are shown a week

later. Our dataset contained a minimum of 2,919 captions per cartoon, and a maximum of 13,314 captions.

Our goal was to choose 5 cartoons based on how successfully a machine could predict that a given caption belonged to a cartoon. To test how well a machine matched a cartoon with a caption, we downloaded the 992,050 captions for the 177 cartoons in our sample. We removed any stopwords (e.g., "the", "a", "an") from the captions, and held out 30% of our data as a test set. We used Doc2Vec to create a 300 column vector space out of the training data, with a row for each caption. We then used three logistic regression, a neural network, and a random forest to predict which captions in our test set belonged to each cartoon. A naïve model assigning captions to cartoons at random would assign $\frac{1}{\#\ Cartoons}$ accurately. With 177 cartoons in our sample, this meant a naïve model would correctly assign a caption to a cartoon about 0.5% of the time.

We chose the five cartoons with the average highest F-Scores when using our three models, where the F score is: $2 * \frac{Precision * recall}{precision + recall}$. Using an F-Score as the criterion for selection balances false positives and false negatives. We then paid the New Yorker for licensing their cartoons in our experiment.

3.4 Operationalization

We manipulate difficulty by changing the type of image. Difficult images to caption are political cartoons because they need to be entertaining enough that the subjects might be more engaged than they would be with other material, and are currently beyond the abilities of AI to caption well. The goal of a political cartoon is usually humor, often in the form of irony or sarcasm. although not always. We will tell subjects that the goal is to be humorous. Political cartoon

caption contests are also an activity that laypeople can engage in for fun. They can garner hundreds of entrants, and captioning political cartoons is an elite profession (Donnelly 2019).

The easy creative tasks will be to caption simple pictures of landscapes. Landscapes are usually easy enough for even children to caption well. For example, most parents can attest to hearing their young children narrate the landscape as they go on a long drive (e.g., “Mom, it’s a herd of cows!”).

We measure confidence by asking subjects whether they believe they are in the top 10% of submissions. We measure cognitive effort using time elapsed to make a caption.

In addition to those baseline experimental procedures, we use a unique measure of belief change. In the dissertation proposal, we proposed several ways of measuring belief change, because there is no best way to measure change in text. In the proposal we stated that we may use other MTurk workers to assess how much a caption has changed or use an algorithm such as Doc2Vec to assess how much text has changed. However, in our design of the experiment we realized that reviewers may have pushed back against these measures of belief change, so we chose to use a four-option multiple choice answer with an optional free response, see Figure 3.2

Please provide a caption of this image. The most frequent caption for this cartoon, surveying 5,000 ***other people*** *was* ***"a stone bridge in a forest."***

You previously said "A bridge over a river"

Keep my first caption

Use the recommended caption

I want to tweak the recommended caption to improve my chances of receiving a bonus

I want to create a new caption to improve my chances of receiving a bonus

Figure 3.2. Example of Multiple Choice Questions Used to Determine Weight on Advice

3.5 Experimental Procedures

We follow the methodology outlined in sections 1.2, 1.3, and 1.4.

3.5.1 AMT Procedure

Subjects are instructed that they will be given images to caption. In the hard (easy) condition, subjects will be told that their captions should be funny (accurate). Once subjects submit a caption, they are taken to another screen with the same image. For the easy images of landscapes, subjects then see a recommended caption written by the first author. For the hard images, the caption that won The New Yorker caption cartoon contest for the associated cartoon was used. After viewing the advice, subjects are asked to write the best caption they can think of and are told that writers of the best captions will receive a bonus. Captions need to be at least three characters long, and each subject assesses 10 images.

3.6 Results

We deviated slightly from our preregistration, which indicated that we would gather results from at least 470 participants. After removing participants for failing the attention checks, manipulation check, and subjects who stated that they answered at random, we had 419 subjects in our sample. An analysis of the strength of our effects indicated that it was unlikely that any effect that was not significant would become significant if we gathered another 51 responses, so we did not gather additional responses. We conducted this analysis by calculating the what the p values for all effects would be if we had 470 subjects as the number of people in our sample while keeping the effect size the same. Using this analysis, no effects that were not significant would have become significant if we had gathered more responses.

Subjects were significantly less confident in their final answer ($\beta = -0.33$, $p < 0.001$) and initial answer ($\beta = -0.48$, $p < 0.001$), and took more time when writing their first caption ($\beta = 21.82$, $p <$

0.001), when they were captioning hard questions relative to easy questions. Thus, we conclude that subjects perceived the difficult and easy questions differently.

When analyzing belief change and confidence change, use a cumulative link mixed model (CLMM) from the ordinal package in R to fit each model. This is a deviation from the dissertation proposal, because in the dissertation proposal we had a different dependent variable in mind. Using a cumulative link mixed model is appropriate when the dependent variable is an ordered categorical variable. Our analytical approach was:

$$logit\big(P(Y_i \leq\ j)\big) = \theta_j - \beta_1(AdviceSource_i) - \beta_2(Difficulty_i) - \beta_3(AdviceSource_i * Difficulty_i) - \mu(Subject_i)$$

i = 1,…,n, j = 1,…, J – 1

This models the cumulative probability that caption *i* is in the answer *j* or below, where *i* index the observations and *j* index the possible answer choices (J = 4) (Christensen 2019). We assume the subject effects are random and that the subject effects are independently and identically distributed normally (Christensen 2019).

Across all models the effect of algorithmic advice is non-significant. While we did not formally hypothesize about this effect in the proposal, we believed we would see that subjects would change their beliefs more when receiving non-algorithmic advice. In a CLMM, a positive and statistically significant coefficient for algorithmic advice would indicate that subjects relied more on algorithmic advice than advice from a crowd. Across all models there is no significant effect of the interaction between algorithmic advice and difficulty, indicating no support for H7. These effects are displayed in Table 3.2.

3.6.1 Belief Change

Table 3.2. Weight on Advice Models

	Model 1	P Value	Model 2	P Value	Model 3	P Value
1|2	**0.994**	< 0.001	**0.825**	**< 0.001**	-0.757	0.398
	(0.097)		(0.168)		(0.896)	
2|3	**1.217**	< 0.001	**1.048**	**< 0.001**	-0.510	0.569
	(0.098)		(0.169)		(0.896)	
3|4	**1.409**	< 0.001	**1.240**	**< 0.001**	-0.296	0.741
	(0.099)		(0.169)		(0.895)	
Algorithmic Advice	0.101	0.454	0.087	0.523	0.518	0.117
	(0.136)		(0.136)		(0.331)	
Difficulty	**0.220**	**0.026**	**0.221**	**0.0258**	-0.190	0.600
	(0.099)		(0.099)		(0.362)	
Algorithmic Advice * Difficulty	-0.002	0.989	-0.003	0.985	0.042	0.774
	(0.139)		(0.139)		(0.146)	
Remote Associates Test			-0.039	0.220	-0.021	0.695
			(0.032)		(0.053)	
TRI					-0.006	0.731
					(0.017)	
Initial Confidence					**-0.691**	**<0.001**
					(0.055)	
English					0.428	0.582
					(0.778)	
Round					0.006	0.637
					(0.013)	
Algorithmic Advice * Remote Associates Test					-0.116	0.104
					(0.072)	
N	4190.000		4190.000		4190.000	
Log likelihood	-3476.531		-3475.782		-3307.781	
AIC	6967.06		6967.563		6643.562	

The dependent variable is a categorical variable based on a subject's final answer. TRI is a subject's score on the Insecurity Component of the TRI. Confidence is on a Likert scale from 1-4. Algorithmic Advice is a categorical variable (algorithmic condition = 1, human advice = 0) and Difficulty is a categorical variable (hard questions = 1, easy questions = 0). Standard errors are in parentheses. There were N = 419 subjects.

3.6.2 Confidence

Next, we observe the effects on confidence. In the model below, the dependent variable is the final level of confidence stated by the subject. As expected, subjects were less confident (p = in their final answers for hard problems (political cartoons), than easy problems (landscapes), across all models. Subjects exposed to algorithmic advice were not statistically significantly

more confident than subjects exposed to advice from other people, across all models. Most importantly, subjects did not rely more on algorithmic advice or advice from other people as questions changed in difficulty, and this effect was not significant across all models.

Table 3.3. Confidence Models

	Model 1	**P Value**	**Model 2**	**P Value**	**Model 3**	**P Value**
1|2	**-4.838**	**< 0.001**	0.260	0.488	0.431	0.303
	(0.185)		(0.376)		(0.418)	
2|3	**-2.458**	**< 0.001**	**3.080**	**< 0.001**	**3.257**	**< 0.001**
	(0.166)		(0.374)		(0.417)	
3|4	**1.057**	**< 0.001**	**7.225**	**< 0.001**	**7.415**	**< 0.001**
	(0.160)		(0.388)		(0.430)	
Algorithmic Advice	-0.047	0.834	-0.041	0.799	0.370	0.317
	(0.224)		(0.160)		(0.370)	
Difficulty	**-1.064**	**< 0.001**	**-0.254**	**0.012**	**-0.267**	0.027
	(0.094)		(0.101)		(0.120)	
Algorithmic Advice * Difficulty	-0.122	0.349	-0.146	0.284	-0.146	0.287
	(0.130)		(0.137)		(0.137)	
Initial Confidence			**2.070**	**< 0.001**	**2.061**	**< 0.001**
			(0.062)		(0.063)	
TRI			-0.006	0.765	-0.007	0.694
			(0.019)		(0.019)	
Remote Associates Test			-0.057	0.153	-0.005	0.937
			(0.040)		(0.059)	
Initial Time					0.000	0.437
					(0.000)	
Algorithmic Group * Remote Associates Test					-0.099	0.216
					(0.080)	
N	4190.000		4190.000		4190.000	
Log Likelihood	-4014.003		-3382.595		-3380.311	
AIC	8042.007		6785.190		6786.621	

The dependent variable is a categorical variable based on a subject's final stated level of confidence. TRI is a subject's score on the Insecurity Component of the TRI. Confidence is on a Likert scale from 1-4. Algorithmic Advice is a categorical variable (algorithmic condition = 1, human advice = 0) and Difficulty is a categorical variable (hard questions = 1, easy questions = 0). Standard errors are in parentheses. There were N = 419 subjects.

3.6.3 Time

For our models related to time, we used the percent change in time spent on a question as the dependent variable, to remain in accordance with the analysis plan from the first experiment. This dependent variable forced us to use a linear mixed effects model rather than a cumulative linked mixed model. Similar to the results related to time in experiment one, we did not observe

significant effects related to time, across any models for any variables. Most importantly, we did not observe a significant interaction between difficulty and advice source, indicating no support for H9.

Table 3.4. Time Models

	Model 1	p Value	Model 2	p Value	Model 3	p Value
(Intercept)	**-0.361**	**<0.001**	**-0.356**	**<.001**	**-0.534**	**0.046**
	(0.04)		(0.063)		(0.267)	
Algorithmic Advice	-0.068	0.227	-0.069	0.226	-0.115	0.298
	(0.056)		(0.057)		(0.11)	
Difficulty	**-0.181**	**0.001**	**-0.181**	**0.001**	**-0.188**	**0.001**
	(0.055)		(0.055)		(0.056)	
Algorithmic Advice * Difficulty	0.036	0.639	0.036	0.639	0.036	0.640
	(0.078)		(0.078)		(0.078)	
SD (Intercept)	0.134		0.135		0.138	
	(NA)		(NA)		(NA)	
SD (Observation)	1.257		1.257		1.257	
	(NA)		(NA)		(NA)	
Remote Associates Test			-0.001	0.923	-0.007	0.667
			(0.011)		(0.017)	
Initial Confidence					-0.015	0.511
					(0.023)	
English					0.253	0.301
					(0.244)	
Algorithmic Advice * Remote Associates Test					0.011	0.639
					(0.023)	
N	4190		4190		4190	
Log Likelihood	-6932.438		-6935.997		-6941.340	
AIC	13876.877		13885.995		13902.681	

The dependent variable is a categorical variable based on the change in time that a subject spent on a question. TRI is a subject's score on the Insecurity Component of the TRI. Confidence is on a Likert scale from 1-4. Algorithmic Advice is a categorical variable (algorithmic condition = 1, human advice = 0) and Difficulty is a categorical variable (hard questions = 1, easy questions = 0). Standard errors are in parentheses. There were N = 419 subjects.

3.7 Post-Hoc Results

We expected that TRI and RAT would strongly predict belief change and confidence, but we did not observe either to be a significant predictor, regardless of the dependent variable. This was

especially surprising because the Remote Associates Test, which requires significant thought, had a high average score: 4.12 out 6.00. Given this high score, we thought this was strong evidence that the subjects were paying attention and were especially paying attention when answering questions related to the control variables. However, the results indicate that both TRI and RAT are not predictive of reliance on advice.

We also expected that the RAT would moderate the effect of advice source on reliance on advice, although this was not a significant effect for any dependent variable.

3.8 Limitations and Future Directions

It is possible that the design of the experiment, with multiple-choice answers that allow for free response for some of the answers but not others, created incentives for the subjects to be lazy, and choose only the multiple choice responses that did not require a free response. Further research could conduct a similar experiment without this confound.

Further post-hoc analysis could explore the time variable in greater depth. It is possible that outliers are greatly influencing our findings, and that a standardization of the time variable could provide more insight into how people spend time on problems.

Furthermore, this research is the first we are aware of to examine cognitive effort, measured using time, and confidence in *creative* tasks informed by algorithmic and human advice. Further research should examine whether there are other factors that moderate the effect of advice source on reliance on advice in creative tasks.

3.9 Discussion

This paper shows that the type of task has a strong effect on the willingness to rely on algorithmic advice. In the intellective task there was a direct effect of algorithmic advice, and subjects relied more on algorithmic advice as problems became harder. In the creative task, both

the direct effect and interaction effect of algorithmic advice changed to become non-significant. This is almost assuredly due to the type of task, rather than other explanations. We believe that the type of task is the reason these effects changed, because the experiments were nearly identical except for the manipulation of difficulty and the type of task. The manipulation of difficulty was successful subjects took more time and were less confident on the questions we believed would be more difficult, just like in experiment one. These experiments were conducted within several months of each other, on the same platform, with different subjects from the same pool, and nearly identical manipulations of advice source. Thus, we conclude that type of task is the cause of the difference between experiments, because other factors are either identical (e.g. advice source), or very successfully transformed (e.g. difficulty).

Although our results are not significant, we believe this research open new frontiers for IS research. We study a context that is dramatically underexplored: human responses to recommendations from crowds and algorithms across different levels of difficulty in *creative* tasks. Understanding how humans respond to algorithms compared to human crowds in creative tasks will help in designing social media platforms, e-commerce platforms, and conversational agents.

Chapter 4

Algorithmic Appreciation in Mixed-Motive Tasks

4.1 Introduction

We study bail decisions for an example of a mixed motive task. Bail decisions have significant societal implications, and are enacted millions of times each year (Kleinberg et al. 2017). Bail judgments are racially biased against black defendants, and this effect is stronger among inexperienced and part-time judges (Arnold et al. 2018). The best economic estimates of the cost of being detained indicate that defendants lose approximately $30,000 in lost wages and government benefits (Dobbie et al. 2018). Bail decisions are determined by mostly untrained judges who have little interaction with the defendant (Arnold et al. 2018), indicating that algorithms might be a preferable societal alternative, if unbiased algorithms could be developed.

In the context of bail, the two parties with conflicting interests are the alleged malefactor and the broader society. The alleged deserves to not be unjustly held, but society deserves a criminal justice system that does an adequate job protecting citizens from the alleged offender who might not willingly return for their trial. Judges may have biases when weighing these two interests – and algorithms could correct these biases.

The goal of bail judges is well defined: “to set bail conditions that allow most defendants to be released while minimizing the risk of pretrial misconduct” (Arnold et al. 2018). The underlying idea is that people will be likely to return for their trial if they have a financial incentive to do so. Judges must also keep in mind that some alleged offenders are likely to believe that they are so likely to be convicted that it would be utility-maximizing for them to flee.

Unfortunately, a review of algorithmic risk assessment tools used by state governments indicates they create racist outcomes (Stevenson 2017). This can occur for several reasons. The outputs of the risk assessment tool can be misunderstood, ignored, or used off-label by judges (Stevenson 2017). Judges may also misuse the algorithm in self-serving ways when elections are near (Stevenson 2017). However, we suspect that the general public is unaware of the problems with algorithmic risk assessment in bail decisions.

Bail decisions are a thorny problem to evaluate using machine learning models. Because the observable outcomes are a consequence of human decision-makers, it is hard to know whether algorithms improve on outcomes. Thus, model evaluation is more difficult, as outcomes are not randomly sampling the population of cases (Lakkaraju et al. 2017). Thus, we never observe the crime outcomes for people who are not given bail (Kleinberg et al. 2017). Recent advances in model evaluation theory have led to the development of algorithms that dominate human judgment. New research suggests algorithmic bail could reduce post-bail crime by 24.8% with no change in jailing rates, or reduce jail populations by 42% with no increase in crime rates, while also decreasing the percentage of African-Americans and Hispanics in jail (Kleinberg et al. 2017).

4.2 Operationalization of Corollaries as Hypotheses

The mixed motive task has been operationalized as an assessment of how much bail to request of a hypothetical alleged criminal. Our following hypotheses are informed by our research into the intellective and creative task experiments conducted previously. As a result of these experiments, we changed away from the corollaries we originally submitted in the dissertation proposal. We preregistered the hypotheses at the Open Science Foundation.[2] The new hypotheses are:

[2] https://osf.io/7v4p3/?view_only=82beaf8b852a4a87865ae55cc444340f

Hypothesis 10: *When determining the level of bail for a defendant, people rely more on advice that is labeled as from an algorithm than that labeled as from a large number of other people.*

Hypothesis 11: *The stated difficulty of a question will positively moderate the effect of advice source on WOA.*

We did not hypothesize about confidence and time in the preregistration. Due to the prior two experiments having very similar effects between confidence and WOA, and no significant effect related to time, we decided to leave these variables for post-hoc analysis.

4.3Experimental Design

4.3.1 Scenario Acquisition

We listened to publicly available bail hearings from a major metropolitan city in the United States and took contemporaneous notes about each hearing. We noted the alleged crime, the criminal history, the scales describing the probability of fleeing and of reoffending, and any mitigating information offered by the alleged criminal's lawyer. In the jurisdiction there are two common outcomes: cases where the alleged criminal would have to pay money up-front for pre-trial release (typically for more serious crimes), and cases where the alleged would not have to provide funds unless they failed to appear in court (typically for less serious crimes). We included only scenarios for which the defendant would have to pay in order to be released from jail. For each scenario, we removed data that would easily identify the defendant (e.g., the defendant's name, jurisdiction of the crime, defendant's hometown, etc.). For data relevant to the bail decision that we could make less identifying, we de-specified the data. For example, each subject's age is listed as being within a bracket (e.g., "Defendant is in his 30s"). An example of a scenario, along with the questions we asked each subject, is in Appendix 4: Bail Scenario.

4.3.2 Design

We use a 2 by 2 between-and-within subjects design (Table 4.1). The conditions are: (1) whether the subject received the algorithmic or crowd recommendation, and (2) whether the task is easy or difficult.

Table 4.1: Mixed-Motive Experimental Design

Task difficulty	**Algorithmic Recommendation**	**Crowd Recommendation**
Stated Difficulty	An algorithm, trained on 5,000 similar cases, concluded that the correct bail is $X.	5,000 people reviewing the same case concluded, on average, that the correct bail is $X.

We delivered this advice at both the top and bottom of the page the second time a subject saw a scenario, and we asked them to "incorporate the advice." We reminded subjects what their prior amount of bail set for this scenario was by stating it directly above the text box where subjects were asked to input the bail amount for the second time.

4.3.3 Operationalization

To operationalize the different advice sources, we use different labels, as shown in the table above. To operationalize difficulty, we state whether a judge thought the bail decision was difficult. During pilot testing, we tried seven methods to operationalize the difficulty of a bail decision, before finally deciding that subjects would be told the decision was easy or difficult.

4.3.4 Pilot Tests

First, we tested whether subjects believed scenarios were more difficult based on a contrast between the severity of the crime and the likelihood of fleeing. We believed that when the severity of a crime was high (low) and the probability of fleeing was low (high), subjects would perceive a bail determination as difficult compared to when these factors were either both low or both high. We created three categories: Unambiguous High, for alleged felons who were high

flight risks; Ambiguous, for alleged criminals with moderate flight risk; and Unambiguous Low, for alleged criminals who committed misdemeanors and had low flight risk. See examples of each in Table 4.2.

Table 4.2 Pilot One Example Scenarios

Unambiguous High	Ambiguous	Unambiguous Low
Imagine an alleged criminal who is: 32 years old, a wealthy socialite, the owner of two passports, and accused of stealing money from investors (a felony). If you were the judge, at what level would you set the bond?	Imagine an alleged criminal who is a father, with some means, will be fired if he can't post bail, and accused of robbing a 711 (a misdemeanor). If you were the judge, at what level would you set the bond?	Imagine an alleged criminal who: is 91, has no criminal history, has two grandchildren in town he lives in, and is charged for disorderly conduct (a misdemeanor). If you were the judge, at what level would you set the bond?

Our results indicated that this manipulation was moderately successful in causing subjects to perceive scenarios as different levels of difficulty. See Table 4.3.

Table 4.3 Perceived Difficulty of Questions in Pilot One

Scenario	Perceived Difficulty
Ambiguous 1	2.34
Ambiguous 2	2.34
Unambiguous High 2	2.31
Unambiguous High 5	2.20
Ambiguous 5	2.17
Ambiguous 3	2.17
Unambiguous High 4	2.06
Unambiguous High 3	2.03
Ambiguous 4	2.0
Unambiguous High 1	1.86
Unambiguous Low 3	1.86
Unambiguous Low 2	1.48
Unambiguous Low 4	1.48
Unambiguous Low 5	1.41
Unambiguous Low 1	1.37

Second, we tested whether subjects believed scenarios were more difficult when they had more information about an alleged criminal. For example, we would present no mitigating information for some scenarios, and we suspected that would make a scenario easier. Our results indicate that subjects did not believe these were easier.

Third, we tested whether subjects believed scenarios were more difficult if they received an example amount of bail, providing them with some initial guidance as to what a normal amount of bail would be for a sample crime. Our results indicate that subjects did not report scenarios were easier when they received this guidance.

Fourth, we tested whether subjects believed scenarios were more difficult based on receiving numbers, called scales, that quantified the likelihood of fleeing or of committing another offense. Subjects did not report scenarios were easier when they received this information.

Fifth, we tested whether subjects believed scenarios were more difficult when they received both an example and scales. Subjects did not report scenarios were easier when they received this information.

Sixth, we tested whether scales made questions easier if they were delivered within-subject rather than between-subject. Our results indicate that subjects did not consistently report scenarios with scales were easier.

Finally, we tested whether subjects believed scenarios were more difficult based on the stated difficulty of a question. For each question, we would say whether a judge believed that a scenario was easy or hard. In fact, no judge had made these assessments and they were randomly assigned. This was a between-subjects condition. For five of the eight scenarios, subjects perceived the question labeled as difficult as harder than the easy question. We used the four

scenarios where the difference between hard and easy was in the right direction and strongest (Scenario 1, Scenario 3, Scenario 4, and Scenario 5).

We outline the results of tests two through seven in Table 4.4.

We also spoke with three judges responsible for bail decisions in the United States. They indicated that they all bail decisions to be approximately the same degree of difficulty. However, these judges always receive cases with all the facts and with no experimental manipulations. Although the judges find all scenarios approximately equally difficult, we believe that our manipulation of difficulty was both coherent *a priori* and confirmed *ex post facto*, through our results indicating that subjects treated easy and hard questions differently.

Table 4.4 Pilot Tests

	Test 2, 3, 4, and 5					Test 6		Test 7	
Scenario	Both	No Example No Scale	Example	Scale	Scales + No Mitigating	Scales (within subject)	No Scales (within subject)	Hard	Easy
1	2.57	2.50	2.70	2.00	2.44	2.91	2.90	3.5	3.16
2	2.70	3.00	2.00	2.33	3.00	2.92	2.71	2.25	3.4
3	2.43	3.00	2.50	2.45	2.38	2.66	2.44	2.83	2.57
4	2.33	2.30	2.60	2.27	2.80	3.11	3.08	2.83	1.4
5	2.50	2.63	2.00	2.73	2.63	1.60	2.73	3.2	2.57
6	3.00	2.50	1.75	2.33	2.67	2.81	2.63	2.66	2.5
7	1.56	2.89	3.00	2.82	2.78	3.33	2.91	3.14	3.28
8	2.27	2.22	3.00	2.20	2.30	2.18	2.72	2	2.44
Average Difficulty	2.42	2.63	2.44	2.39	2.62	2.69	2.77	2.80	2.66
N	14	14	14	17	15	34	34	12	13

Scenarios highlighted in yellow were used in the experiment.

4.4 Results

We define commonly used terms in the results tables in Table 4.5.

Table 4.5 Common Variables in Results Tables

Term	Definition	Distribution	Type
WOA	Weight on Advice	0-1	Continuous
Initial Confidence	Initial level of confidence in a scenario	1-4 (Likert Scale)	Ordered Categorical
Algorithmic Advice	Whether the subject was in a group receiving algorithmic advice	0 (Crowd) or 1 (Algorithmic)	Categorical (Binary)
Difficulty	Whether the scenario was described as easy or hard	0 (easy) or 1 (hard)	Categorical (Binary)
Technology Readiness Index (TRI)	A subject's score on the Insecurity Questions of the Technology Readiness Index	4-20 (Sum of Likert Scales)	Sum of Ordered Categorical
Identification With Criminal Others (ICO)	A subject's score on the ICO component of Modified Criminal Sentiments Scale	6-30 (Sum of Likert Scales)	Sum of Ordered Categorical

4.4.1 Randomization and Manipulation Check

We first check that the manipulations were successful by comparing the averages of each group, with a series of two sample t-tests. Subjects in the algorithmic advice condition had similar levels of initial confidence in their answers as subjects in the crowd condition ($p = 0.53$), indicating that subjects were identical prior to seeing the treatment. Subjects also spent the same amount of time on questions when answering them for the first time ($p = 0.96$), reinforcing that subjects were identical between the groups prior to exposure to the treatment. Subjects spent more time answering a question for the first time when they were told the question was hard ($p < 0.001$). In practical terms, subjects spend about 31% longer on questions labeled as difficult compared to questions labeled as easy. This result indicates that subjects noticed the stated difficulty and is evidence the manipulation worked.

4.4.2 Testing For Multicollinearity

We test for multicollinearity of independent variables using the variance inflation factor (VIF) test. Generally, a VIF above 5.0 indicates problematic multicollinearity (Sheather 2009). We calculated the VIFs for all variables in following models. The maximum VIF for any variable was 2.91, indicating multicollinearity was not likely driving bias. VIFs are not a perfect criterion to judge multicollinearity, and even models with low VIFs can have spurious false positives (Goodhue et al. 2017). However, because both of the hypotheses are not supported, false positives that might occur despite low VIFs are not a concern.

We also checked the pairwise correlations for all independent variables see Table 4.6. The absolute value of the largest correlation was 0.15, further indicating that multicollinearity was not problematic. If we had problematic multicollinearity then we could have used either principal components analysis or a penalty function such as LASSO or Ridge regression to handle multicollinearity (Tibshirani 1996).

Table 4.6 Correlations

	WOA	Algorithmic Advice	Difficulty	TRI	ICO	Initial Confidence	Order	Accuracy	Initial Time
WOA	1.00								
Algorithmic Advice	0.04	1.00							
Difficulty	-0.04	-0.01	1.00						
TRI	0.03	0.04	0.02	1.00					
ICO	-0.09	0.03	0.01	-0.02	1.00				
Initial Confidence	-0.12	0.02	-0.07	0.09	0.03	1.00			

Order	-0.05	-0.01	0.00	0.00	0.01	-0.03	1.00		
Accuracy	0.05	-0.01	0.01	-0.01	-0.04	-0.06	-0.04	1.00	
Initial Time	0.01	0.00	0.02	0.04	-0.03	0.02	-0.15	0.00	1.00

4.4.3 Statistical Models

We fit two linear mixed effects models (Model 1 and Model 2) with WOA as the dependent variable (Bates et al. 2014) (Table 4.7). Model 1 tests the hypothesized effects without control variables, Model 2 adds control variables. Theoretically, a mixed effects model would indicate that the slope is the different across subjects and the deviation around the mean is different (Kennedy 2008).

Interestingly, our simple model without control variables has a lower AIC. When comparing two similar models, if the difference in AIC is larger than 2, then the model with the lower AIC should be considered superior (Burnham and Anderson 2004). The simplicity of the first model overwhelms the benefits of the control variables, although this might be surprising because the ICO, Initial Confidence, and Accuracy variables are all statistically significant predictors of WOA. However, because Model 1 has a superior fit, the coefficients we report on below are from Model 1.

We fail to reject the null hypothesis for H10, that people rely more heavily on algorithmic advice ($\beta = 0.029$, $p = 0.394$). The coefficients for Algorithmic Advice in Experiment 1 and the coefficient in this experiment can be directly compared, because the DV and independent variables are identical and on the same scale. For this mixed-motive task, the effect size ($\beta = 0.029$) is about one quarter of the size we observe in the best fitting model for the intellective task ($\beta = 0.108$).

We also fail to reject the null hypothesis for H11, that difficulty moderates the effect of algorithmic advice, (β = -0.003, p = 0.944). Both H10 and H11 are not supported regardless of whether Model 1 or Model 2 are used. Clearly, task type matters!

Table 4.7 WOA Model Results

	Model 1	P Value	Model 2	P Value
(Intercept)	**0.475**	**< 0.001**	**0.687**	**< 0.001**
	(0.024)		(0.081)	
Algorithmic Advice	0.029	0.394	0.032	0.343
	(0.034)		(0.034)	
Difficulty	-0.029	0.279	-0.035	0.189
	(0.027)		(0.027)	
Algorithmic Advice * Difficulty	-0.002	0.966	-0.003	0.933
	(0.039)		(0.039)	
SD(Intercept)	0.188		0.183	
	(NA)		(NA)	
SD(Observation)	0.341		0.339	
	(NA)		(NA)	
Technology Readiness Index			0.004	0.210
			(0.003)	
Identification With Criminal Others			**-0.009**	**0.013**
			(0.003)	
Initial Confidence			**-0.059**	**0.000**
			(0.015)	
Order			-0.015	0.107
			(0.009)	
Accuracy			**0.001**	**0.025**
			(0.000)	
Initial Time			0.000	0.410
			(0.000)	
Mixed Effects	Yes		Yes	
N	1254		1254	
Log Likelihood	-569.422		-585.644	
AIC	1150.843		1195.289	

The dependent variable is weight on advice. Standard errors are in parentheses. There were N = 353 subjects.

4.5 Post-Hoc Analyses

4.5.1 Confidence

As a post-hoc analysis, we test whether subjects were more likely to rely on algorithmic advice, manifested by being more confident in their final answer or by taking less time in answering the question (Table 4.8). We again use mixed effects models for Model 1 and Model 2, adding control variables in Model 2. The AIC and log likelihood of Model 2 indicate it is the model of

best fit. Subjects were less confident for hard questions, across both models, demonstrating the salience of the difficulty manipulation. Our models detect no significant difference between people who receive algorithmic advice compared to advice from a crowd ($\beta = 0.040$, $p = 0.483$) on how confident they are in their final answer. Our models also do not detect any moderating effect of difficulty on the relationship between algorithmic advice and confidence ($\beta = 0.030$, $p = 0.609$). There is some evidence that indicates that TRI and ICO are predictive of confidence. Lastly, we also run a Tobit model instead of mixed effects model, because the dependent variable is an ordered categorical variable (Kennedy 2008). The results do not change.

Table 4.8. Confidence Results

	Model 1	**P Value**	**Model 2**	**P Value**
(Intercept)	**3.173**	**< 0.001**	**1.959**	**< 0.001**
	(0.048)		(0.136)	
Algorithmic Advice	0.048	0.486	0.040	0.483
	(0.070)		(0.057)	
Difficulty	**-0.147**	**0.000**	**-0.113**	**0.005**
	(0.041)		(0.040)	
Algorithmic Advice * Difficulty	0.037	0.535	0.030	0.609
	(0.059)		(0.058)	
TRI			**0.015**	**0.009**
			(0.006)	
ICO			**-0.016**	**0.012**
			(0.006)	
Initial Confidence			**0.369**	**< 0.001**
			(0.025)	
Order			-0.017	0.204
			(0.014)	
Accuracy			**0.002 *****	**< 0.001**
			(0.000)	
Random Effects	Yes		Yes	
N	1254		1254	
Log Likelihood	-1223.454		-1139.274	
AIC	2458.907		2300.548	

The dependent variable is the final level of confidence. Standard errors are in parentheses. There were N = 353 subjects.

4.5.2 Time

We also conducted a post-hoc analysis on the time spent on a problem. Like the models using WOA and Confidence as a dependent variable, we use a mixed effects model without control variables as Model 1. We use the same model with control variables as Model 2. Model 2 had the lower AIC, so the effects we report are from Model 2, although the results between the two models are similar. The dependent variable for all models is the amount of time spent on a question after receiving advice. The effects for algorithmic advice ($\beta = -2.640$, $p = 0.544$) and the interaction between algorithmic advice and difficulty ($\beta = 4.243$, $p = 0.479$) were not statistically significant. The order of a problem was predictive ($\beta = -6.831$, $p < 0.001$), meaning that subjects spent less time on problems in later rounds than earlier rounds.

Table 4.9 Time Models

	Model 1	P Value	Model 2	P Value
(Intercept)	**28.215**	**< 0.001**	**45.390**	**<0.001**
	(3.078)		(10.169)	
Algorithmic Advice	-2.252	0.609	-2.640	0.544
	(4.403)		(4.351)	
Difficulty	-7.608	0.0726	-7.858	0.060
	(4.233)		(4.172)	
Algorithmic Advice * Difficulty	3.984	0.5138	4.243	0.479
	(6.099)		(6.003)	
TRI			0.449	0.232
			(0.375)	
ICO			-0.334	0.398
			(0.396)	
Initial Confidence			0.302	0.879
			(1.999)	
Order			**-6.831**	**< 0.001**
			(1.426)	
Accuracy			**-0.072**	**0.030**
			(0.033)	
Initial Time			**0.030**	**0.004**
			(0.011)	
N	1254		1254	
Log Likelihood	-6781.745		-6763.904	
AIC	13575.489		13551.808	

The dependent variable is the time, in seconds, spent on the question after receiving advice. Standard errors are in parentheses. There were N = 353 subjects.

4.6 Limitations

This research has several limitations. Manipulating the difficulty of a bail decision is challenging and might require more creative approaches so that it is correctly perceived by the subjects rather than stated as a treatment condition. Moral questions such as those involving punishment, and occasionally capital punishment, are perhaps so inherently equivocal that making difficulty a problem feature is near impossible. Furthermore, it is possible there is another way to establish or manipulate difficulty that produces a difference between treatments.

It is also possible that the subjects recruited were not motivated enough to thoughtfully consider the scenarios. Although we removed observations when subjects went too quickly, or failed the attention or manipulation check, it is still a possibility that they were simply not sufficiently motivated, though the same subject pool was used for the first experiment.

Lastly, it is possible that the manipulation of the advice source, while salient enough to result in subjects passing the manipulation check, was not sufficiently realistic. However, computer recommendations are oftentimes offered with similarly salient labels, such as when Zillow says, "This house is for you", (algorithmic advice) or when a dating application allows your friend to recommend you to someone else on the platform (social advice).

4.7 Discussion

This experiment makes several contributions. First, it provides important context around how people perceive algorithmic appreciation (aversion) in mixed-motive tasks. Bail decisions are a suitable candidate for algorithmic recommendations; already-designed algorithms are superior to the judgment of judges (Kleinberg et al. 2017). However, these bail decisions are often misused or underutilized (Stevenson 2017). Future research could investigate how to illustrate to the public and judges the efficacy of algorithmic bail decisions.

Second, this chapter, in tandem with the chapter on the intellective task, indicates that the moderating effect of difficulty and direct effect of algorithmic advice is contingent on the type of task. This is valuable information and adds knowledge above and beyond what we learned from the creative task.

Future work could also examine the effect of training on algorithmic appreciation in mixed motive tasks. It is possible people are unwilling to be persuaded by the efficacy of algorithms in mixed motive tasks, because they are worried about implicit biases of the algorithm makers, they

are aware that algorithms are trained on datasets in which the data does not include counterfactuals, or it might be that humans are not ready to accept AI advice for mixed motive tasks. AI is currently a black box and acceptance requires faith in the objectivity of algorithm without an explanation of how it works.

Chapter 5

5.1 Introduction

In this chapter we review the results of a combined model that includes observations from each task. First, we discuss the methods for this combined model, including a forward feature selection model that provides atheoretical support for identifying the most important variables in our models. Then, we discuss the results of our overall models on WOA and confidence dependent variables. The final sections of this chapter outline future research that would augment these findings and increase the likelihood of publishing, the limitations of this research, and the implications for managers and theory.

5.2 Methods

We test three linear mixed-effects models to determine whether there is a difference between task types and reliance on algorithmic advice.

5.2.1 Data Wrangling

Prior to running those models, we manipulate the data to make it more comparable between experiments. The data from the creative task was substantially different from the data from the other two tasks. For the creative task, we do not have a calculable initial answer, because it is a text caption. Thus, in our models combining the results of all three experiments we cannot use accuracy, which includes the distance between the initial answer and the correct answer, as an independent variable in the model. Furthermore, for the creative task our dependent variable was an ordered categorical variable. To make the data comparable, we change the dependent variable of the creative task to be scaled between zero and one. When subjects chose to keep their original caption, we assign that a WOA of zero. We changed the option to tweak a caption to have a

WOA of 0.33. We changed the option to create a new caption to a WOA of 0.66, and we changed the option to use the recommended caption to a WOA of 1. As a robustness check, we also coded any change as a WOA of 1, while keeping the subject who chose to keep their original caption as a WOA of 0. This did not meaningfully change the results.

Furthermore, of the three experiments we used for the intellective task, we only include the data from the first experiment in this analysis, because that was the only experiment using a between-subjects condition with high-quality advice.

5.2.2 Model Selection: Forward Feature Selection

We use a step-wise forward feature selection model to determine whether an atheoretical analysis supports our theory-driven understanding (Ferri et al. 1994). Forward selection runs a linear model on all possible combinations of the independent variables. It begins by choosing the model of best fit with one predictor, then moves to the model of best fit with two predictors, until all predictors are used (Figure 5.1). The leftmost graph indicates that the highest adjusted R^2 value occurs when we include six independent variables, although the adjusted R^2 is similar when using five independent variables or seven independent variables. The middle plot indicates that the Bayesian Information Criteria (BIC) is lowest with five predictors, closely aligned with the adjusted R^2, as expected. The rightmost plot tells us the most important information. It indicates how much the adjusted R^2 improves with each new feature included in the model. The most important predictor of WOA across all variables in all experiments is initial confidence. A simple model with two terms, initial confidence and an intercept term, gives an adjusted R^2 of 0.041. The next most important term is whether the task was intellective. Adding this term increases the adjusted R^2 by more than 50 percent, to 0.064. Adding the mixed motive variable

increases the R^2 to 0.084. The final two useful terms of the model are Algorithmic Advice and Difficulty. Including these increase our R^2 to 0.095. Adding the order of the questions and the time spent on a question prior to advice do not increase the R^2 by more than 0.001.

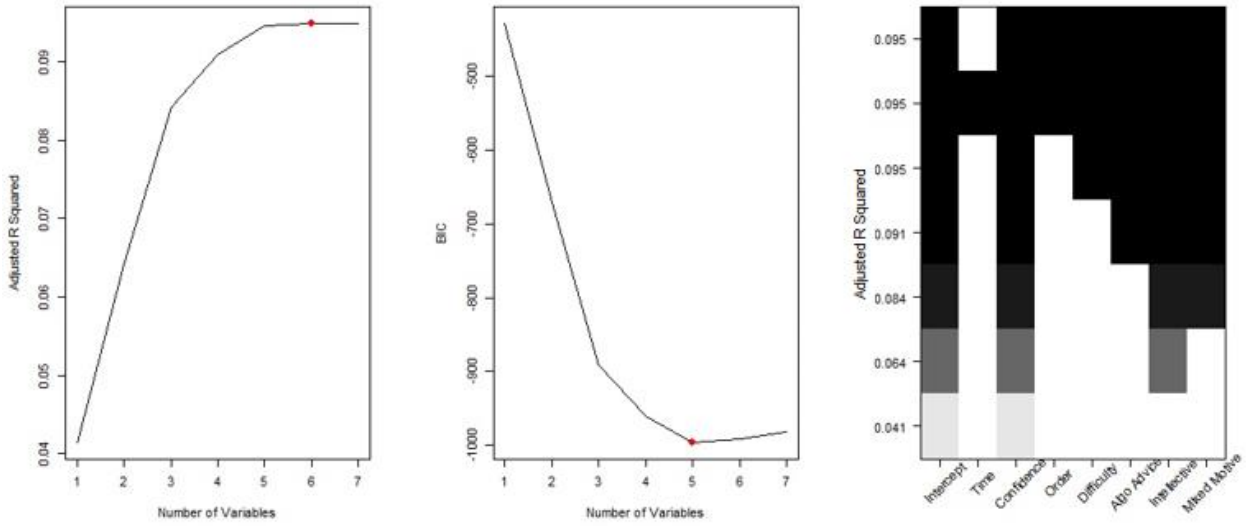


Figure 5.1 Forward Feature Selection Results

These plots have several shortcomings that we address later in this chapter. First, forward feature selection tends to overfit and does not effectively predict out-of-sample data. Second, our forward feature selection procedure uses a simple linear model, rather than a mixed effects model.

5.2.3 Analytical Approach

Our main model is similar to our model from the intellective and mixed-motive tasks, except this time we introduce the Task categorical variable and an interaction between task and algorithmic advice:

$$y_{ik} = \beta_{0i} + \beta_1 \text{AlgoCondition}_i + \beta_2 \text{Difficulty}_k + \beta_3 \text{AlgoCondition}_i \times \text{Difficulty}_k + \beta_4 Task_k + \beta_5 Task_k \times AlgoCondition_i + \beta X_{ik} + \varepsilon_{ik}$$

In this model, y_{ik} is a dependent variable for participant i and question k; β_{0i} is the coefficient for participant i; AlgoCondition_i is a categorical variable indicating the advice condition and Difficulty_k is a categorical variable indicating the problem difficulty. X_{ik} is a vector of control variables.

5.3 Results

For each dependent variable we run three models. The first model uses the variables that are of primary interest to this dissertation, the second model adds control variables, and the third model adds an interaction between task and algorithmic advice.

5.3.1 Weight on Advice

We use an unweighted linear mixed effects model to test whether we observe different reliance on advice across experiments. This is appropriate for seeing whether any variables that interact with the experimental condition is different from one another, but is not useful when looking at terms such as the effect of algorithmic advice alone, because we have far more observations from the intellective task and creative task than from the mixed-motive task. Thus, if we looked directly at the effect of algorithmic advice, for example, our results would be heavily skewed away from the results of the experiment with a smaller sample. We find that both the intellective and mixed motive experiments have positive and significant coefficients in all three models, indicating that subjects in those experiments relied more on advice than subjects in the experiment for the creative task. Lastly, we observe a positive and significant interaction between algorithmic advice and intellective tasks, indicating that subjects relied more heavily on algorithmic advice in that task than on correct advice from the creative task. See Table 5.1.

Table 5.1 Summary Results

	Result	P Value	Result	P Value	Result	P Value
Intercept	**0.221**	**< 0.001**	**0.484**	**< 0.001**	**0.510**	**< 0.000**
	(0.013)		(0.021)		(0.026)	

Algorithmic Advice	**0.052**	**< 0.001**	**0.056**	**< 0.001**	0.006	0.772
	(0.014)		(0.014)		(0.021)	
Difficulty	**0.079**	**< 0.001**	**0.033**	**< 0.001**	**0.033**	**< 0.000**
	(0.010)		(0.010)		(0.009)	
Intellective	**0.195**	**< 0.001**	**0.165**	**< 0.001**	**0.112**	**< 0.000**
	(0.014)		(0.014)		(0.019)	
Mixed Motive	**0.183**	**< 0.001**	**0.207**	**< 0.001**	**0.198**	**< 0.000**
	(0.017)		(0.018)		(0.024)	
Algorithmic Advice * Difficulty	0.021	0.132	0.020	0.137	0.019	0.147
	(0.014)		(0.014)		(0.013)	
Initial Confidence			**-0.094**	**< 0.001**	**-0.095**	**< 0.000**
			(0.005)		(0.005)	
Order			**0.003**	**0.011**	**0.003**	**0.011**
			(0.001)		(0.001)	
Initial Time			0.000	0.724	0.000	0.747
			(0.000)		(0.000)	
Algorithmic Advice * Intellective					**0.106**	**< 0.000**
					(0.027)	
Algorithmic Advice * Mixed Motive					0.001	0.639
					(0.003)	
N	10527		10527		10527	
Log Likelihood	-4695.704		-4540.000		-4536.967	
AIC	9407.409		9102.000		9099.935	

The dependent variable is weight on advice. Standard errors are in parentheses. Statistically significant effects are denoted in bold. There were N = 1,302 subjects.

As a robustness check we also run the models above using random sampling from the intellective and mixed motive tasks, so that there were 1,254 observations in each of the three groups (Table 5.2). This ensures that the experiments with more observations were not biasing our results. We highlight any values that are different between the two models in yellow. In the models with equal observations for each experiment, difficulty becomes non-significant ($\beta = 0.02, p = 0.201$), and so does order ($\beta = 0.004, p = 0.086$).

Table 5.2 Equally Weighted Experiments

	Result	P Value	Result	P Value	Result	P Value
(Intercept)	**0.234**	**< 0.001**	**0.437**	**< 0.001**	**0.452**	**< 0.001**
	(0.018)		(0.032)		(0.034)	
Algorithmic Advice	**0.054**	**0.006**	**0.058**	**0.003**	0.033	0.262
	(0.020)		(0.020)		(0.030)	
Difficulty	**0.052**	**0.002**	0.022	0.187	0.022	0.201
	(0.017)		(0.017)		(0.017)	
Intellective	**0.199**	**< 0.001**	**0.176**	**< 0.001**	**0.137**	**< 0.001**
	(0.019)		(0.019)		(0.027)	
Mixed Motive	**0.185**	**< 0.001**	**0.210**	**< 0.001**	**0.213**	**< 0.001**
	(0.020)		(0.021)		(0.028)	
Algorithmic Advice * Difficulty	0.011	0.638	0.012	0.599	0.012	0.613
	(0.024)		(0.024)		(0.024)	
Initial Confidence			**-0.078**	**< 0.001**	**-0.078**	**< 0.001**
			(0.008)		(0.008)	
Order			0.004	0.087	0.004	0.086
			(0.002)		(0.002)	
Initial Time			0.000	0.661	0.000	0.684
			(0.000)		(0.000)	
Algorithmic Advice * Intellective					**0.077**	**0.042**
					(0.038)	
Algorithmic Advice * Mixed Motive					-0.008	0.843
					(0.039)	
N	3762		3762		3762	
Log Likelihood	-1782.7		-1753.3		-1755.0	
AIC	3581.4		3528.6		3536.1	

The dependent variable is weight on advice. Standard errors are in parentheses. Statistically significant effects are denoted in bold. Effects that are different from the effects in table 5.1 are highlighted in yellow. There were N = 1,247 subjects.

5.3.2 Confidence

When we evaluate the effects on confidence, we find that subjects relied on advice less in the intellective task than in the creative task, and more in the mixed motive task than in the creative task. This might indicate that creative task was inherently harder than the intellective task, or it may indicate that subjects are simply less confident in their creative skills than their intellective skills. We also find that the interaction between algorithmic advice and the intellective task is positive and significant, indicating that subjects were more confident in advice from algorithms

in the intellective task than in the creative task. The interaction between algorithmic advice and the mixed motive task was not significant, indicating that subjects were equally confident in advice from algorithms for the creative and mixed-motive tasks.

Table 5.3 Confidence Results

	Result	P Value	Result	P Value	Result	P Value
(Intercept)	**3.128**	**< 0.001**	**1.483**	**< 0.001**	1.516	**< 0.001**
	(0.035)		(0.034)		(00.037)	
Algorithmic Advice	0.067	0.063	0.040	0.086	-0.026	0.481
	(0.036)		(0.023)		(0.036)	
Difficulty	**-0.403**	**< 0.001**	**-0.141**	**< 0.001**	-0.141	**< 0.001**
	(0.017)		(0.015)		(0.015)	
Intellective	**-0.270**	**< 0.001**	**-0.092**	**< 0.001**	-0.150	**< 0.001**
	(0.039)		(0.024)		(0.033)	
Mixed Motive	**0.171**	**< 0.001**	**0.096**	**0.001**	0.062	0.123
	(0.045)		(0.029)		(0.040)	
Algorithmic Advice * Difficulty	0.008	0.744	0.016	0.427	0.016	0.438
	(0.024)		(0.021)		(0.021)	
Initial Confidence			**0.544**	**< 0.001**	0.544	**< 0.001**
			(0.008)		(0.088)	
Order			0.003	0.165	0.003	0.165
			(0.002)		(0.002)	
Initial Time			0.000	0.187	0.000	0.192
			(0.000)		(0.000)	
Algorithmic Advice * Intellective					0.117	**0.013**
					(0.047)	
Algorithmic Advice * Mixed Motive					0.068	0.232
					(0.057)	
N	10527		10527		10527	
Log Likelihood	-10935.073		-9169.495		-9170.571	
AIC	21886.147		18360.990		18367.142	

Due to the tasks requiring significantly different amounts of reading and comprehension, we do not formally test which tasks resulted in more time spent on a problem.

As a final result of the dissertation, we list which hypotheses were supported and not supported in Table 5.4.

Table 5.4 Summary of Results

Hypothesis	Supported?
For intellective tasks, the effect of algorithmic advice on belief change will be stronger than the advice of a crowd.	Yes
For intellective tasks, the effect of algorithmic advice on belief change will be stronger for a more difficult task.	Yes
For intellective tasks, algorithmic advice will result in less cognitive effort compared to the advice of crowds.	No
For intellective tasks, the effect of algorithmic advice on cognitive effort will be stronger for a more difficult task.	No
For intellective tasks, algorithmic advice makes humans more confident in their decisions than the advice of crowds.	Yes
For intellective tasks, the effect of algorithmic advice on decision confidence will be stronger for a more difficult task.	Yes
For creative tasks, the effect of algorithmic advice on belief change will be stronger for a more difficult task.	No
For creative tasks, the effect of algorithmic advice on cognitive effort will be stronger than the effect of crowd advice for a more difficult task.	No
For creative tasks, the effect of algorithmic advice on confidence will be weaker for a more difficult task.	No
For mixed-motive tasks, the effect of algorithmic advice on belief change will be greater than the effect of the advice of a crowd on belief change.	No
For mixed-motive tasks, the effect of algorithmic advice will be moderated by task difficulty.	No

5.4 Limitations

One of the significant limitations of this research is that it does not directly inform whether an individual will rely more on an algorithm or on advice from peers for a specific task. Rather, we

evaluate what occurs when a focal task gets harder and how that changes based on the type of task. Thus, our research is complementary to other efforts that more directly answer whether an individual will rely on algorithmic advice for a task.

Like all experimental research, the effects we observe may be biased because subjects knew that they were being observed. Additionally, sampling from Mechanical Turk may have biased our results, if people on Mechanical Turk have different propensities to rely on algorithmic advice, or different interpretations of how difficulty changes in tasks. In the future research section below we outline several ways to address this bias.

5.5 Future Research

An avenue for future research is to measure cognitive effort more directly by using NeuroIS techniques such as EEG or fMRI technology. Future collaborations exploring this opportunity are in progress, and they could be helpful in understanding why we observed largely null effects in the creative experiment. Incorporating other types of creative tasks and using NeuroIS techniques would likely result in high-quality journal publications.

Future research could also investigate whether we face a social acceptability confound. It is possible that we subjects rely more on advice when they thought it was socially acceptable. In tasks like counting, there is no risk of appearing callous when relying on an algorithm. Without an instrument measuring social acceptability we do not have clear evidence proving or disproving this possibility. In the mixed-motive task, on the other hand, relying heavily on an algorithm may cause someone to feel either callous or be self-conscious about their willingness to rely on advice. Using other relevant variables, such as whether the decision is made via Zoom or in person, could result in publishable findings.

Although the research from the intellective task is published, further research into how cultures affect algorithmic appreciation could shed new light on the phenomena. Specifically, if other cultures have greater individualism or collectivism, then they may be more willing to rely on advice from a crowd, which could influence the relative difference between crowd and algorithmic advice.

Another way we could build on the results from the intellective task would be to observe how people rely on algorithmic advice using archival data. Platforms that recommend products, such as Amazon, Netflix, or Zillow would have perfect data to compare the relative effect of algorithmic versus crowd recommendations.

5.6 Discussion

This dissertation focuses on how difficulty affects algorithmic appreciation and algorithmic aversion across three types of tasks. We found that the task type affects the degree to which subjects rely on WOA, and that subjects relied significantly more on algorithmic advice in the intellective task. An important takeaway from this project is that self-reported confidence is strongly positively correlated with WOA, and that time spent on a problem is not correlated with either weight on advice or self-reported confidence.

This research can inform several important questions relevant to managers. First, it can help managers realize when they may utilize algorithmic or human advice. This may have utility in answering questions such as when to leverage humans or machines in providing help to a client, such as through a help desk. It could also inform managers of technology platforms whether they want to frame a recommendation as stemming from their algorithms or from other people. Of course, the recommendations would be the same, because the algorithms are trained from other customer's decisions, but the *framing* of the recommendation can be informed by this research.

Although this may require some minor deception, or would at least require the platform to gloss over the similarities between a crowd recommendation and an algorithm trained based on crowd behavior, we believe this could be accomplished through a simple A/B test that companies run routinely. Tests that check how people respond to different framings, colors, or other aesthetics are run constantly by the major tech platforms. If there were still concerns about the ethics of this slight deception, subjects could even be informed of the results after the experiment concluded.

Our research also has important implications for theory. This is the first paper we are aware of that looks at cognitive effort, behavioral change, and confidence in tandem. Prior research has used physiological measurements such as heart rate to measure cognitive load, a construct closely related to cognitive effort (Alexander et al. 2018). Our results highlight the noisiness of measuring cognitive effort using time spent, and the strong correlation between behavioral change and confidence. Future research could use NeuroIS techniques to determine the relationship between time spent and physiological measures such as brain activity or heart rate. We are also the first paper to investigate algorithmic appreciation in a creative task, and the first paper researching algorithmic appreciation to use McGrath's Circumplex Model to inform our task choices.

We believe one explanation of these results is that the public is ready to accept advice for a broad swathe of intellective tasks. This is demonstrated by reliance on algorithms for finance, weather forecasting, or games such as chess. Similarly, our results explain why people are so surprised at the effectiveness of algorithms in other types of tasks, such as when apps like Spotify are able to recommend music so well (Gershgorn 2019). As a society, we are not yet aware of how our music tastes can be effectively distilled into a numerical calculus.

We already know that task objectivity predicts whether an individual will prefer human or algorithmic advice (Castelo et al. 2019). These experiments are capturing a snapshot in time for the public's trust in algorithms. Decades ago, it would have been unheard of to use algorithms to count people in a crowd, generate meaningful text, or resolve conflicts, and if these experiments had been conducted then we would have likely observed results indicating algorithmic aversion across all tasks. Given the relentless pace of progress on algorithms, and increasingly large datasets on which to train, we expect the public will eventually embrace algorithmic assessments in decisions such as mixed-motive and creative tasks. The public have heard of many bad examples of inequity in mixed-motive or otherwise subjective tasks, such as hiring. However, like most facets of news, our attention is drawn more to failures than to successes. As the success stories permeate and become commonplace, the public acceptance will likely also grow.

There are several paths that could result in algorithms gaining acceptance in new task types. One way is that algorithms remain biased, but their biases are unknown or hidden. A more sanguine perspective is that future developers may be able to make less biased algorithms, or that the public is introduced to successful algorithms in new task types with lower stakes, such as through music recommendations. If less biased algorithms can be developed, then algorithms may be meaningfully embraced by society in mixed-motive tasks focusing on equity and fairness, potentially followed by acceptance in creative tasks.

References

Abeliuk, A., Benjamin, D. M., Morstatter, F., and Galstyan, A. 2020. "Quantifying Machine Influence Over Human Forecasters," *Scientific Reports* (10:15940).

Alexander, V. 2017. "Kanye West Rap Verses," *Kaggle*. (https://www.kaggle.com/viccalexander/kanyewestverses, accessed April 2, 2020).

Alexander, V., Blinder, C., and Zak, P. 2018. "Why Trust an Algorithm? Performance, Cognition and Neurophysiology," *Computers in Human Behavior* (89), pp. 279–288.

Ames, D., and Fiske, S. 2015. "Perceived Intent Motivates People to Magnify Observed Harms," *Proceedings of the National Academy of Sciences* (112:12), pp. 3599–3695.

Arnold, D., Dobbie, W., and Yang, C. 2018. "Racial Bias in Bail Decisions," *The Quarterly Journal of Economics* (133:4), pp. 1885–1932.

Asch, S. 1951. "Effects of Group Pressure Upon the Modification and Distortion of Judgments," in *Groups, Leadership, and Men*, H. Guetzkow (ed.), Pittsburgh: Carnegie Press, pp. 222–236.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Bonnefon, J.-F., and Rahwan, I. 2018. "The Moral Machine Experiment," *Nature* (59:64), pp. 59–64.

Baker, P. 2019. "'I Think This Guy Is, Like, Passed Out in His Tesla,'" *The New York Times Magazine*.

Bates, D., Mächler, M., Bolker, B., and Walker, S. 2014. "Fitting Linear Mixed-Effects Models Using Lme4," *ArXiv Preprint*.

Bauer, A., Eisenbeis, R., Waggoner, D., and Zha, T. 2003. "Forecast Evaluation with Cross-Sectional Data: The Blue Chip Surveys," *Economic Review* (88:2), pp. 17–31.

Becker, J., Brackbill, D., and Centola, D. 2017. "Network Dynamics of Social Influence in the Wisdom of the Crowds," *Proceedings of the National Academy of Sciences* (114:26), pp. E5070–E5076.

Benbasat, I., and Lim, L.-H. 1993. "The Effects of Group, Task, Context, and Technology Variables on the Usefulness of Group Support Systems: A Meta-Analysis of Experimental Studies," *Small Group Research* (24:4), pp. 430–462.

Berinsky, A., Margolis, M., and Sances, M. 2013. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self Administered Surveys," *American Journal of Political Science* (58:3), pp. 793–853.

Bickel, J. E., and Kim, S. D. 2008. "Verification of the Weather Channel Probability of Precipitation Forecasts," *American Meteorological Society* (136), pp. 4867–4881.

Bikhchandani, S., Hirshleifer, D., and Welch, I. 1992. "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades," *Journal of Political Economy1* (100:5), pp. 992–1026.

Blinder, A., and Morgan, J. 2005. "Are Two Heads Better than One? Monetary Policy by Committee," *Journal of Money, Credit, and Banking* (37:5), pp. 789–811.

Bonaccio, S., and Dalal, R. 2006. "Advice Taking and Decision-Making: An Integrative Literature Review,

and Implications for the Organizational Sciences," *Organizational Behavior and Human Decision Processes* (101:2), pp. 127–151.

Bonnefon, J.-F., Shariff, A., and Rahwan, I. 2016. "The Social Dilemma of Autonomous Vehicles," *Science* (352:6293), pp. 1573–1576.

Brin, S., and Page, L. 1998. "The Anatomy of a Large-Scale Hypertextual Web Search Engine," in *Proceedings of the Seventh International Conference on World Wide Web*, pp. 107–117.

Brown, N., and Sandholm, T. 2019. "Superhuman AI for Multiplayer Poker," *Science* (365:6456), pp. 885–890.

Buhrmester, M., Kwang, T. N., and Gosling, S. 2011. "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?," *Perspectives on Psychological Science* (6:1), pp. 3–5.

Burnham, K., and Anderson, D. 2004. "Multimodel Inference: Understanding AIC and BIC in Model Selection," *Sociological Methods & Research* (33:261).

Camerer, C., and Hogarth, R. 1999. "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework," *Journal of Risk and Uncertainty* (19), pp. 7–42.

Carlson, K., and Wu, J. 2012. "The Illusion of Statistical Control: Control Variable Practice in Management Research," *Organizational Research Methods2* (15:3), pp. 413–435.

Castelo, N., Bos, M., and Lehmann, D. 2019. "Task-Dependent Algorithm Aversion," *Journal of Marketing Research* (56:5), pp. 809–825.

Chafkin, M., and Verhage, J. 2018. "Betterment's Low-Fee Evangelist Has a Retirement Algorithm for You," *Bloomberg*. (https://www.bloomberg.com/news/features/2018-10-11/betterment-s-low-fee-evangelist-has-a-retirement-algorithm-for-yo, accessed March 12, 2020).

Chandler, J., Mueller, P., and Paolacci, G. 2014. "Nonnaïveté among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers.," *Behavioral Research Methods* (46), pp. 112–130.

Chandler, J., and Shapiro, D. 2016. "Conducting Clinical Research Using Crowdsourced Convenience Samples," *Annual Review of Clinical Psychology* (12), pp. 53–81.

Christensen, R. H. B. 2019. "A Tutorial on Fitting Cumulative Link Mixed Models with Clmm2 from the Ordinal Package," *R Vignette*. (https://cran.r-project.org/web/packages/ordinal/vignettes/clmm2_tutorial.pdf).

Connolly, T., Jessup, L., and Valacich, J. 1990. "Effects of Anonymity and Evaluative Ton on Idea Generation in Computer-Mediated Groups," *Management Science* (36:6), pp. 689–703.

Dance, A. 2015. "News Feature: How Online Studies Are Transforming Psychology Research: The Samples Are Large and Diverse, but Will This Trend Strengthen the Field or Merely Introduce New Sources of Error?," *Proceedings of the National Academy of Sciences* (112:47), pp. 14399–14401.

Dastin, J. 2018. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women," *Reuters*.

Dawes, R. 1979. "The Robust Beauty of Improper Linear Models in Decision Making," *American Psychologist* (34:7), p. 571.

Dawes, R., and Corrigan, B. 1974. "Linear Models in Decision Making," *Psychological Bulletin* (81:2), pp.

95–106.

DeBruine, L. 2002. “Facial Resemblance Enhances Trust,” *Proceedings of the Royal Society of Biological Sciences* (269:1498).

Dellarocas, C. 2003. “The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms,” *Management Science* (49:10), p. 2003.

Dennis, A. 1996. “Information Exchange and Use in Group Decision Making: You Can Lead a Group to Information, but You Can’t Make It Think,” *Management Information Systems Quarterly* (20:4), pp. 433–457.

Dennis, A., and Garfield, M. 2003. “The Adoption and Use of GSS in Project Teams: Toward More Participative Processes and Outcomes,” *Management Information Systems Quarterly* (27:2), pp. 289–323.

DeSanctis, G., and Gallupe, B. 1987. “A Foundation for the Study of Group Decision Support Systems,” *Management Science* (33:5), pp. 589–609.

Diab, D., Pui, S.-Y., Yankelevich, M., and Highhouse, S. 2011. “Lay Perceptions of Selection Decision Aids in U.S. and Non-U.S. Samples,” *International Journal of Selection and Assessment* (19:2), pp. 209–216.

Dietvorst, B., Simmons, J., and Massey, C. 2015. “Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err,” *Journal of Experimental Psychology: General* (144:1), pp. 114–126.

Dietvorst, B., Simmons, J., and Massey, C. 2016. “Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them,” *Management Science* (64:3).

Dobbie, W., Goldin, J., and Yang, C. 2018. “The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges,” *American Economic Review* (108), pp. 201–240.

Dockrill, P. 2017. “In Just 4 Hours, Google’s AI Mastered All The Chess Knowledge in History,” *Science Alert*.

Donnelly, L. 2019. “How I Became A New Yorker Cartoonist,” *New Yorker*. (https://www.newyorker.com/culture/culture-desk/how-i-became-a-new-yorker-cartoonist).

Dzindolet, M., Pierce, L., Beck, H., and Dawe, L. 2002. “The Perceived Utility of Human and Automated Aids in a Visual Detection Task,” *Human Factors* (44:1), pp. 79–94.

Eastwood, J., Snook, B., and Luther, K. 2011. “What People Want from Their Professionals,” *Journal of Behavioral Decision Making* (25:5), pp. 458–468.

Fama, E. 1970. “Efficient Capital Markets: A Review of Theory and Empirical Work,” *Journal of Finance* (25:2), pp. 383–417.

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. 2007. “G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences,” *Behavior Research Methods* (39:2), pp. 175–191.

Ferri, F. J., Pudil, P., Hatef, M., and Kittler, J. 1994. “Comparative Study of Techniques for Large-Scale

Feature Selection," *Machine Intelligence and Pattern Recognition* (16), pp. 403–413.

Field, H., and Lapowsky, I. 2020. "Coronavirus Is AI Moderation's Big Test. Don't Expect Flying Colors.," *Protocol*. (https://www.protocol.com/ai-moderation-facebook-twitter-youtube).

Gallupe, B., DeSanctis, G., and Dickson, G. 1988. "Computer-Based Support for Group Problem-Finding: An Experimental Investigation," *Management Information Systems Quarterly* (12:2), pp. 277–296.

Galton, F. 1907. "Vox Populi (The Wisdom of the Crowds).," *Nature* (75:1949), pp. 450–451.

Gelderman, M. 2002. "Task Difficulty, Task Variability and Satisfaction with Management Support Systems," *Information & Management* (39:7), pp. 593–604.

George, J., Easton, G., and Nunamaker, J. 1990. "A Study of Collaborative Group Work With and Without Computer-Based Support," *Information Systems Research* (1:4), pp. 394–415.

Gershgorn, D. 2019. "How Spotify's Algorithms Knows Exactly What You Want to Listen To," *OneZero*.

Gigerenzer, G., and Gaissmaier, W. 2011. "Heuristic Decision Making," *Annual Review of Psychology* (62), pp. 451–482.

Gino, F. 2008. "Do We Listen to Advice Just Because We Paid for It? The Impact of Advice Cost on Its Use," *Organizational Behavior and Human Decision Processes* (107:2), pp. 234–245.

Gino, F., and Moore, D. 2007. "Effects of Task Difficulty on Use of Advice," *Journal of Behavioral Decision Making* (20:1), pp. 21–35.

Gladwell, M. 2008. *Outliers*, Little, Brown and Company.

Goodhue, D., Lewis, W., and Thompson, R. 2017. "A Multicollinearity and Measurement Error Statistical Blind Spot: Correcting for Excessive False Positives in Regression and PLS," *Management Information Systems Quarterly* (41:3), pp. 667–684.

Grove, W., and Meehl, P. 1996. "Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures: The Clinical-Statistical Controversy," *Psychology, Public Policy, and Law* (2), pp. 293–323.

Grove, W., Zald, D., Lebow, B., Snitz, B., and Nelson, C. 2000. "Clinical versus Mechanical Prediction: A Meta-Analysis," *Psychological Assessment* (12:1), pp. 19–30.

Gruber, K. 2019. "Is the Future of Medical Diagnosis in Computer Algorithms?," *Lancet* (1), pp. 15–16.

Gueorguiva, R., and Krystal, J. 2004. "Move over ANOVA: Progress in Analyzing Repeated-Measures Data and Its Reflection in Papers Published in the Archives of General Psychiatry," *Archives of General Psychiatry* (61:3), pp. 310–317.

Guilbeault, D., Becker, J., and Centola, D. 2018. "Social Learning and Partisan Bias in the Interpretation of Climate Trends," *Proceedings of the National Academy of Sciences* (115:39), pp. 9714–9719.

Harvey, N. 1997. "Confidence in Judgment," *Trends in Cognitive Sciences* (1:2), pp. 78–82.

Harvey, N., and Fischer, I. 1997. "Taking Advice: Accepting Help, Improving Judgment, and Sharing Responsibility," *Organizational Behavior and Human Decision Processes* (70:2), pp. 117–133.

Harvey, N., Harries, C., and Fischer, I. 2000. "Using Advice and Assessing Its Quality," *Organizational*

Behavior and Human Decision Processes (81:2), pp. 252–273.

Hickey, A. 2019. "How Coffee Meets Bagel Leverages Data and AI for Love," *CIODIVE*. (https://www.ciodive.com/news/coffee-meets-bagel-dating-technology-ai-data/548395/#:~:text=The company's matching algorithm runs,a day to decide on., accessed March 12, 2020).

Idrees, H., Saleemi, I., Seibert, C., and Shah, M. 2013. "Multi-Source Multi-Scale Counting in Extremely Dense Crowd Images," in *IEEE International Conference on Computer Vision and Pattern Recognition*.

Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maddeed, S., Rajpoot, N., and Shah, M. 2013. "Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds," in *Proceedings of IEEE European Conference on Computer Vision*, Munich, Germany.

Janis, I. 1972. *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes*, Boston: Houghton, Mifflin.

Johnson, E., Payne, J., and Bettman, J. 1988. "Information Displays and Preference Reversals," *Organizational Behavior and Human Decision Processes* (42), pp. 1–21.

Johnson, and Payne. 1985. "Effort and Accuracy in Choice," *Management Science* (31:4), pp. 394–414.

Kageki, N. 2012. "An Uncanny Mind: Masahiro Mori on the Uncanny Valley and Beyond," *IEEE Spectrum*.

Kahneman, D. 2011. *Thinking, Fast and Slow*, (1st ed.), Farrar, Straus and Giroux.

Kahneman, D., and Tversky, A. 1972. "Subjective Probability: A Judgment of Representativeness," *Cognitive Psychology* (3:3), pp. 430–454.

Kahneman, D., and Tversky, A. 1974. "Judgment under Uncertainty: Heuristics and Biases," *Science* (185:4157), pp. 1124–1131.

Kahneman, D., and Tversky, A. 1979. "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica* (Vol. 47).

Kamps, H. J. 2016. "Microsoft Demos Next-Generation Image-Captioning Captionbot," *Techcrunch*. (https://techcrunch.com/2016/03/30/microsoft-caption-bot/, accessed May 2, 2020).

Karau, S., and Williams, K. 1993. "Social Loafing: A Meta-Analytic Review and Theoretical Integration," *Journal of Personality and Social Psychology* (65:4), pp. 681–706.

Kawaguchi, K. 2020. "When Will Workers Follow an Algorithm? A Field Experiment with a Retail Business," *Management Science* (67:3), pp. 1670–1695.

Kennedy, P. 2008. *A Guide to Econometrics*, (6th ed.), Cambridge, Mass: MIT Press.

Keppel, G. 1991. *Design and Analysis: A Researcher's Handbook*, Prentice-Hall Inc.

Kerras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. 2019. "Analyzing and Improving the Image Quality of StyleGAN," *ArXiv*.

Klayman, J., Soll, J., González-Vallejo, and Barlas, S. 1999. "Overconfidence: It Depends on How, What, and Whom You Ask," *Organizational Behavior and Human Decision Processes* (79:3), pp. 216–247.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. 2017. "Human Decisions and Machine Prediction," *Quarterly Journal of Economics* (133:1), pp. 237–293.

Komiak, S., and Benbasat, I. 2008. "A Two-Process View of Trust and Distrust Building in Recommendation Agents: A Process-Tracing Study," *Journal of the Association for Information Systems* (9:12), pp. 727–747.

Kruger, J., and Dunning, D. 1999. "Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments," *Journal of Personality and Social Psychology* (77:6), pp. 1121–1134.

Lakkaraju, H., Kleinberg, J., Leskovec, J., and Ludwig, J. 2017. "The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, Canada, pp. 275–284.

Laughlin, P. 2011. *Group Problem Solving*, Princeton, NJ: Princeton University Press.

Lazer, D., Baum, M., Benkler, Y., Berinsky, A., Greenhill, K., Menczer, F., Metzger, M., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S., Sunstein, C., Thorson, E., Watts, D., and Zittrain, J. 2018. "The Science of Fake News," *Science* (359:6380), pp. 1094–1096.

Liberman, V., Minson, J., Bryan, C., and Ross, L. 2012. "Naive Realism and Capturing the 'Wisdom of Dyads,'" *Journal of Experimental Social Psychology* (48:2), pp. 507–512.

Logg, J., Haran, U., and Moore, D. 2018. "Is Overconfidence a Motivated Bias? Experimental Evidence," *Journal of Experimental Psychology: General* (147:10), pp. 1145–1465.

Logg, J., Minson, J., and Moore, D. 2019. "Algorithmic Appreciation: People Prefer Algorithmic to Human Judgment," *Organizational Behavior and Human Decision Processes* (151), pp. 90–103.

Lorenz, J., Rauhut, H., Schweitzer, F., and Helbing, D. 2011. "How Social Influence Can Undermine the Wisdom of Crowd Effect," *Proceedings of the National Academy of Sciences* (108:22), pp. 9020–9025.

MacGillis, A. 2019. "The Case Against Boeing," *The New Yorker*.

Malkiel, B. 1973. *A Random Walk Down Wall Street*, W W. Norton & Company.

Mannes, A. 2009. "Are We Wise about the Wisdom of Crowds? The Use of Group Judgments in Belief Revision," *Management Science* (55:8), pp. 1267–1279.

McCormick, D. 2014. "Virtual Tween Passes Turing Test," *IEEE Spectrum*. (https://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/virtual-tween-passes-turing-test).

McGrath, J. 1984. *Groups: Interaction and Performance*, Upper Saddle River, NJ: Prentice Hall.

Mednick, S. A. 1968. "The Remote Associates Test," *The Journal of Creative Behavior* (2:3), pp. 213–214.

Meehl, P. 1954. *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*, Minneapolis, Minn. : University of Minnesota Press.

Milgram, S., Bickman, L., and Berkowitz, L. 1969. "Note on the Drawing Power of Crowds of Different Size," *Journal of Personality and Social Psychology* (13:2), pp. 79–82.

Minson, J., and Mueller, J. 2012. "The Cost of Collaboration: Why Joint Decision Making Exacerbates Rejection of Outside Information," *Psychological Science* (23:3), pp. 219–224.

Mlot, S. 2017. "AI Beats George R.R. Martin to Writing 'GoT' Book Six," *Geek*. (https://www.geek.com/tech/ai-beats-george-r-r-martin-to-writing-got-book-six-1714350/, accessed April 2, 2020).

Moore, D., and Healy, P. 2008. "The Trouble with Overconfidence," *Psychological Review* (115:2), p. 502.

Moravec, P., Minas, R., and Dennis, A. 2019. "Fake News on Social Media: People Believe What They Want to Believe When It Makes No Sense at All," *Management Information Systems Quarterly* (43:4), pp. 1343–1360.

Moray, N., and Lee, J. 1994. "Trust, Self Confidence, and Operators' Adaptation to Automation," *International Journal of Human-Computer Studies* (40:1), pp. 153–184.

Mosier, K., and Skitka, L. 1996. "Human Decision Makers and Automated Decision Aids: Made for Each Other?," *Automation and Human Perforamcne: Theory and Applications*.

Nickerson, R. 1998. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises," *Review of General Psychology* (2:2), pp. 175–220.

Noble, S. U. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*, NYU Press.

Paese, P., and Sniezek, J. 1998. "Influences on the Appropriateness of Confidence in Judgment: Practice, Effort, Information, and Decision-Making," *Organizational Behavior and Human Decision Processes* (48:1), pp. 100–130.

Paolacci, G., and Chandler, J. 2014. "Inside the Turk: Understanding Mechanical Turk as a Participant Pool," *Current Directions in Psychological Science* (23:3), pp. 184–188.

Parasuraman, A., and Colby, C. 2015. "An Updated and Streamlined Technologies Readiness Index: TRI 2.0," *Journal of Service Research* (18:1), pp. 59–74.

Parasuraman, R., and Dietrich, M. 2010. "Complacency and Bias in Human Use of Automation: An Attentional Integration," *Human Factors: The Journal of Human Factors and Ergonomics Society* (52:3), pp. 381–410.

Parasuraman, R., and Riley, V. 1997. "Use, Misuse, Disuse, and Abuse.," *Human Factors: The Journal of Human Factors and Ergonomics Society* (39:2), pp. 230–253.

Payne, J., Bettman, J., and Johnson, E. 1993. *The Adaptive Decision Maker*, New York, New York: Cambridge University Press.

Petter, S., DeLone, W., and McLean, E. 2013. "Information Systems Success: The Quest for the Independent Variables," *Journal of Management Information Systems* (29:4), pp. 7–61.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., and Wallach, H. 2018. *Manipulating and Measuring Model Interpretability*. (http://arxiv.org/abs/1802.07810).

Promberger, M., and Baron, J. 2006. "Do Patients Trust Computers?," *Journal of Behavioral Decision Making* (19:5), pp. 455–468.

Qiu, L., and Benbasat, I. 2009. "Evaluating Anthropomorphic Product Recommendation Agents: A Social Relationship Perspective to Designing Information Systems," *Journal of Management Information*

Systems (25:4), pp. 145–181.

Rai, A. 2020. "Explainable AI: From Black Box to Glass Box," *Journal of the Academy of Marketing Science* (48:1), pp. 137–141.

Schaeffer, J., Burch, N., Björnsson, Y., Kishimoto, A., Müller, M., Lake, R., Lu, P., and Sutphen, S. 2007. "Checkers Is Solved," *Science* (317:5844), pp. 1518–1522.

Schwartz, L. M., Woloshin, S., Black, W. C., and Welch, H. G. 1997. "The Role of Numeracy in Understanding the Benefit of Screening Mammography," *Annals of Internal Medicine* (127:11), pp. 966–972.

Servan-Schreiber, E., Wolfers, J., Pennock, D. M., and Galebach, B. 2004. "Prediction Markets: Does Money Matter?," *Electronic Markets* (14:31), pp. 243–251.

Sheather, S. 2009. *A Modern Approach to Regression with R*, New York, New York: Springer.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. 2017. "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm," *ArXiv*.

Silver, N. 2012. *The Signal and the Noise*.

Simmons, J., Nelson, L., and Simonsohn, U. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science* (22:11), pp. 1359–1366.

Simon, H. 1956. "Rational Choice and the Structure of the Environment," *Psychological Review* (63:2), pp. 129–138.

Simon, H. 1979. "Rational Decision-Making in Business Organizations," *American Economic Review* (69), pp. 493–513.

Simourd, D. 1997. "The Criminal Sentiments Scale-Modified and Pride in Delinquency Scale: Psychometric Properties and Construct Validity of Two Measures of Criminal Attitudes," *Criminal Justice and Behavior* (24:1), pp. 52–70.

Sniezek, J., and Buckley, T. 1995. "Cueing and Cognitive Conflict in Judge-Advisor Decision Making," *Organizational Behavior and Human Decision Processes* (62:2), pp. 159–174.

Sniezek, J., and Van Swol, L. 2001. "Trust, Confidence, and Expertise in a Judge-Advisor System," *Organizational Behavior and Human Decision Processes* (84:2), pp. 288–307.

Soll, J., and Mannes, A. 2011. "Judgmental Aggregation Strategies Depend on Whether the Self Is Involved," *International Journal of Forecasting* (27:1), pp. 81–102.

Sparrow, B., Liu, J., and Wegner, D. 2011. "Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips," *Science* (333:6043), pp. 776–778.

Staffelbach, M., Semploinski, P., Kijewski-Correa, T., Thain, D., Wei, D., Kareem, A., and Madey, G. 2015. "Lessons Learned from Crowdsourcing Complex Engineering Tasks," *PLOS ONE* (10:9).

Stevenson, M. 2017. "Assessing Risk Assessment in Action," *Minnesota Law Review* (103:303).

Straus, S. 1999. "Testing a Typology of Tasks: An Empirical Validation of McGrath's (1984) Group Task

Circumplex," *Small Group Research* (39:2), pp. 166–187.

Strunk, W., and White, E. 2005. *The Elements of Style*, (3rd ed.), Penguin Books.

Surowiecki. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*, Doubleday;Anchor.

Sutskever, I., Martens, J., and Hinton, G. 2011. "Generating Text with Recurrent Neural Networks," in *Proceedings of the 28th International Conference on Machine Learning*, pp. 1017–1024.

Tan, B., Wei, K.-K., Watson, R., Clapper, D., and McLean, E. 1998. "Computer-Mediated Communication and Majority Influence: Assessing the Impact in an Individualistic and a Collectivistic Culture," *Management Science* (44:9), pp. 1263–1278.

Taylor, S. 2014. *The Tending Instinct: Women, Men, and the Biology of Nurturing*, Times Books.

Tewari, U. 2019. "Game of Thrones Episode Script Generation Using LSTM and Recurrent Cells in Tensorflow," *Medium*. (https://medium.com/analytics-vidhya/game-of-thrones-episode-script-generation-using-lstm-and-recurrent-cells-in-tensorflow-c0c40d415a8b, accessed April 2, 2020).

"Thispersondoesnotexist.Com." 2019. (thispersondoesnotexist.com, accessed May 2, 2020).

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society* (58:1), pp. 267–288.

Tiwana, A., and Kim, S. 2019. "From Bricks to an Edifice: Cultivating Strong Inference in Information Systems Research," *Information Systems Research* (30:3), pp. 1029–1036.

Todd, P., and Benbasat, I. 1999. "Evaluating the Impact of DSS, Cognitive Effort, and Incentives on Strategy Selection," *Information Systems Research* (10:4), pp. 356–374.

Torbet, G. 2019. "Chrome Will Use AI to Describe Images for Blind and Low-Vision Users," *Engadget*. (https://www.engadget.com/2019/10/10/chrome-image-descriptions-accessibility/, accessed May 2, 2020).

Van de Ven, A. 2007. *Engaged Scholarship: A Guide for Organizational and Social Research.*, New York: Oxford University Press.

Wallace, D. F. 1996. *Infinite Jest*, Little, Brown and Company.

Wang, W., and Benbasat, I. 2007. "Recommendation Agents for Electronic Commerce: Effects of Explanation Facilities on Trust Beliefs," *Journal of Management Information Systems* (23:4), pp. 217–246.

Wang, W., and Benbasat, I. 2008. "Attributions of Trust in Decision Support Technologies: A Study of Recommendation Agents for e-Commerce," *Journal of Management Information Systems* (24:4), pp. 249–273.

Wickham, H. 2016. *Elegant Graphics for Data Analysis*, New York: Springer-Verlag.

Xiao, B., and Benbasat, I. 2007. "E-Commerce Product Recommendation Agents: Use, Characteristics, and Impact," *Management Information Systems Quarterly* (31:1), pp. 137–209.

Xu, J., Benbasat, I., and Cenfetelli, R. 2014. "The Nature and Consequences of Trade-off Transparency in the Context of Recommendation Agents," *Management Information Systems Quarterly* (38:2), pp.

379–406.

Yaniv, I., and Kleinberger, E. 2000. “Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation,” *Organizational Behavior and Human Decision Processes* (83:2), pp. 260–281.

Yao, Y., Viswanath, B., Cryan, J., Zhang, H., and Zhao, B. 2017. “Automated Crowdturfing Attacks and Defenses in Online Review Systems,” in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1143–1158.

Yeomans, M., Shah, A., Mullainathan, S., and Kleinberg, J. 2019. “Making Sense of Recommendations,” *Journal of Behavioral Decision Making* (32:4), pp. 403–414.

Zajonc, R. 2001. “Mere Exposure: A Gateway to the Subliminal,” *Current Directions in Psychological Science* (10:6), pp. 224–228.

Zigurs, I., and Buckland, B. 1998. “A Theory of Task/Technology Fit and Group Support Systems Effectiveness,” *Management Information Systems Quarterly* (22:3), pp. 313–334.

Zuckerman, G. 2019. *The Man Who Solved the Market*, New York, NY: Portfolio Penguin.

Appendix 1:

7.1Kanye West RNN Generated Lyrics

Right here, history
Where the day it's face in. I really want some Marie weed
Hotel to aton will reach
If your cousing off on my free
You can't even though too much on the old flowed

7.2George R.R. Martin RNN Generated Lyrics

"'I feared Master Sansa, Ser,' Ser Jaime reminded her. "She Baratheon is one of the crossing. The second sons of your onion concubine.'"

(Mlot 2017)

Appendix 2: Attention Check

This attention check was used in (Yeomans et al. 2019). Participants who failed the attention check will not be not counted in our recruitment totals and were not allowed to complete the survey.

First, tell us about yourself!

To help us understand how people think about different activities, please answer this question correctly. Specifically, we are interested in whether you actually take the time to read the directions; if not, the results would not be very useful. To show that you have read the instructions, please ignore the items below about activities and instead type 'I will pay attention' in the space next to 'Other'. Thank you.

[ ] Watching Athletics

[ ] Attending Cultural Events

[ ] Participating in Athletics

[ ] Reading Outside of Work or School

[ ] Watching Movies

[ ] Travel

[ ] Religious Activities

[ ] Needlework

[ ] Cooking

[ ] Gardening

[ ] Computer Games

[ ] Hiking

[ ] Board or Card Games

[ ] Other: _____

Appendix 3: Manipulation check

This was used at the end of the study to confirm that participants had processed the information they were given about the source of the recommendations they had received.

_____

Answer a quick question about the experiment you just took part in, to make sure you were paying attention. What was the source of the recommendations in your study?

[ ] An algorithm

[ ] An average of human estimates

[ ] Didn't specify

Appendix 4: Bail Scenario

Charge: Armed habitual criminal and aggravated battery

State Attorney's reasons for high bail:

- Routine traffic stop
- Officers observed defendant put something in someone else's lap
- Officers did a pat down of defendant and felt a gun
- Officers tried to cuff defendant, who pushed officer in chest while fist was balled up
- There was a struggle and all officers fell to ground

- Officers suffered scrapes and lacerations
- Officers recovered a gun from defendant
- Someone else in car had 9 grams of cocaine
- Defendant was discharged from parole 4 days ago
- Defendant has 4 prior felony convictions (2015 possession of firearm by gang member, 2015 aggravated fleeing, 2011 unlawful use of weapon by felon, 2006 aggravated criminal sex abuse)
- Loaded weapon found on person

Public Defender's reasons for low bail:

- Defendant is in his 30s
- Lifelong resident of the county
- Lives with a friend for last 6 days
- Has a child who is 5 years old
- Has GED
- Has certificate from local college
- Also has asthma, which could make spending time in prison complicated because of COVID
- Can post $500 towards bail