Mortal Versus Machine: Investigating Interpersonal Advice and Automated Advice

By

Andrew Prahl

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Communication Arts)

at the

UNIVERSITY OF WISCONSIN-MADISON

2018

Date of final oral examination: 04/23/2018

This dissertation is approved by the following members of the Final Oral Committee:

Lyn M Van Swol, Professor, Communication Arts
Michael Xenos, Professor, Communication Arts
Zhongdang Pan, Professor, Communication Arts
Douglas Wiegmann, Professor, Industrial and System Engineering
Daniel Bolt, Professor, Educational Psychology

Dedicated to my loving and inspiring family: Mom, Dad, Coco & Molly

Acknowledgements:
My endlessly helpful advisor: Lyn Van Swol
My supportive committee members: Dan Bolt, Zhongdang Pan, Doug Wiegmann & Michael Xenos
The inspiring educators throughout my life: Roger Bass, Sal Khan, Ralph Kline & Kevin McMahon
My fellow graduate students who were always there when the going was tough: Paul Ahn, Michael Braun, Cassandra Carlson, Mina Choi, Larisa Doroshenko, Valerie Kretz, Sangwon Lee & Jihyun Paik
My friends who nudged me on my way to and through graduate school: Chase Caster, Julie Collins, Jessica Demarre Asay, Sabrina Gogol, Jeanine Hemel, Isaac Hoffer, Yukon Jenkins, Amanda Keppler, Lindsay Pour, Caleb, Charlie and Linda Sattgast & Sarah Thompson
My encouraging family members: Brian, Caroline, Lachlan & Logan
And those who helped me in this research: John D Judge & Allison Sattgast

Table of Contents

Abstract:

Chapter I: Introduction 1

Chapter II: Study I: Understanding Algorithm Aversion: When is Advice from Automation Discounted? 25

Chapter III: Study II: People and the Machines: Advice Utilization when a Human Advisor is Replaced by Automation 26

Chapter IV: Study III: Utilized but Despised: Automated Advice versus Human Advice in High and Low Demonstrability Tasks 60

Chapter V: Summary & Conclusion 108

Appendix I: Supplemental Analysis 154

Appendix II: Figures & Tables 163

Abstract

A series of studies are conducted to investigate differences in trust and advice processes in human-human versus human-automation communication. Study I investigates phenomenon that lead to the discounting of automated advice. Study II introduces the manipulation of advisor replacement, such as a human being replaced by automation. Study III manipulates the decision-making task. Results show that human-decision makers are generally less happy receiving advice from automated advisors and perceive automated advisors as lower quality. Furthermore, humans judge automated advisors more harshly than human advisors after the issuance of bad advice. Additionally, the replacement of a human advisor by an automated advisor is found to exaggerate these effects. Results are situated in several theoretical lenses and used to inform emergent theory comparing mortals versus machines.

Chapter I: Introduction

The history of labor is one that is intertwined with the history of technology and automation. Over time, innovations in technology have boosted labor productivity and created new jobs as automation has taken many physical labor demands away from humans, allowing them to dedicate themselves to more complex work (Noble, 2017). But, due to the rapidly increasing capability of automation, many speculate that a large amount human labor in more specialized domains such as law and medicine will be entirely replaced, leaving an unprecedented number of humans permanently out of work (Ford, 2015). Such concerns were put succinctly at an event held just last month, where Israeli labor historian Yuval Noah Harari said that a coming age of algorithms and other forms of automation risked creating a "global useless class," because we will have created machines that are simply, "better than us [humans]" (Freytas-Tamura, 2018, p. 5).

Many clear examples of the rapid advance of automation are visible in our lives every day, and not just in the workplace. In the early 21st century, attempts to develop self-driving cars were of limited success due to the complexity of perception and judgement needed to operate a vehicle (Behringer et al., 2004); a fully autonomous system seemed decades away. Yet, since 2015 the automaker Tesla has been offering their "Autopilot" system to as an option on all of their vehicles. Although Tesla does not claim this system is fully autonomous, it only took a few months before a Tesla vehicle, with autopilot engaged, drove into the side of a commercial truck at highway speed, killing the driver. Just a three weeks ago, a similar fatal accident occurred as a vehicle with autopilot engaged drove into a highway barrier at over 60 miles per hour. Preliminary reports from both of these accidents indicate a dangerous human tendency: despite numerous warnings, neither driver hand their hands on the wheel or took action to avoid the

accidents even though they would have had more than enough time to react had they been alert (Habib, 2017; Siddiqui, 2018).

Driving a vehicle is obviously a complex task that involves attention, perception, coordination, and judgement. But so does piloting an airplane, and the aviation industry has been struggling for decades now with the issue of automation reliance, the phenomenon where humans lose their own skill and are unable to perform their jobs when automation fails (Geiselman, Johnson, & Buck, 2013; Mehta, Rice, & Winter, 2014). Automation reliance is also found in other domains such as industrial operations (J. E. Lee & See, 2004), and the Tesla accidents foreshadow a future where normal people may become reliant on automation to do things like simply drive to work. Reliance is not the only reason not everyone is excited about automation taking over roles filled by humans. For example, with autonomous vehicles, many scholars in fields such as law and ethics are asking questions about how automation should respond when faced with a moral decision (Hevelke & Nida-Rümelin, 2014). Beyond vehicles, there are other areas where automation is advancing where the thought of an automated decision-maker is met with skepticism. For example, in healthcare, some scholars are asking if an automated system can truly understand the subjective aspects of a decision such as quality of life and pain management (Inthorn, Tabacchi, & Seising, 2015). Alternatively, it is it really "fair" to have an automated system decide what level of care someone is deserves to receive (Jha, 2012)? The advance of automation and replacement of human labor continues to elicit varying reactions from scholars and the general public, many of which are observable on op-ed pages every day. Perhaps people are right to be scared about a "robot future" with no work for them – or perhaps these concerns are overly pessimistic and increasing automation will raise the quality of life around the world (Naastepad & Mulder, 2018).

Regardless of what the future holds, one thing is certain: humans will find themselves working increasingly with automation in both personal and professional life in the near future. Even if machines are not fully autonomous, humans will frequently consult automation for advice when in a decision-making situation. But, while seeking automated advice has largely been constrained to decisions suited to mathematical and statistical judgment since the advent of computing, the use of artificial intelligence and machine learning in domains such as healthcare and disaster relief (Garattini, Raffle, Aisyah, Sartain, & Kozlakidis, 2017; Moor, 2006) suggest that machines will increasingly help us make decisions with more subjective consequences. Besides some preliminary investigations, the scholarly community currently lacks understanding of how humans may react to automation entering more subjective domains, even in just an advisory role. There is also a lack of research on how humans are likely to use new automated advisors if they replace human advisors.

The purpose of the research below is to conduct research into both of these questions. I present three studies that address differing aspects of the questions above but share enough common elements such that they can be compared. The primary goal of study 1 is to study the behavior of advice utilization from automated advisors compared to human advisors. Study 2 build upon the results of study 1 to understand the effects of advisor replacement. Finally, study 3 also reports utilization behavior, but primarily investigates the perception of human and automated advice in different types of decisions. Before presenting the studies, we provide a general conceptual overview and review of theoretical guidance from the fields of interpersonal communication and human-automation trust. We start by (1) discussing automation as an advisor, (2) conceptualizing trust and advice-utilization, (3) reviewing interpersonal and

automation advice theory, (4) presenting theoretical guidance to compare human and automated advice, and (5) discussing task demonstrability.

Automation as an Advisor

An important interpersonal process affected by the advance of automation is advice. The solicitation and exchange of advice is typically thought of as between humans, but automated advisors are becoming more common and they are replacing human advisors. For example, algorithmic methods for predicting stock performance have been in use since some of the first computers were commercially available (Alexander, 1995), but only recently have they been thought of as culpable for some turbulence in global financial markets . Currently, automated advisors are being developed for use in a variety of contexts, such as IBM's Watson computing platform being used to help diagnose cancer and develop treatment plans (Cha, 2015). In personal life, algorithms on dating websites advise people on their best match, and virtual agents are advising people on how to troubleshoot IT problems at home (Li, Kizilcec, Bailenson, & Ju, 2016).

Advice has long been studied by interpersonal scholars, producing a substantial body of research (Bonaccio & Dalal, 2006; Yaniv & Kleinberger, 2000). In fact, advice exchange has been identified as one of the primary social processes through which rapport and trust develops between people (MacGeorge, 2016). Given that advice is a social process, a natural question emerges regarding automated advisors – if automation is not human and not social, is it really advice? I believe the answer is yes for two primary reasons. First of all, advice is not necessarily a social process. If someone is looking for advice on what to pack for a summer-long hike of the Pacific Crest Trail, they might read passages from the bestselling novel *Wild: From Lost to Found on the Pacific Crest Trail* (Strayed, 2012), even though the author is not directly

communicating advice to the decision-maker. Alternatively, an employee might seek advice for how to perform a new job duty and consult an "advice for new team members section" in the organizational manual – without an author name listed, the decision-maker has nobody to even feel thankful to for the advice. A large amount of interpersonal advice research has been conducted in such a context where the process of advice is studied in absence of a decision-maker knowing that the advisor intended to advise them personally, such as when a decision maker is supplied a peer's answers to mathematical problems as a form of advice (Van Swol, MacGeorge, & Prahl, 2015).

Second, even if advice is conceptualized as strictly social, there exists a large body of research showing that humans often respond socially to machines and automation. Human-human social scripts are activated (sometimes unconsciously) by interacting with automation (Duffy, 2003; Nass & Moon, 2000) even though humans logically know the machine is not a conscious being. The tendency to socialize machines has is found to extend across various levels of cognitive development and can even result in negative feelings such as guilt and betrayal (Kahn et al., 2012, 2015). Even when machines are not designed to be human, but simply some other living creature like a pet dog, humans can develop intense bonds and feelings of care for the automation (Friedman, Kahn, & Hagman, 2003). Finally, even in the absence of any replicated social cues, humans have been known to anthropomorphize automation (such as bomb disposal robots) to such an extent that they are willing to stand in harm's way to protect it and hold funerals when automation is irreparably broken (Carpenter, 2013).

Both of these reasons for envisioning automation as an advisor have important implications for the study of automated versus human advice. First, when a human receives solicited or unsolicited advice, they must decide whether to use that advice or not. Even in

absence of a social process, decision-makers may consider the usefulness of automated advice differently than they do human advice. We expand further on this question as we discuss perceptions of advisors and performance expectations below. Second, if an automated advisor is responded to as a social actor by a decision-maker, do their feelings differ (compared to receiving advice from a human) given that logically, the decision-maker knows that automation is not human? Both the behavioral and perceptual questions are of great importance as automation replaces humans in organizations. For example, research on layoffs shows that the psychological reaction of workers that "survive" is a key determinant of the success of a company as it moves forward after a period of downsizing (van Dierendonck & Jacobs, 2012). Even when downsizing is beneficial to firm performance, the process can lead to lower organizational commitment and general morale among survivors (Baumol, Blinder, & Wolff, 2005). Thus, a question emerges: what happens when people who interact with a human see that human replaced by automation? There exists a large amount of research on human and automated advice, but none has directly addressed this question of replacement in a controlled experimental setting. To understand the past research that has been conducted on both advisor types, it is helpful to understand the relationship between trust and advice utilization; we explicate this relationship below.

Trust

Trust is a concept that is larger than advice-utilization alone. Trust is a fundamental aspect of social life and interpersonal communication, thus, it has long been a subject of scholarly research in a wide variety of fields from philosophy and law to social psychology and communication (Butler, 1991; Jones & George, 1998; Schoorman, Mayer, & Davis, 2007). This has led to a seemingly innumerable number of conceptualizations of trust. For example, trust has

been defined as simply a set of expectations (Barber, 1983) or as an effect of social comparison (Pruitt & Rubin, 1986). In some organization research, trust is conceptualized as the degree to which one party finds another predictable (Zaltman & Moorman, 1988), or an agreement of a mutual set of goals – a definition that is closer to that of cooperation rather than trust (Mayer, Davis, & Schoorman, 1995). Trust has also been operationalized in a number of different ways, blurring the distinction between trust being a behavioral outcome or a mental state. However, some of the most influential conceptualizations of trust and a number of review articles have defined trust as a mental state that reflects a trustor (decision-maker's) willingness to take risk with a trustee (advisor) (Hoff & Bashir, 2015; Mayer et al., 1995).

The large numbers of theoretical models that consider trust to be a mental state share several commonalities. Most importantly from an advice point of view, the trust process begins with an assessment of the credibility (or, trustworthiness) of the advisor. Different models have identified different aspects of trustworthiness. (Mayer et al., 1995) for example, describe trustworthiness as being composed of a decision-maker's assessment of the advisor's ability, benevolence and integrity. Ability, which refers to the advisor's skills and performance capability, is an element present in nearly every model of trust. Benevolence refers to the perception that the advisor has good intentions, because even an advisor with great ability may not be trusted if their intentions are to hurt the decision-maker. Lastly, integrity refers to a similarity in accepted values and principles between the advisor and decision-maker. Integrity is perhaps easier to understand in the context of a moral decision because an advisor of high ability and good intention may not see the moral consequences of decision similarly to the decision-maker. Two important moderators of trustworthiness assessments are the situational context (i.e., a match between the advisor's ability and the demands of the decision-making task) and the

advisor's self-concept because assessments of an advisor involve comparisons with one's own ability and values (Sitkin & Roth, 1993). The result of the assessment of trustworthiness is what leads to the mental state of trust and finally, trusting behavior (Colquitt, Scott, & LePine, 2007; Sitkin & Pablo, 1992).

The aspect of risk is critical to the conceptualization of trust as it relates to the concept of advice. Interpersonal advice research has frequently operationalized trust as the degree to which a decision-maker shifts towards the recommendation of an advisor from a decision-maker's initial estimate (Bonaccio & Dalal, 2006); in essence the decision-maker or judge is making an assessment of trustworthiness and choosing to expose themselves to the risk that the advice is worse than their own estimate. The judge-advisor system is a widely used example of such an operationalization of trust (Sniezek & Buckley, 1995; Sniezek & Van Swol, 2001), and a large amount of research conducted in both the forecasting and human-automation advice fields utilizes similar experimental methods to measure trust. In this literature and the majority of literature studying the utilization of advice, trust is frequently used interchangeably with advice-utilization. However, this introduces a tension between conceptualization and operationalization because although trust is often conceptualized as a mental state, trust measurement is typically achieved through the analysis of a behavioral outcome (advice-utilization). For the purposes of the research here, we believe that advice-utilization represents the most precise representation in terms of the behavior we wish to study because we will be measuring the degree to which a decision-maker shifts an initial (self-generated) estimate after the receive advice from an advisor. Our measure of advice-utilization is identical to many past measures that have been used as operationalizations of trust, but because a mental state and behavior can differ, our measure of advice-utilization can be understood as a measure of trusting behavior (or "risk taking behavior,"

(Mayer et al., 1995)). This conceptualization of the relationship between trust and advice-utilization allows results from the studies presented here to be compared to the wide variety of studies that measure the concept of "trust" similarly and the equally numerous studies that investigate the same process but call it "advice-utilization" (Bonaccio & Dalal, 2006). In the following sections I use "advice-utilization" in favor of "trust" when discussing the many studies that use the terms interchangeably. Similarly, I use the term "decision-maker" in place of trustor and "advisor" in place of trustee. Additionally, for the studies which discuss trust as a perception or feeling, I am specific in what perception is being discussed because trust is a result of the many perceptions involved in assessing trustworthiness (e.g., perceiving benevolence and integrity).

In addition to providing a connection between trust and advice-utilization, Mayer et al., (1995) explain how the concept of risk also establishes boundary conditions of trust that I extend to the construct of advice-utilization in the studies below. First, in order for risk to be present, there the decision-maker must have a choice to utilize advice. Thus, advice-utilization is not measurable in situations in which a decision-maker feels they have no option but to utilize advice, such as a life and death medical decision that must be made in a short time and one receives advice from a doctor. Additionally, the presence of risk implies that advice-utilization occurs when the decision-maker actually has something to lose if the advice is worse than their own estimate. In an organizational situation for example, there may be control structures in place (such as a supervisor that is favorable to one employee but not others) that would punish someone for issuing bad advice to the decision-maker if the outcome of a decision is not optimal. In this instance, the decision-maker takes no risk by utilizing advice, why would they choose *not* to use advice? The presence of risk has design implications for the experiments presented below.

First, our experiments are designed such that the participant has a clear choice (is not forced) to use advice. Additionally, good performance on the tasks is incentivized such that the participant will consider the potential risk of utilizing advice versus their own judgment.

Interpersonal and Automated Advice Theory

Interpersonal Advice

Advice as an interpersonal process has been an interest of a wide variety of scholars in fields such as communication, psychology and management (Bonaccio & Dalal, 2006; MacGeorge & Van Swol, 2018). People frequently seek the input of others to help them make decisions in their personal and professional lives. Even when people do not seek the advice of others, it is sometimes delivered unsolicited by people trying to help others who are in challenging situations (Feng & Magen, 2016). Because the communication of advice is such a frequent interpersonal process, a large amount of research has resulted in several paradigms and numerous theoretical models (MacGeorge, 2016).

The psychological paradigm is one that is primarily concerned with the utilization of advice (MacGeorge, 2016). Scholars in this area see advice as a means to making a better decision, and a large number of studies have investigated both what leads a decision-maker to utilize advice and if the utilization led to better decisions. Perceptions of advisor ability are critical in this line of research because it is one of the primary drivers of utilization behavior (Yaniv, 2004). Perceptions of advisor ability have been studied as perceptions of advisor expertise, competence and skill (among other constructs) but they all fall under the general idea that a decision-maker perceives the advisor to have the sufficient ability to supply advice that is better than the decision-maker's judgment alone. In this way, the study of advice-utilization, as outlined earlier, largely parallels the study of interpersonal trust and behavioral decision-making

in general. However, advice-utilization scholars also emphasize the quality of decisions, or in other words, *should* a decision-maker utilize advice (Harvey & Fischer, 1997). Consistent with the common adage, "two heads are better than one," utilization researchers typically find that using advice improves decision quality in domains as diverse as financial planning, healthcare and military strategy (Bonaccio & Dalal, 2006; MacGeorge & Van Swol, 2018; Yaniv, 2004).

A different area of focus for some advice scholars is on the characteristics of advice messages themselves and how they affect the interpretation of advice. This research focus shares many concepts of interest with utilization focused research such as evaluations of an advisor and mental states of decision-makers that lead to differing evaluations of advice quality. Research in this "message paradigm" (MacGeorge, 2016) is well suited to studying situations in which the decision being made does not have a clearly correct answer because an empirical measurement of advice utilization is not required to understand how decision-makers feel about receiving advice. Instead, when of interest, utilization is often measured as an intention to use advice (Feng & MacGeorge, 2006). An additional strength of this research area is that it often studies decision-makers as they face actual problems in their lives; even when studied experimentally (MacGeorge, Guntzviller, Hanasono, & Feng, 2013). Unlike utilization focused research which uses similar theories as interpersonal trust, message focused research has produced several theories including Advice Response Theory (MacGeorge et al., 2013) and the Integrated Model of Advice (Feng & MacGeorge, 2010). A major contribution of theory in this area is separating message characteristics (i.e., politeness, clarity) from advisor characteristics (i.e., competence, effort) and decision-maker characteristics (i.e., emotional state, self-efficacy). Such concepts are not new to the understanding and study of advice per say, but theories like advice response

theory provide a set of measurable constructs that are useful for research like that conducted in the studies presented here.

A theme that is consistent across all interpersonal advice research is the importance of situational context and the type of decision being made. When evaluating the expertise of an advisor, for example, such evaluations are domain specific. Whereas an airplane pilot may be perceived as a great advisor if one needed to fly a plane, they probably are not as good a source of advice for how to decorate a cake. For highly serious personal problems, people may feel as if advice messages from people with whom they do not have an intimate relationship are inappropriate and less caring than messages from people they are close to. Essentially every other aspect of the advice process beyond advisor characteristics is affected by context as well – the study of advice therefore requires the careful consideration of the situation, decision, and risk level present in an advice scenario.

Automated Advice

Automation is a broad term that encompasses a wide variety of technologies. A calculator, airborne drone, or social robot are all forms of automation. One of the most influential definitions of automation comes from Parasuraman, Sheridan, and Wickens (2000), "… a device or system that accomplishes (partially or fully) a function that was previously, or conceivably could be, carried out (partially or fully) by a human" (p. 287). This definition is important because it excludes instances of when one existing technology is upgraded with another (modernization), and it highlights the necessity of existing human capabilities that are replaced (or, "automated") by non-human technology.

The body of literature of advice from automated sources is a considerably newer than research on interpersonal advice. However, the past decade has seen an immense research

interest in automated systems, how humans interact with technology and why humans do (or do not) use automation. Much of this research is conducted in similar ways to advice research by investigating decision-maker and advisor traits or characteristics of advice messages and the resulting advice-utilization. Although there is a large amount of research, the fields (e.g., computer science, industrial engineering) in which it is studied tend to emphasize theory less than social science fields such as communication. Consequently, there is a lack of comprehensive theoretical models that directly describe the advice process in human-automation relationships, and scholars rarely conduct research with hypotheses derived from extant theory. There have been attempts to apply interpersonal theories to human-automation relationships (e.g., Lee & See, 2004; Pop, Shrewsbury, & Durso, 2015), but this frequently introduces problems when aspects of interpersonal relationships (i.e. emotion, reciprocity) are not easily (or impossible) for automation to provide (Gefen, Karahanna, & Straub, 2003).

The study of human-automation trust has produced a number of theoretical models that are address aspects of automated advice, and various theories have been developed to describe more specific types of automation, such as human-robot trust/advice. Additionally, a large amount of theory in this area deals with the acceptance of technology and why some technologies are voluntarily integrated into people's lives while other technologies fail (Davis, 1989; Y. Lee, Kozar, & Larsen, 2003). Human-automation trust theories often incorporate (or are developed to inform) the design of future automation and hence incorporate prescriptive recommendations. This has the benefit of driving future research, but the models can feel outdated quickly if they do not incorporate the latest technological capabilities. Perhaps the primary agreement between interpersonal advice and automated advice theories is that advice is

context dependent. Like interpersonal advice, depending on the situation and decision to be made, different advisors and the messages thy communicate with be perceived differently.

A major contribution of automated advice theory (Lee & See, 2004; Parasuraman, Visser, Wiese, & Madhavan, 2014) the highlighting of aspects of automated advice that have no direct analogy in interpersonal advice research. First, an important factor that automation researchers face is the "level" of automation (Parasuraman et al., 2000). Levels of automation refer to the differing amount of human involvement when using an automated system. For example, on a forecasting of future sales task, one automated source of advice might simply provide an estimate for a decision-maker to consider. An alternative system at a higher level of automation may actually *make* a decision (in effect becoming a decision-maker) while still allowing a human (and *final* decision-maker) the ability to override or modify the forecast. At the highest levels of automation are entirely autonomous systems which operate and make decisions without human intervention. Another important "level" of automation is the degree to which a human decision-maker feels responsible for a suboptimal outcome: the motivation to solve a problem can vary with perceived responsibility. But, high levels of automation could lead the decision-maker to feel no responsibility for the outcome ("the system did it, not me"). Decision-maker accountability and phenomenon like the diffusion of responsibility is also an aspect of interpersonal advice research (Harvey & Fischer, 1997; Yaniv, 2004), but blaming a machine which has no feelings is a different consideration for decision-makers than blaming a fellow human who could be hurt (Friedman, 1995). Additionally, there exists large individual differences in if people feel it is morally right to even use automated advice for decisions with ethical implications and such debates are at the center of the rapidly growing field of machine ethics (Coeckelbergh, 2010; Moor, 2006).

In sum, existing automated advice research, like interpersonal advice research, highlights the importance of situational context and decision tasks when studying the advice-utilization process. Additionally, automated advice theory points to two important aspects of study design: the level of automation and the perceived responsibility of the decision-maker. In the studies below that compare interpersonal advice to automated advice, it is important to use lower levels of automation that mirror interpersonal advice situations; in other words, automation that puts the human in a responsible decision-making role and the automation in strictly an "advisor" capacity.

Theoretical Perspective for Comparing Interpersonal Versus Automated Advice

There are many academic fields that examine both interpersonal and automated advice. Given the large amount of research in these areas, there is surprisingly little research that directly compares human advice to automated advice. Additionally, the studies that do exist are difficult to create a cohesive picture with due to the issues outlined above such as differing paradigms and theoretical/applied purposes. Nevertheless, the current body of research points to several robust findings that differentiate automated advice from interpersonal advice especially as it relates to the utilization of advice: higher default utilization of automated advice and less utilization of automated advisors following negative outcomes.

In a series of studies in which human decision-makers received advice from either a human or automated source on repeated trials, it was found that decision-makers typically utilize automated advice more than human advice on initial trials (Dijkstra, Liebrand, & Timminga, 1998; Dzindolet, Pierce, Beck, & Dawe, 2002). The effect was especially strong when both advisors were described as experts ("expert systems" for automation). Importantly, in some studies, the initial bias towards automated advice decreases over time as decision-makers are able to calibrate their level of advice-utilization to the accuracy of the tool. Similar results have

been found in the forecasting literature (Önkal, Goodwin, Thomson, Gönül, & Pollock, 2009). Conversely, there are a handful of studies that indicate an opposite effect: a general mistrust of automation initially (Lerch, Prietula, & Kulik, 1997; Wærn & Ramberg, 1996). However, one key aspect of these contradictory studies is that the participant is given no information about the expertise of the human or the system. This suggests that humans may be more "suspicious" of automated advice when they know nothing about the source as compared to humans, and if so, this would be consistent with the general human tendency to trust other humans by default (Levine, 2014; Waal, 2009). In sum, however, most evidence is consistent with humans having a default bias towards utilizing automated advice (Madhavan & Wiegmann, 2007b).

A second consistent effect in automated advice research is the greater decrease in automated advice utilization following the issuance of poor advice. For example, when the reliability of an automated and human advisor is held constant, trust declines more rapidly in automated advisors as reliability is reduced. When automated advisors malfunction, it can lead to a rapid and nearly complete disuse of automated advice (Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003). The most thorough explanation of the bias to trust automation more initially, and the harsher judgment of automation following a mistake is the perfect automation schema (Madhavan & Wiegmann, 2007b). It suggests that humans have different assumptions about the underlying qualities of humans and automation. Most importantly, automation is expected to be ridged, reliable and consistent. Humans are expected, on the other hand, to be adaptable but also susceptible to judgment error. In short, people think automation will be perfect, but not humans.

One communication theory that can explain some of the effects common in both interpersonal and automated advice research, as well as the effects that seem to differentiate them is Expectancy Violations Theory (EVT) (Burgoon, 2015; Burgoon & Jones, 1976). Initially

developed to explain interpretations of nonverbal behavior, EVT has subsequently been applied to other communication processes such as verbal communication, deception and conflict management (Bachman & Guerrero, 2006; Houser, 2006). Recently, the one of the developers of EVT used it as a framework to study impressions of virtual agents (i.e., avatars), showing the applicability of the theory to the world of human-automation communication (Burgoon et al., 2016a). The first key component of EVT is expectations. People have expectations about how a person will act in an interpersonal encounter, for strangers these expectations are largely determined by social norms and stereotypes that the recipient of communication has regarding the sender (Burgoon, 2015). When these expectations are not met, it results in the second key component of EVT: expectancy violations. When expectations are met, it is an expectancy confirmation. Expectancy violations are typically the most cognitively arousing as they are inconsistent with a receivers beliefs about what should occur, and thus, they affect communication processes such as liking, persuasion and impression formation to a greater degree than expectation confirmations (Bond et al., 1992; Burgoon, 2015). A third key component is violation valence: violations can lead to both more positive and more negative evaluations of a communicator.

Components of interpersonal advice and automated advice theories such as evaluations of trustworthiness and message politeness are widespread. These evaluations lead to expectations about the usefulness and appropriateness of advice messages from and advisor. Furthermore, because the perfect automation schema is a well-established and robust effect, decision-makers seem to have fundamentally different expectations regarding automated versus human advisors. In the framework of EVT, the expectation of consistent, high performance from an automated advisor would lead to greater advice utilization in general than a human advisor who is expected

to not perform as well. On the other hand, when automation is not perfect, it represents a more negative expectancy violation than it does for a human advisor. The increased negative valence of this experience is then reflected in decreased utilization of the automated advisor.

The application of expectancy violations theory to understand human versus automated advice allows for the careful consideration of a wide variety of effects that have been found in research using EVT. For example, (Afifi & Burgoon, 2006) found that initial negative violations in and interaction could be discounted because they only created uncertainty about a communicator, but second negative violations confirmed a negative evaluation and affected behavior. This is an element (multiple instances of bad advice) that has been incorporated into study 2 of this research project. The application of EVT also provides some insight as to what may happen when one advisor type (human/automation) replaces the opposite type. When one advisor is replaced by the opposite type, it could lead to a cognitive contrast effect that heightens the salience of these differing expectations and thus exaggerates biases resulting from them. A decision-maker may expect perfection from automation, but this expectation of perfection could be made extra salient when the decision-maker has previously received advice from a human whom they expected to be fallible. Contrast effects occur when the evaluation of one target is affected by the evaluation of a preceding target; contrast effects are one of the most widely studied cognitive phenomena (Palmer & Gore, 2014). The research presented in studies 2 and 3 below are some of the first investigations that examine the successive contrast of replacing a human advisor with an automated advisor.

Task Demonstrability

The differing underlying qualities of human advisors vs. automated advisors are affected by other factors than just the perfect/fallible divide implied by the perfection automation schema.

People judge the suitability of automated advice as more or less superior than human advice based on the context of a decision being made – the most important contextual factor being the decision type (Dijkstra et al., 1998). This is similar to findings in interpersonal advice literature indicating that different advisors are perceived as more or less credible or expert depending on the problem domain (Carlson, 2014; Sniezek & Van Swol, 2001; Van Swol, 2011). In regards to automation, for example, (Parry, Cohen, & Bhattacharya, 2016) note, "AI-based decision systems are notoriously poor at evaluating subjective criteria (e.g., in determining what constitutes art)" (p. 585). In a more general sense, automation is expected to have qualities that lead to good performance on tasks requiring *quantitative processing*, but inferior to humans on tasks that require *qualitative values* (Parry et al., 2016). Quantitative versus qualitative is one aspect of a widely used framework for understanding different decision-types that was proposed by decision-making researchers: task demonstrability (Laughlin & Ellis, 1986). Task demonstrability and explicated by (Laughlin & Ellis, 1986) is anchored at one end by tasks that are intellective (they have a clear correct answer, like 2+2) and at the other end by judgmental tasks that have no clear answer (such as if a music performance is good or if an animal is cute).

Regarding differing types of tasks and automation, researchers in human-robot interaction have been especially focused on the issue of what qualities humans expect robots to have, presumably because such evaluations are affected by design elements. For example, the degree to which a robot resembles a human can affect emotional responses to interacting with the robot and affect it's perceived suitability for certain job roles in society (Duffy, 2003; Katz & Halpern, 2014). Therefore, it is not always a question of human vs. automation – there can instead be a continuum of *humanness* in automated systems, particularly robots. But these social responses to automation do not mean the psychological experience of interaction is the same as with humans

or we expect the same things from automation when it comes to issues such as consciousness and morality (Kahn, Ishiguro, Friedman, & Kanda, 2006). A large amount of research shows that even humanoid robots are not given the same moral standing as a human being, and there are qualitative attributions of human inherent worth that we do not apply to even the most lifelike technology (for review, see (Kahn et al., 2006; Malle & Scheutz, 2015). Furthermore, this effect is strong enough that the lack of moral agency given to automation is even present in young children who we might assume lack the cognitive maturity to differentiate human from machine (Kahn et al., 2012).

Many scholars may argue a contrary point: that humans do in fact assign moral worth to automation. However, most of the literature indicating that automation is judged as having moral qualities is focused on placing *moral blame* in an automated decision as opposed to saying the automation *itself has moral standing* (Malle & Scheutz, 2015; Voiklis, Kim, Cusimano, & Malle, 2016). Furthermore, many of the studies showing automation is blamed for moral wrongs is done with the automation being a type of robot, a much higher level of automation than is studied in a typical automated advice scenario, and robots are typically designed for some level of anthropomorphism (Billings et al., 2012). Anthropomorphism may increase the amount that humans respond to automation socially (such as placing blame on the robot), but this does not equate to the robot having inherent moral worth or the capacity to morally reason (Coeckelbergh, 2010; Kahn et al., 2012).

In sum, the extant literature clearly shows that humans experience interaction and see the fundamental qualities of humans and machines differently – especially as it relates to moral agency. Unfortunately, however, the actual utilization of automation to make less demonstrable decisions with moral consequences is primarily the domain of philosophers and ethicists – there

exists very little quantitative study of this topic. The most scientifically organized study of this area was conducted far back in the 1970's and 1980's when the clinical vs. actuarial decision-making debate was occurring in clinical psychology. Although essentially no data was collected during this time, discussions with clinicians indicated that the lack of moral reasoning was one of the primary reasons for resisting the use of actuarial advice. Actuarial (automated) systems inherently "dehumanize" a patient and are unable to empathize (Dawes, Faust, & Meehl, 1989). On the other hand, proponents of automated advice argued that the superiority of it compared to human judgement meant that automation was indeed the correct ethical choice (Grove, Zald, Lebow, Snitz, & Nelson, 2000; Meehl, 1986). Today, a similar debate about the dehumanizing nature of automated decision making (versus the potential for increased accuracy using automated advice) continues in areas such as law and healthcare (Jha, 2012; Murdoch & Detsky, 2013).

Studying automated advice in the context of moral decisions is difficult because often the decision-making task is not comparable to existing studies that investigate automated advice when the consequences have less qualitative (and more demonstrable) consequences. For example, forecasting researchers frequently look at quantitative outcomes like the likely price of a stock or inventory levels (Alvarado-Valencia & Barrero, 2014). On the other hand, decisions with more moral consequences, such as should a military drone launch or not launch a missile that could result in civilian casualties, are not comparable to past research in a how a decision-maker moves through the decision-process or weighs potential outcomes (Cummings, 2006; Grossman, 2009). Furthermore, even studying such topics in a controlled setting is difficult due to the potential for psychological distress to study participants. As a result, it is difficult to

compare experimental studies looking at automated advice to more philosophical or case study-based research on automated advice with moral consequences.

The difficulties of comparing task types is clear, but the importance of a task manipulation along the demonstrability continuum is also clearly critical to understanding the difference between human and automated advice. In study 3 presented below, I believe I have found an ideal balance between task similarity but differing demonstrability between two decision making tasks. First, the tasks are forecasting tasks, which immediately precludes them from being purely intellective because a future state is unknowable. At the same time, there will eventually be a correct answer once the future state is current, which allows for an evaluation of the accuracy and quality of advice. My manipulation of task demonstrability is largely reliant on the perceived consequences of a poor decision. In one set of tasks, people make forecasts about financial and management related decisions, such as how to split a budget between ordering two office products with the goal of having no unused inventory and wasted expenditure. In the other task, decision-makers look at the same graphs and data but are under the impression they are making forecasts for a humanitarian agency, such as how much food to allocate between two refugee camps to minimize starvation deaths. The consequences of a wrong decision therefore vary from demonstrable (dollars) to indemonstrable (the value of human life).

To confirm perceived differences in tasks, I conducted a brief pilot study using Amazon's Mechanical Turk service using US citizens over the age of 18. Participants were randomly assigned into either the humanitarian (n = 25) or management condition (n = 25). Participant's made four forecasts (with no advice) using the same stimuli to be used in studies 2 & 3 below. After completing the tasks, participants filled out a brief questionnaire.

Independent t-tests, management vs. humanitarian task.

Item	*t*(50)	*p*
1: The decisions I made were more about money than human life	14.551	0.000
2: The decisions I made were more about human life than money	13.310	0.000
3: There is no right answer to these problems	7.515	0.009
4: The best one can do on these sorts of problems is just find the best answer according to the numbers	5.548	0.024
5: I think a machine, like a statistical computer program would find these decisions difficult to make	1.589	0.213
6: I think someone needs to be a compassionate person to make the best decisions on the forecasts I saw	14.105	0.000
7: I think a computer program would value the same things I did when making my decisions	3.291	0.076
8: If the same decisions were made by one of my family members, I think they would be really similar to the decisions that I made	0.458	0.502
9: If the same decisions were made by one of my close friends, I think they would be really similar to the decisions that I made	0.229	0.634
10: If the same decisions were made by a fellow human, I think they would be really similar to the decisions that I made	0.018	0.893
11: If the same decisions were made by a statistical software, I think they would be really similar to the decisions that I made	8.382	0.006
12: If the same decisions were made by artificial intelligence, I think they would be really similar to the decisions that I made	5.220	0.027

Overall, I felt like the pilot results indicated the tasks were perceived differently. Interestingly, the two questions most directly measuring demonstrability "there are no right answers…," and "The best one can do…" were significant, but not the question about computer programs finding the decisions difficult. However, the questions regarding the similarity of another decision maker's decision showed significant effects if the decision-maker was human or automated. Some items from this pilot study were used in the studies below; manipulation checks between tasks are presented in study 3.

Studies 1, 2 & 3 follow. Study 1 conducts preliminary investigations into human versus automated advice. Study 2 adds in manipulations of advisor replacement. Study 3 adds the manipulation of task demonstrability. Following the three studies, a summary of additional findings is presented as potential future directions for research. Finally, results of all three studies are discussed and situated in communication theory.

Chapter II: Study I

Prahl, A., & Swol, L. V. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting, 36*(6), 691-702. doi:10.1002/for.2464

Chapter III: Study II

People and the Machines: Advice Utilization when a Human Advisor is Replaced by Automation

People and the Machines: Advice Utilization when a Human Advisor is Replaced by Automation

Andrew Prahl

University of Wisconsin- Madison

Abstract

This research studies the effects of replacing an automated or human advisor with a different advisor over the course of 20 forecasting trials. Participants first completed 10 trials with one type of advisor (human/automated), then evaluated the advisor. For the following 10 trials, the participant was provided with a new advisor who was either the same type or different type as before (human/automated). Results show that automated advisors are utilized more immediately after they replace a human advisor compared to automated advisors that replace other automation. We also find that automated advisors are utilized less following the issuance of bad advice if the advisor has replaced a human. Additionally, automated advisors that replaced humans were rated as issuing lower quality advice, and human advisors that replaced automated advisors were rated as providing better quality advice. Results are discussed in the context of expectancy violations, contrast effects, interpersonal trust, and human-automation trust theory.

Introduction

A recent report suggested that automation in the workforce would result in about 5% of all human jobs in Canada being replaced by automation over the next two decades. This was the most conservative estimate, with other analyses suggesting that over 40% of the workforce could lose their jobs to automation (Sunil & Thirgood, 2016). Similar reports estimated the effect on the US workforce in similar terms (Benedikt Frey & Osborne, 2013). The threat of automation to jobs has gathered a fair amount of coverage in the popular press, the most dire accounts often discuss a future of humanity that is "post-work," where the vast majority of humans are left unemployed (Thompson, 2015). The focus of much of this speculation is the doom and gloom outlook for the jobless "victims" of automation, but far less attention is paid to how the "survivors" feel. Organizations contain complex communication and social networks between human workers who interact, exchange advice, and offer social support to one another (Lee, Bachrach, & Lewis, 2014; Tasselli, Kilduff, & Menges, 2015). How will a worker react when a fellow human is replaced with a machine?

Many jobs involve the giving and receiving of advice on work related tasks. The social process of advice exchange has been identified as one of the primary ways that organizational culture, communication networks, and status relationships are established in organizations (Agneessens & Wittek, 2012; Brooks, Gino, & Schweitzer, 2015). Automation in the modern workplace is reaching a point where it is frequently used as an advisor, the most visible example being in the financial sector where much has been written about the "robo-advisor" and more advanced systems that use algorithms to trade autonomously (Ji, 2017; Stone, 2012). Medical care is another industry that has seen a rapid advance of automated advising systems, the IBM Watson artificial intelligence platform is now advising oncologists on treatment plans, and

physicians decide on treatments for critically ill patients with the use of algorithms (Cha, 2015; Doyle-Lindrud, 2015).

Clearly, it is becoming increasingly important to understand not only what differentiates the utilization of automated compared to human advice but how the effects of such differences will play out when a human advisor is replaced by an automated advisor. In this study, we put participants through a series of forecasting tasks where they received advice from a human or an automated advisor. Midway through the trials, the advisor was replaced by either another of the same type of advisor (human/automated) or the other type. In this way, we can analyze the effects of a human advisor being replaced by automation compared to if the human is replaced by a fellow human or the automation simply replaces other automation. Below, we review past human versus automated advice research and generate hypotheses using expectancy violation theory and contrast effects. Results are discussed and situated in emerging human-automation trust theory; implications for the accelerating automation of labor are also discussed.

Past Investigations of Automated vs. Human Advice

Both interpersonal and human-automation advice giving are well-studied phenomenon. The body of literature is so large that several influential review articles covering interpersonal advice have been written (Bonaccio & Dalal, 2006; Yaniv, 2004) and a recent handbook has been devoted to interpersonal advice (MacGeorge & Van Swol, 2018). In addition, there are numerous reviews and meta-analyses reporting research in human-automation trust (Hancock et al., 2011; Hoff & Bashir, 2015). However, there are relatively few studies which have studied both human and automated advice in the same experimental design or real-world context. Within these studies, there is conflicting evidence as to what differentiates the two sources of advice. Generally, it is found that advisor accuracy, decision-maker attributes, and environmental context

are the most important determinants of advice-utilization (Hoff & Bashir, 2015), but the effect of each operates in different ways for human versus automated advisors.

The first scholars to consider the value of automated versus human advice were clinical psychologists in the mid-20th century (Dawes, 1979; Meehl, 1986). The discovery that mathematical (i.e., actuarial) methods for diagnosing mental health disorders were considerably more accurate than human judgment sparked a decades long debate as to how much a clinician should trust advice from such models. Although no studies were done to directly measure advice-utilization, per say, it became clear that equivalent or superior accuracy of algorithmic advice was not enough to overcome a general human aversion to using automated advice for mental health diagnosis and treatment (Katsikopoulos, Pachur, Machery, & Wallin, 2008).

This early comparison of human versus automated advice is rarely examined in human-automation trust research that began to accelerate with the proliferation of automation into the workplace near the end of the 20th century. Then again, the decisions studied by human factors and engineering scholars is considerably different than a clinical psychologist judging the psychological state of a patient. For example, a foundational study comparing human advice to automated advice, for example, used a target-detection task – where a simple yes/no answer is needed after viewing an image and identifying a hidden object (Dzindolet et al., 2002). Subsequent studies have used similar tasks, such as screening luggage x-ray images for weapons and identifying the next number in a numerical pattern (Madhavan & Wiegmann, 2007a; Sutherland, Harteveld, & Young, 2016).

In contrast to early clinical research, the sum of research in the human factors field suggests that human decision-makers exhibit a general tendency to trust automated advisors more than human advisors, especially in the initial trials of a repeated measures experiment

(Dijkstra et al., 1998; Dzindolet et al., 2003). This phenomenon along with a tendency to become complacent and reliant on automated advice is known as automation bias (Bahner, Hüper, & Manzey, 2008; Goddard, Roudsari, & Wyatt, 2012). This effect is moderated by the perceived expertise of the advisor, however, as some research suggests "expert" human advisors are trusted more than "expert" automated advisors (Wærn & Ramberg, 1996). Although we know of no research that compares "novice" automated advisors to novice human advisors, such a situation is likely to elicit automation bias because in the absence of any expertise information, human advisors may be perceived to be more novice in general than automated advisors (Madhavan & Wiegmann, 2007b). In contrast to automation bias, human bias, a general bias for human advice when compared to automation, has been reported in a number of past studies as well. Forecasting researchers have found that human advice is preferred over that of a statistical method advisor, both when the human and automated advice are presented separately or simultaneously (Dietvorst, Simmons, & Massey, 2015; Önkal et al., 2009). In addition, when presented with an automation generated forecast, decision-makers make more adjustments to the forecast than when the forecast is provided by a human (Önkal et al., 2009). This research echoes the earlier research from clinical psychology about actuarial versus clinical judgment, where clinician decision-makers were resistant to using statistical methods to diagnose or assess risk for their patients. Perhaps one explanation of the differing bias towards automated or human advice is that the task moderates the bias. In diagnostic tasks such as target detection (in which there is or is not a signal present at the time), automation is preferred; but in tasks that are predictive (i.e., forecasting), human judgment is preferred.

A more consistent effect in both human-automation trust research and other fields like forecasting is the decreased utilization of automated advice following an error (in the case of a

diagnostic task) or advice that causes the decision-maker to perform worse than their unaided judgment in forecasting. Decision-makers tend to judge automation more harshly when it makes a mistake and decrease the utilization of an automated advisor far more than a human advisor who makes the same mistake (Dzindolet et al., 2002; Madhavan & Wiegmann, 2007a; Parasuraman et al., 2014). Some recent forecasting research suggests that this effect remains even when humans make more (or worse) errors than automation and decision-makers are aware of this (Dietvorst et al., 2015). The human bias to disuse automation following even minor errors is observable outside of the lab as well, ignoring automation following malfunctions is a contributing cause in many real-world accidents such as power plant shutdowns and airplane crashes (Mehta et al., 2014; Parasuraman & Riley, 1997; Wise, Rio, & Fedouach, 2011).

Decision-maker feelings also play a role in trusting the advice of humans versus automation. Professional forecasters may be less inclined to use the advice of forecasting support systems because they feel that using the system admits fault in their own human judgment (Goodwin, Fildes, Lawrence, & Nikolopoulos, 2007; Goodwin, Gönül, & Önkal, 2013). This is not unlike case studies from other industries that suggest workers may resist using new technology, or purposely misuse it, because they feel it is designed to replace their jobs (Naweed, Bearman, Dorrian, Rose, & Dawson, 2013). The comfort level of a decision-maker with technology in general also may cause differences in human and automated advice-utilization. (Wærn & Ramberg, 1996) found that Indian car mechanics rated advice from an automated system as less trustworthy than that from a human, but the opposite was true for American mechanics, who may be more familiar with technology. Again however, results are mixed in this area because in contrast to Waern et al., (Önkal et al., 2009) used graduate business students, who presumably are very familiar with forecasting technology, and still found a preference for

human advice. Perhaps the participants in Onkal et al. were “too” familiar with forecasting technology and knew of some of its flaws.

A third factor emerging from human versus automation advice studies is the effect of task environment. Humans may be introduced to automated advice in workplaces where they already have a well-established routine (Hoff & Bashir, 2015). If the automated advice is not delivered at a time or place that is consistent with a preferred workflow, it may be disused. This is in contrast to a new human advisor who, because trust is a social and reciprocal process, may be attended to because the decision-maker feels obligated to reciprocate the effort put forth. Decision-maker workload also affects trust in automated systems. (Lyons & Stokes, 2012) found that utilization of automated advice increased when decision-maker workload was high, but only if the advice was delivered in a simple format. More complex presentations of advice from automation were, on the other hand, utilized less as workload increased. This would likely differ for human advisors because human advisors constantly adjust the advice messages to the receiver and situation, and presumably such adjustment of advice messages would extend to the perceived emotional state and workload of the decision-maker (MacGeorge, 2016). Additionally, because automated advisors are expected to be more predictable and consistent than human advisors (Dzindolet et al., 2003), decision-makers may ignore automated advice more because they feel like they already know what the advisor would say.

There is clearly a fair amount of evidence suggesting that automated advice is more or less likely to be utilized than human depending on a number of factors, but no review of literature would be complete without mentioning a growing number of studies that find no difference in the use of automated versus human advice. King et al., (2014) found that automated health advice compared to human peer advice resulted in similar levels of exercise routine

adoption over the course of 12 months. [Removed] did not find a general preference for human advisors over automated advisors in a forecasting task when both advisors were performing well. Regarding bad advice and trust repair, de Visser et al., (2016) showed that manipulating the humanness of the automation (i.e., virtual agent with human avatar versus no avatar) has the potential to make trust repair equivalent between human and automated advisors. In sum, past direct comparisons of human versus automated advice are far less plentiful than studies of either advisor type in isolation. Within these studies, there is ample evidence that decision-makers evaluate and utilize each advisor differently depending on a number of conditions. The most comprehensive explanation for these differences proposed thus far is the perfect automation schema (Madhavan & Wiegmann, 2007b). Below, we use expectancy violation theory to explain the perfect automation schema and contrast effects literature to formulate our hypotheses.

Expectancy Violations and Contrast Effects

The perfect automation schema (Madhavan & Wiegmann, 2007b) assumes that people have differing expectations of humans and automation: whereas a decision-maker expects automation to be nearly perfect on some tasks (Dzindolet et al., 2002), the same decision-maker understands that humans make mistakes and can adjust future advice to account for those mistakes. Thus, the expectation of perfection leads to inflated perceptions of the quality of advice for automation during the initial trials of an experiment. Expectancy violations theory (Burgoon & Hale, 1988), though never explicitly intended for use to describe human-automation communication, offers a similar explanation for an initial bias for either advisor type. Expectancy violations theory describes how when humans interact, we have certain expectations about how the other party will behave – and violations of those expectations have greater weight (positive or negative) on the appraisal of a target than behaviors that conform to expectations. A key

proposition of expectancy violations theory is that the interaction context determines what behaviors are expected. In decision tasks that are perceived to be more suited to mathematical and unemotional judgement, decision-makers would expect an automated advisor to be superior, even perfect. However, for decisions that a decision-makers feels are more intuitive, a fellow human would be trusted more because intuition is a quality that humans are expected to have.

Expectancy violations and the perfect automation schema also explain reactions to advisor mistakes. Because of the expectation of automation as being consistent and reliable especially for mathematical tasks, mistakes are more unexpected (Madhavan & Wiegmann, 2007b). Furthermore, expectancy violations vary in their valence (both positive and negative but also how much so). Because humans are expected to be fallible, even experts, an error constitutes a less severe violation and is perceived as less negative. Additionally, a decision-maker may adjust their future expectations of the automation more following a mistake, as the decision-maker sees the mistake as a "red flag" that the automation will make other similar mistakes in the future because the automation is presumably pre-programmed and cannot adjust (Dzindolet et al., 2003). On the other hand, decision-makers may be disappointed that a human advisor led them astray, but they are not as quick to think the advisor is worthless. Compounding this tendency to judge automation more harshly is the general tendency to trust other humans (Levine, 2014; Waal, 2009); humans are more resistant to say they distrust another human compared to automation because distrusting a human is a may be more unpleasant thought (Jian, Bisantz, & Drury, 2000). In sum, the fundamental expectations that decision-makers have about automated versus human advisors depend on the context of the interaction. The bias towards automation at the start of an experiment appears to be sensitive to changes in the decision task. On the other hand, the expectation that automation is consistent appears to be less sensitive to

task manipulations, thus any mistake by an automated advisor leads to decreased utilization of automated advice following the issuance of bad advice.

In the study proposed here, we use similar forecasting tasks as used in previous research (Dietvorst et al., 2015; Önkal et al., 2009) because forecasting tasks represent a balance between mathematical calculation and intuition about future trends. Our design is unique in that we will replace one advisor with another in one of the middle trials of the experiment; replacing one advisor type with a different type offers new opportunities to understanding the effects of the differing expectations. Our design is also notable in that it is similar to real world situations where human advisors are being replaced by automation.

In the study below, over the course of ten forecasting trials, decision-makers will become familiar with their advisor and (at the end of the first set of forecasting trials) fill out a survey evaluating their advisor. Generally, familiarity with an advisor, and trust, build over time (Lewicki, Tomlinson, & Gillespie, 2006), and familiarity leads to more perceived similarity and liking (Strauss, Barrick, & Connerley, 2010). Therefore, the replacement advisor is at a disadvantage in comparison. So, any initial bias towards trusting automation will compete with the unfamiliarity of the automation being new when it is the replacement advisor. Additionally, the bias towards automated advice has only been demonstrated in the initial trials of experiments – it is not clear that a decision-maker who has gained experience from previous trials with any sort of advisor will exhibit the bias to trust automation more. A human advisor that replaces automation or who replaces even another human also faces the same hurdle in that the decision-maker has no previous experience with the advisor. Because our forecasting task is less likely to elicit an automation bias to start with in comparison to previous research using target/signal detection tasks, and because the effect has only been observed in the very first trials of an

experiment, the decrease in advice utilization for the "new" replacement advisor should occur regardless of whether the new advisor human or automated.

H1: When an advisor (either human or automation) is replaced by a different type of advisor, utilization of advice will decrease compared to conditions in which the advisor replaces the same type of advisor.

If we find support for a decrease in utilization following replacement, there still may be differences between human and automated advisors. If the bias towards automated advice holds even when the automation is a new advisor, then automated advisors will be utilized more than human advisors when both advisors are replacing their opposite advisor type.

H2: Automated advisors that replace human advisors will be utilized more than human advisors who replace automated advisors on the first trials following replacement.

Another aspect of the bias towards trusting automation more than human advisors is that the bias disappears after initial trials because repeated interactions (with high quality advice) have a tendency to increase advice-utilization - or, in automation research, *reliance* on the advisor (Bonaccio & Dalal, 2006; Hancock et al., 2011; Hoff & Bashir, 2015). We expect that human advisors that replace automation will incur the greatest decrease in utilization due to increased uncertainty with a new advisor, but that decrease may be short-lived this uncertainty is decreased if advice is high quality (as it will be for the first 6 trials). Additionally, we have little understanding of how long the bias towards automation lasts. Previous research using much faster tasks than forecasting (i.e., image recognition), have been unable to expand on just how long the bias towards automation lasts. For example, in (Madhavan & Wiegmann, 2007a), participants completed 200 trials in 5 distinct blocks, only the mean of the first 25 trials provided evidence of automation bias. Our experiment uses 10 trials with feedback after each trial,

allowing the participant to quickly adjust to advisor performance and therefore we have an ideal environment to explore if, and how quickly, automation bias fades. The resulting question has important real-world applications: does automation bias decrease over time?

RQ1: If automation bias is found for automated advisors that replace human advisors, does the bias last for longer than the initial trials after replacement?

Contrast Effects

The replacement of an advisor also raises important questions regarding the greater decrease in advice utilization following mistakes for automated advisors. In the experiments proposed, both the human advisor and the automated advisor will deliver bad advice on one trial and marginally bad advice on another. Additionally, participants will fill out a survey assessing their perceptions of the first advisor before they are presented a replacement advisor. With the qualities of their first advisor salient, a replacement advisor may elicit a contrast effect. There is a large body of literature on contrast effects in social comparison situations (Tourangeau & Rasinski, 1988; Amos Tversky & Kahneman, 1974). When a judge is presented several targets for evaluation, the evaluation of one target is affected by the evaluation of a previous target or other targets presented simultaneously. Contrast effects are related to the anchoring effects that also have a wide body of supporting literature (Epley & Gilovich, 2006).

Palmer and Gore, (2014) specified several conditions under which contrast effects are most likely to occur, and three of them are present in our experiment. First, contrast effects are most likely when decision-makers have ample cognitive resources because they have resources to divert to actual comparison processes instead of quick judgment. Our experiment is not particularly complex, and we set no time limit for each forecast. Second, contrast effects are most likely when the context is homogenous between evaluation conditions. Our only

manipulation during the replacement phase is the advisor themselves - differences in advice utilization and advisor evaluation are unlikely to be explained by contextual factors. Finally, negative events, consistent with the negativity effect (Rozin & Royzman, 2001), are more likely to elicit contrast differences. This is important because reactions to bad advice are clearly the participant reacting to a negative event that, as established earlier, is perceived more negatively for automation. In total, our experiment is likely to make differing advisor expectations more salient due to a cognitive contrast effect. When participants gain experience with one advisor type and evaluate the advisor, the replacement of that advisor will lead to a process of social comparison which makes the differing qualities of both advisors more salient. With this increased salience we expect the effects of such expectations (automation bias and decreased utilization following mistakes) to be increased as well.

H3a: After giving bad advice, automated advisors that replace human advisors will be utilized less than automated advisors that do not replace humans.

Similarly, when a human advisor replaces an automated advisor the human's underlying trait of adjustability is likely more salient because they replaced an advisor with fundamentally different qualities. When compared to a previous automated advisor, mistakes by the human will result in less of an expectancy violation and judged less harshly.

H3b: After giving bad advice, human advisors that replace automated advisors will be utilized more than human advisors that do not replace automation.

Finally, our primary goal is to test differences in utilization between advisors; however perceptions of advice quality are also of interest because these self-reported items are frequently paired with behavior utilization data in both human-automation trust literature (Hoff & Bashir, 2015) and interpersonal advice literature (Bonaccio & Dalal, 2006). Furthermore, decision-

maker emotions are quickly being identified as a key factor leading to advice-utilization especially in human-automation trust literature (Merritt, 2011; Schaefer, 2016). Two questions are of interest: first, does perceived advice quality or emotional reactions to receiving advice predict advice utilization for either advisor type? Second, are contrast effects also detectable in ratings of the advice – is automated advice perceived differently if it replaces human advice compared to if it replaces advice from a different automated advisor?

RQ2: Do emotions or perceptions of advice quality predict advice utilization?

RQ3: Are advisors rated differently and/or do decision-makers experience different emotions if the advisor replaces a different type of advisor compared to if the new advisor replaces a similar type of advisor?

Method

Participants and design

A total of 356 participants completed the study. Participants were randomly assigned to the two first advisor conditions (human/automated) and at the midpoint of the survey, assigned to a replacement advisor condition (human/automated). In total, n=82 automated advisor replaced by automation (ArA), n=81 human advisor replaced by human (HrH), n=90 automated advisor replaced by human advisor (ArH), n=85 human advisor replaced by automated advisor (HrA). After the data was cleaned for incomplete or fraudulent responses or failure to answer several attention check questions, the responses for analysis were n=77 (A), n=76 (H), n=82 (ArH), n=73 (HrA), (N=308). All 20 graphs and 20 forecasting scenarios were randomly assigned into four sequences using a random number generator, the sequences were then randomly assigned within each advisor condition. Participants were recruited through Amazon's Mechanical Turk service and were required to be United States citizens over the age of 18. Seven percent of

participants were between the ages of 18-25, 75% between the ages of 25-55, the remaining 18% were over the age of 55. 59% of participants were 60% male and 40% female, two participants did not report their sex.

Task

Participants completed 20 forecasting tasks consisting of a graph of past data for a variety of issues related to operating room management in a hospital setting. Each graph presented two lines representing the percentages of two items used or event occurrences, the two lines symmetrically trended in opposition to one another. An example of a forecast to make would be "There are two sizes of latex gloves, what percentage of gloves ordered next month should be for hand sizes XS-M versus L-XXL?" The graph would display two lines representing the percentage of all gloves that were consumed in either size, monthly for the past several years. See the Appendix II for screenshots of the task. Participants were then asked to make an initial allocation forecast. A judgmental forecasting task, as opposed to intellective task, was used so results could easily be related to the large amount of previous research on forecasting support systems (e.g., Onkal et al., 2009, Lawrence et al., 2008).

Most forecasting tasks were specific to the operating room unit of the hospital. This is to control for participants having personal intuition for what the outcomes would be. Previous forecasting studies have often used students enrolled in economics classes to forecast stock prices. This may lead to overconfidence and egotistic discounting of advice as the participants consider themselves experts, or the participants have specific knowledge about the shortcomings of forecasting support systems (Goodwin, Fildes, Lawrence, & Nikolopoulos, 2007; Önkal et al., 2009). Recipients may underutilize advice when they perceive high self-efficacy in the domain

(Lawrence, Goodwin, O'Connor, & Önkal, 2006), and our choice of forecasting tasks minimizes this risk.

Procedure

Participants signed up for the study via the MTurk portal. After digitally signing a consent form, participants read a short introduction about the types of tasks they would be given. The manipulation was simple and similar to past studies (Dzindolet et al., 2002; Madhavan & Wiegmann, 2007a; Önkal et al., 2009). Participants were either told at the opening that the advice would come from an algorithmic software program, or an experienced surgeon at the hospital, Logan. Each advisor was introduced to the participant with a short paragraph introducing them as expert advisors, we did this to replicate real world scenarios in which advisors are unlikely to be sought unless they are experienced. After the first 10 trials, the participants filled out a survey assessing their advisor and then were told that they would be getting a new advisor for the remaining 10 forecasts. Advisor descriptions were as followed:

> Your advisor today is an algorithmic computer program called OptiLytics. OptiLytics is a software program used by the Gain Healthcare System to help with forecasting. The statistical models in OptiLytics have been built using 10 years of past Gain Healthcare data, as well as some data from the Center for Disease Control in the United States.

> Your advisor today is: Logan Girard. Logan is a medical doctor who has been working for Gain for 10 years doing operating room management and hospital operations. Prior to joining Gain, Logan gained experience in healthcare management with the Center for Disease Control in the United States.

When advisors were replaced midway through the survey, the introduction text was preceded by “Due to time constraints, we were not able to get OptiLytics/Logan’s advice for every forecast. Therefore, you will have a new advisor to help you on the remaining tasks. Your new advisor is…” In the conditions where a human advisor was replaced by another human advisor or an automated advisor with another automated advisor, the names of the new advisors were OperationAid/Cameron, and the corresponding advisor introductions were nearly identical to those above.

Participants were compensated at the federal minimum wage ($7.15/hr) rate assuming the survey would take 45 minutes. In order to incentivize accuracy, participants were informed that they could win a $100 Amazon giftcard based on their performance if they finished all of the forecasting tasks with the lowest mean percentage error among all participants. No performance record or implementation history was provided or implied for either advisor because past performance and the perceived reasons for system implementation may cause participants to start the experiment with biased expectations (Dietvorst et al., 2015; Madhavan & Wiegmann, 2007a).

After viewing the graph of past data, participants entered an initial forecast (text entry). Then the screen advanced to an “advice” screen in which the advice of either the human or algorithmic advisor was presented. The advice was presented in a “point forecast” format (i.e., an exact number). Participants then had the opportunity to consider the advice and enter a revised forecast on the screen.

After each task was complete, participants received feedback on the accuracy of their forecast. A screen appeared that showed the absolute percentage error of the participant's revised forecast as well as the advice forecast (this allowed them to directly compare their performance

to the advisor). Furthermore, the participant's mean absolute percentage error across all previous revised forecasts was computed and displayed on the feedback screen in a very large red font. Participants were reminded after several trials on the feedback screen that the participant with the lowest mean absolute percentage error would win a $100 gift card. Thus, it is expected that the participants would feel some degree of frustration/gratitude if the advice caused them to decrease/improve their forecast accuracy. Finally, the participants percentage error was multiplied by 1000 and presented as: “This forecasting error is estimated to have cost the hospital $3600.” This was to reinforce that the decisions had financial consequences instead of more emotional consequences, such as patient safety.

Participants were walked through the process of making a forecast and entering their confidence during one “warm up” forecast. Directions appeared on the screen reminding participants of the proper format for entering forecasts and also directing their attention to where and when the advice would appear. After the warm up, participants proceeded to the first of 10 forecasting trials. The survey was programmed such that the advice was always the exact percentage better (or worse) than the participants forecast on each trial, regardless of which graph or forecasting scenario was randomly presented. This avoided a problem common in repeated measures studies in that some participants simply, through luck or skill, perform particularly good or bad on early measures. In our study, every participant experienced the exact same accuracy of advice on each trial.

The first 10 trials were performed to set the stage for the second group of 10 trials where our research interest in advisor replacement lies. For the first five trials after advisor replacement (1-5) and trials 7, 9 and 10, participants were given advice that was always better than their own forecast. On the 6th and 8th trials we implemented a bad advice intervention. We gave advice that

was worse than the participants' forecast. Thus, any utilization of the advice would make their forecast worse (and this was clear on the feedback screen that the advice hurt accuracy). After this bad advice intervention, advice returned to being beneficial for the final two trials (9th and 10th). The first block of 10 trials was performed to set the stage for the second block of trials where our research interest in advisor replacement lies.

Measures

To measure advice utilization, a "SHIFT" variable used in previous forecasting studies was used to directly compare human versus automated advice utilization. We chose to use this measure first to maximize commonality with previous forecasting research that used it as a measure of trust and (Önkal et al., 2009) also because the shift variable is a variant of the "weight of advice" (WOA) variable commonly used in advice research (Bonaccio & Dalal, 2006; Harvey & Fischer, 1997). SHIFT is computed via the following equation:

$$\frac{\text{Judge revised forecast} - \text{Judge initial forecast}}{\text{Advisor forecast} - \text{Judge initial forecast}}$$

After shift values were calculated, z-scores were calculated. Any shift score that was more than 4 standard deviations from the mean was checked for obvious type-o's (i.e., 661 instead of 66.1). After type-o's were fixed, the shift scores were then truncated to the theoretical range of 1 (full advice utilization) and 0 (no advice utilization) (Bonaccio & Dalal, 2006). This was needed because occasionally a participant would overshoot their adjustment in the direction of the advice resulting in a shift score over 1 (or alternatively, move away from the advice to a negative shift score). Consistent with past research using shift scores, such occurrences were less than 5% of the data, but such scores can bias mean estimates of advice utilization, especially

outlier scores. Furthermore, this measure is our operationalization of advice utilization, it is a continuous measure with full advice utilization on one end and no utilization on the other; we do not believe that moving away from an advisor's estimate (i.e., initial forecast is 50, advice is 52; forecast is revised to 49) constitutes something beyond "no" advice utilization.

A questionnaire was administered after the first set of ten trials and after the final ten trials with the replacement advisor. The questionnaire measures were identical and measured perceptions of advice quality and advisor quality on a sematic differential scale. Additionally, Likert survey questions measured emotions when receiving advice, trust of advisor, similarity (value, social norm, and thought process) to advisor, and perceptions of advisor effort. The survey was constructed from items from previous advice literature (MacGeorge et al., 2013).

Positive emotions were measured with four Likert questions on a 1 (*not at all*) to 5 (*extremely*) scale for four positive emotions: Appreciative, Happy, Grateful, Thankful. The negative emotion scale was composed of Mad, Frustrated, Annoyed, Irritated. The four positive and negative emotion questions produced sufficient reliability (positive: $\alpha = 0.940$, negative: $\alpha = 0.943$), and the mean was used as an index of positive/negative emotion. 8 sematic differential questions were used to measure advice quality (e.g., thoughtful, careless); the 8 items achieved sufficient reliability ($\alpha = 0.811$) and is hereon presented as an index. Finally, in order to keep the survey a reasonable length, a pair of questions was asked to assess feelings of reciprocity to the advisor (i.e., "I feel like I owe something to my advisor for their help").

Results

Hypothesis 1 stated that when a replacement advisor was different (human to automated advisor or automation to human) the replacement advisor would be utilized less than if the

replacement advisor was the same (human to human or automation to automation). An independent samples *t*-test was carried out with advisor type replacement (different/same) as the grouping variable. Levine's test for the equality of variances indicated the condition variances were equal ($F = 0.976$ $p = 0.324$). There was no significant difference in utilization between the different advisor replacement ($M = 0.515$, $SD = 0.202$) and same advisor replacement ($M = 0.495$, $SD = 0.221$) conditions; $t(312)=0.857$, $p = 0.392$. Thus, hypothesis 1 is not supported when combining both advisor types. However, when analyzed by advisor type, human advisors that replace automation ($M = 0.490$, $SD = 0.186$) were not utilized less than if they replaced other humans ($M = 0.518$, $SD = 0.216$); $t(156) = -0.871$, $p = 0.385$. Automated advisors who replaced human advisors ($M = 0.553$, $SD = 0.218$) were not utilized significantly less than if they replaced another automated advisor ($M = 0.487$, $SD = 0.231$); $t(148) = 1.404$, $p=0.074$, $d = 0.296$. Thus, hypothesis 1 is not supported. Additionally, the means are in the opposite direction as expected, the largest difference we see in utilization is that automated advisors that replace humans are utilized more. These results are found when looking at the mean of all of the 10 trials with the replacement advisor, but our results to hypotheses 2 below show that utilization is more nuanced.

Hypothesis 2 stated that for advisors that replaced a different type of advisor, automated advisors would be utilized more on the initial trials after replacement than human advisors. We created an index of the average utilization in the first three trials following advisor replacement and ran independent samples *t*-tests comparing advisors (within the different replacement condition only). On these first three trials following replacement, automated advisors ($M = 0.536$, $SD = 0.288$) were utilized significantly more than human advisors ($M = 0.443$, $SD = 0.250$); $t(153) = 2.126$, $p = 0.035$, $d = 0.417$. However, further tests carried out on the individual trials suggested that the first ($M_{diff} = 0.099$, $t(153) = 1.603$, $p = 0.111$, $d = 0.444$) and especially

second trial (M_{diff} = 0.189, t(153) = 2.820, p = 0.008, d = 0.848) were driving this result, as the third trial following replacement was nonsignificant between the two types of replacement advisors (M_{diff} = -0.010, t(153) = -0.171, p = 0.865). To investigate further, we performed a 10 x 2 repeated-measures ANOVA on the 10 trials with the replacement advisor (human or automation) as the between subjects independent variable. There was a significant interaction between trial number and advisor type, F(9,157) = 2.403, p = 0.014, d = 0.251. We then specified polynomial contrasts and results suggested a quadratic effect interaction, F(1,153) = 3.308, p = 0.071, d = 0.294. Although this result is marginal, a quadratic curve provides the easiest way to visualize differences in utilization initially after replacement: automated advisors are utilized more in the first trial and utilization increases even more compared to human advisors in the second. However, by the 3rd trial, utilization between both advisors is not significantly different. On the whole, hypothesis 2 is supported, automated advisors that replace human advisors are utilized more in the first few trials following replacement as opposed to human advisors that replace automation. The quadratic effect also provides an answer to RQ1, advice levels tend to equalize after the first two trials following replacement.

Hypothesis 3a posited that automated advisors who replace humans will incur a greater decrease in advice utilization immediately after delivering bad advice compared to automated advisors that replace other automation. Hypothesis 3b posited the opposite effect (more advice utilization) for human advisors that replace automation as opposed to human advisors that replace other humans. We conducted a 10 x 2 x 2 repeated measures ANOVA on the ten shift values for the post-advisor replacement trials with the advisor type (human/automated) and replacement type (similar/different) as independent variables. Results (using the Greenhouse-Geisser adjusted values) indicated a three-way interaction between trial, advisor type, and

replacement type, $F(9, 2542) = 2.826$, $p = 0.003$. We then specified "difference" contrasts (compares one trial to the previous trial) to look at individual trials with the same analysis and the three way interaction was present for utilization on trial 7 (representing changes from 6 o 7), $F(1, 304) = 4.404$, $p = 0.037$, $d = 0.245$; and trial 8, $F(1, 304) = 3.770$, $p = 0.053$, $d = 0.225$. We also specified difference contrasts (compares one trial to the mean to the previous trials) and the significant interaction remained on trial 7, $F(1, 304) = 10.676$, $p = 0.001$, $d = 0.382$. Examining the graph of estimated marginal means for different advisor replacement (fig. 1.1) and similar advisor replacement (fig. 1.2), it is clear that between trial 6 (bad advice is given) and trial 7 (participants now are aware of the bad advice), utilization drops more for automated advisors that replace humans than automated advisors who replace other automation. This effect is slightly reversed for humans that replace automated advisors who see less of a decrease in advice utilization compared to those human advisors that replaced other humans.

To investigate further and directly test hypothesis 3a and hypothesis 3b, we created a variable composed of the average amount of advice utilization over the first six trials (good advice period) and last four trials (unreliable advice period). Then we subtracted the mean of the first six from the last four; a negative score on this index of the good to unreliable period indicates that the participant utilized advice less, on average, during the unreliable advice period than during the good advice period. As would be expected, values were negative across both advisor and replacement conditions. An independent samples *t*-test in the automated advisor condition indicated that utilization dropped more for automation that replaced a human ($M = -0.139$, $SD = 0.295$) compared to replacing other automation ($M = -0.075$, $SD = 0.324$), $t(148) = -1.780$, $p = 0.077$, $d = 0.298$. Given the marginal significance of this result and the significant difference contrast from the repeated measures test, we created a similar variable to compare

utilization on the single trial (7th) immediately following bad advice with the mean of the previous 6 trials. Results showed a significantly greater drop in utilization for automated advisors that replaces humans than those that replaced other automation (different advisor replacement M = -0.274, SD = 0.358; similar advisor replacement M = -0.114, SD = 0.362), $t(148) = -2.718$, p = 0.007, d = 0.455. Within the human advisor condition, the difference between the first 6 trials average and last 4 was not significant between replacement conditions $t(156) = 0.306$, $p = 0.760$. However, the difference between trial 7 utilization and the mean of the previous 6 trials was significant between different advisor replacement (M = -0.076, SD = 0.431) and similar advisor replacement M = -0.203, SD = 0.406), $t(156) = 1.982$, $p = 0.049$, d = 0.331. This means that humans that replace automation incur less decrease in advice utilization compared to humans that replace other humans on the immediate trial after giving bad advice. On the whole these results indicate that automated advisors that replace human advisors incur a larger decrease in advice utilization following the issuance of bad advice compared to if the same automated advisor had replaced other automation. On the other hand, humans that replace automation incur less of a decrease in advice utilization if they replaced an automated advisor instead of another human. We did notice the significant interaction between trials 7 and 8 (when specifying repeated contrasts). Because this was not a planned comparison, we first conducted an omnibus test for an interaction between the two trials (just 7 to 8, not including the other 10 trials) by advisor and replacement type. The interaction was nonsignificant $F(1,304) = 1.308$, $p = 0.472$. Because our study design delivered two instances of bad advice almost in succession, with few trials after, we were limited in our ability to investigate the question of utilization recovery further.

Our survey results provided insights into RQ3 and RQ4. In general, our results revealed some strong differences in the evaluation of human versus automated advice quality, but none of these measures were significant predictors of advice utilization in any replacement/advisor condition or phase of the experiment, leaving us with no reason to investigate RQ3 further. RQ4 asked if the ratings of advice and emotions experienced is affected by the replacement advisor being of the same type (automated/human) or different type than the first advisor. Our results suggested the replacement type matters: in the human replaces human advisor condition had, as would be expected, there were no significant differences between the ratings of advice quality. However, in the automation replaces automation condition, paired *t* tests revealed that decision-makers reported more negative emotions ($M_{diff} = 0.141$, $t(75) = 2.204$, $p = 0.031$, $d = 0.511$) when receiving advice from their new automated advisor and decreased perception of advice quality ($M_{diff} = 0.194$, $t(73) = 2.392$, $p = 0.019$, $d = 0.564$) from the new automated advisor. In the different advisor replacement conditions, there were no differences in emotions reported, however the automation replacing a human advisor resulted in the automated advice being perceived as lower quality (i.e., helpful, useful) than the human advice ($M_{diff} = 0.213$, $t(71) = 2.559$, $p = 0.013$, $d = 0.589$). Feelings of reciprocity did significantly differ between advisor types when one type replaced the other: decision-makers felt more reciprocity to human advisors than automated advisors when human advisors replaced automation ($M_{diff} = -0.880$, $t(70) = -4.270$, $p < 0.001$, $d = 1.011$) and less reciprocity to an automated advisor that replaced a human advisor ($M_{diff} = 0.900$, $t(79) = 5.089$ $p < 0.001$, $d = 1.204$).

Discussion

Our first hypothesis stated that replacing a human/automated advisor with the other type of advisor would result in a decrease in advice utilization, this was not supported. Rather, for

automated advisors we found evidence for the opposite; they were utilized more if they replaced a human as opposed to replacing other automation. Our second hypothesis stated a bias towards automation would be observed when automated advisors replaced humans, but only in comparison to human advisors that had replaced automation. This hypothesis was supported. Further, to answer our first research question, the bias towards automation is the strongest in the first three trials following advisor replacement and after that utilization equalizes between the two conditions. We also find support for our third hypothesis as automated advisors that replaced humans suffered a larger decrease in advice utilization following bad advice compared to automated advisors that did not replace humans. The opposite effect was found for human advisors, who suffer less decrease in advice utilization following bad advice if they replaced automation compared to human advisors that replaced other humans. Finally, the results answer our research questions regarding both the evaluation of advice quality and decision-maker emotions: although neither is a good predictor of advice utilization, they appear to be affected by the advisor replacement being of a similar or different advisor type. Overall, our results replicate previous research showing a bias towards automation in the initial trials of a repeated measures experiment and a harsher judgment of automation following a mistake (Madhavan & Wiegmann, 2007b; Merritt, Unnerstall, Lee, & Huber, 2015). Our study is the first we know of to test and show that the effects can become stronger when the automated or human advisor replaces an advisor of a different type. These results further our understanding of contrast effects in decision-maker reactions to automation replacing a human advisor, make important theoretical contributions for the continued use of expectancy violation theory in advice research, and have important real-world implications as automation continues to replace humans in personal and professional life. We elaborate on these themes below.

Our results did not suggest that there was a consistently preferred advisor between automation and humans on our forecasting tasks. At various stages of our experiment, we found differences in advice utilization between advisor types, but these effects were constrained to several adjacent trials. For example, our findings that automated advisors were utilized more after replacing a human advisor was driven by the first few trials following replacement and not a consistent effect throughout all ten trials. Additionally, the rapid decrease in automated advice use only occurred after bad advice was issued. This result is not dissimilar to previous advice comparing humans to automation that frequently only finds significant differences in utilization due to an intervention such as changing the reliability of advice (de Visser et al., 2016). Our results are consistent as well with most human-automation trust theories that include situational factors as moderators of trust (Hoff & Bashir, 2015; Lee & See, 2004). Our results suggest that the task of management forecasting is not one where human or automated advice is preferred. Instead, because the effects we found are a result of the manipulation of advisor replacement and advice quality, the management forecasting environment provides an ideal context to explore future advisor type and advice quality manipulations.

The suggestions of an opposite effect than stated in hypothesis 1 is unsurprising given our strong results supporting hypothesis 2. The manipulation of advisor replacement type may have produced a contrast effect in the initial trials following replacement as automated advisors who replaced humans were utilized more than the same automated advisor that had replaced other automation. There are several reasons related to the perfect automation schema that may explain why this effect was constrained to the few initial trials after replacement, all have important theoretical implications for the continued use expectancy violations theory in advice research. First, if a contrast effect did occur, a theoretical implication is that expectations should not be

thought of as being simply present or not present, the salience of those expectations is also subject to change due to preceding stimuli. For example, a decision-maker might expect an automated advisor to provide higher quality advice than a human and this expectation may be more salient (and reflected in behavior) when the automated advisor is "new" and compared to a previous human advisor. Previous research in both interpersonal and human-automation trust have typically manipulated advisor characteristics such as the expertise of the advisor in order to create certain performance expectations, but we only changed the advisor type and kept descriptions of the advisor's expertise constant. For continuing theory development regarding human versus automated advice, it is therefore important to recognize that differing decision-maker expectations come not only from immediate information gained/provided about the advisor, but also underlying assumptions about the qualities of automated versus human advisors. These underlying assumptions and expectations (made stronger when a contrast effect is present) may explain why we witnessed increased utilization of automated advisors following the replacement of a human.

The contrast effect and underlying assumptions explanation would be consistent with the perfect automation schema (Madhavan & Wiegmann, 2007b) as well, but so would an alternative explanation that participants in the automation-replaces-automation condition utilized the new automated advisor less because the previous automated advisor had broken the expectation of perfection. This alternative explanation also sheds light on why participants in this condition reported more negative emotions with the second advisor and decreased advice quality, even though it was the same type of advisor. After experiencing unreliable advice from a first automated advisor, participants become increasingly frustrated as the second automated advisor turns out to not be any better. Expectancy violations theory would suggest, however, that if the

expectation of perfection is broken for the second automated advisor, then the delivery of bad advice would produce less negative emotions because it is not an expectancy violation. Neither of these explanations can be ruled out with our data and they should be investigated by scholars moving forward. Our unique contribution to this literature is showing advisor expectations might be changed simply by changing the type of advisor from human to automated or vice-versa.

The differences in advice utilization following a mistake for automated versus human advisors is also consistent with past research on the perfect automation schema and we believe they result from the underlying assumptions about advisor qualities. For example, human advisors are expected to be fallible because at some level, any decision-maker knows that no human is perfect. Although advice utilization for automated and human advisors that replaced the opposite advisor type was similar in the trials preceding the bad advice, the reaction to it was not. Our results are particularity interesting because bad advice is delivered on the 6th trial, well after the expectation effects witnessed in the first three trials is no longer present. From a theoretical perspective, this suggests that certain expectations are also more persistent that others. It also importantly points out that there are multiple dimensions to the perfect expectation of automation – perfection means both performing better than a human and being consistent in that better performance. Another interesting aspect of our findings regarding the reactions to bad advice is that we did not find significant effects after the second delivery of bad advice. This is likely an effect of our study design where bad advice is delivered in 2 out of 3 consecutive trials. We simply did not have enough trials after the first instance of bad advice to understand how quickly utilization may recover to previous levels. Other scholars are studying trust repair processes in human-automation and interpersonal communication (de Visser et al., 2016; Tomlinson & Mayer, 2009), our result points to the need for more research in this area.

The decreased perception of advice quality when an automated advisor replaces a human advisor also points to the need for further research. It is possible that because decision-makers feel more reciprocity ("I owe something to my advisor") to human advisors, this produces a thankful feeling which is reflected in better evaluations of the advice. However, if true, this thankful feeling is not reflected in the behavioral advice utilization data. The social processes involved in interpersonal advice are thoroughly studied in advice situations on more personal situations in which a decision-maker is can benefit from advice as a form of social support (Feng & Burleson, 2008; MacGeorge, 2016), but even though our experiment involved essentially no social interaction we cannot rule out the possibility that social scripts were activated. When decision-makers in this study envisioned their advisor, they may have thought of a human going to the effort of creating advice whereas a non-sentient machine was putting forth no effort. We also cannot ignore that automated advisors that replaced automated advisors also produced negative emotions and evaluations for decision-makers. Our study design is limited in that we only administered the survey evaluation after the decision makers had experienced unreliable advice quality – we cannot therefore truly evaluate the decision-makers perceptions when the advice is good. The results may be a result of the expectation of good automation that replaces an unreliable human not coming to fruition, as opposed to just liking a human advisor more. In sum, both the behavioral data and survey results suggest there are multiple processes at play when one advisor replaces another – the study above is the first show that advisor replacement affects expectations of the new advisor that subsequently lead to behavioral and perceptual effects.

We also want to note some methodological contributions of our study. First, because we did not find a consistent preference for either advisor type, management forecasting may provide a largely neutral task in which to study further manipulations (although it should be noted that

advisor preferences for similar tasks have been found previously (Dietvorst et al., 2015; Önkal et al., 2009)). Second, our use of performance feedback after each trial may have attenuated the automation bias more quickly than it would occur in a more uncertain real-world environment where immediate feedback is not available. Third, the timing of our survey may explain some of the disconnect we see between our behavioral and psychological measures. Such discrepancies are not new to expectancy violation research, as it is difficult to capture what a person's actual expectations are in a given interaction, instead, expectations are often inferred by investigating the way participants respond to violations (Burgoon & Hale, 1988). Additionally, these findings suggest that the oft-cited computers are social actors (CASA) research (Nass & Moon, 2000) - that bases its claims of similarity between human and automated actors largely on behavioral data – may not be capturing the full picture of interacting with automation. Indeed, it appears that capturing behavioral and self-reported data reveals different dimensions of understanding interaction with automation compared to humans.

Our results have serious implications outside our experimental context. Automation is replacing humans in a tremendous number of professional roles such as wealth management and healthcare (Benedikt Frey & Osborne, 2013; Ford, 2015; Inthorn et al., 2015). We even see automated advice working its way into personal life, such as dating sites that suggest matches or algorithms on shopping sites that advise a shopper on gifts for family and friends. For business leaders who are introducing automation in place of human workers, they should understand that the remaining human workers who must work with this new automation may report negative perceptions of it, but this does not necessarily result in them not using it. On the other hand, any malfunction of the automation has serious consequences for continued use of it. Automation replacing humans seems to be something that everyone is talking about, but few people believe it

will be as simple as automation effortlessly taking tasks away from humans. People will have to work with their new machine companions and will often receive advice from them, the study here is the first to begin shedding light on what a complex process that will be. This is surely an important and rich area for continued theoretical development – we do not know what the future will bring – but we know there won't be any simple answers.

Limitations

Our study has several limitations. First, we were unable to have our participants actually interacting with human advisors, so the social aspect of the advice exchange was weak. Also, we did not manipulate the expertise level of the advisors, which has been found to affect automation bias in past research (Dzindolet et al., 2003). Third, the task used in this experiment, while predictive, is likely not perceived as being strongly intuitive in nature. Participants might actually think that automation is better suited to a forecasting task by applying statistical modeling compared to a human's simple perception of trends – we failed to measure this in our survey. Fourth, in any real organization replacing humans with automation, there is likely to be strong elements of organizational culture and much deeper relationships with human coworkers that will add layers of complexity to any analysis. Finally, we did not include a condition in which our participants kept the same advisor for all trials, making it difficult to model the effects of practice at the task or general utilization trends over time in absence of advisor replacement.

Chapter IV: Study III

Perception of Advisors and Utilization when Task Demonstrability is Manipulated

Utilized but Despised: Automated Advice versus Human Advice in High and Low Demonstrability Tasks

Andrew Prahl

University of Wisconsin-Madison

Abstract

We investigated the effects of task demonstrability and replacing a human advisor with an automated advisor (or replacing an automated advisor with a human advisor) on advice utilization and perception of advice. Participants were randomly assigned to make a series of forecasts dealing with either humanitarian relief planning (low demonstrability) or operations management (high demonstrability). In each condition, they received advice from either an automated advisor the whole time, a human advisor the whole time, or their advisor was replaced with the other type of advisor midway through. Automated advisors were used more at the start of the low demonstrability task, but there was no effect of advisor replacement on utilization of advice. Decision-makers rated humans as more expert, delivering more useful advice, and as more similar in thought and value similarity to the decision maker than the automated advisor. These perception effects were strongest when a human advisor was replaced by automation. Decision makers experience more negative emotions, lower feelings of reciprocity, and more desire to fault their advisor for mistakes when automation replaces a human. Results are related to other research comparing human-human communication with human-automation and situated in communication theory.

Introduction

On August 13th, 2014, a video titled *Humans Need Not Apply* was uploaded to Youtube and exploded in popularity, gathering over 1 million views within 3 days; it currently stands at over 10 million views (CGPGrey, 2014; Pagano, 2014). Detailing a future in which human labor is irrecoverably taken over by automation, the 15-minute video on the evolution of labor – and why the robot revolution is different – was described as "terrifying" and foretelling "an economic horror movie" by some commentators (Roggeveen, 2014, p. 1). Since then, public interest in the continued automation of human labor has only increased; it is becoming difficult to read the business section of a popular newspaper and *not* encounter an article discussing the future of work and automation.

Over time, discussion of automation has gradually shifted from topics like economics (i.e., Ford, 2015) to more subjective subjects such as the ethics or moral implications of machines making certain decisions (Freytas-Tamura, 2018; Prahl & Enright, 2017). The latter topic is also getting more attention by academic scholars in a number of fields including human factors, psychology, and philosophy (Coeckelbergh, 2010; Kahn et al., 2012; Malle & Scheutz, 2015). Certainly, automation is increasingly being used for decisions that have more serious and perhaps ethical implications, such as how to treat severe cases of cancer (Malin, 2013). Rarely, though, is automation put in a sole decision-making role in these situations, instead it can best be thought of as an advisor to a human decision-maker.

The scholarly community in a number of fields such as communication and social psychology has conducted considerable research on the interpersonal process of advice giving and receiving (for review, (Bonaccio & Dalal, 2006; MacGeorge & Van Swol, 2018) More recently, the study of automated advice has also accelerated (Alvarado-Valencia & Barrero,

2014; Hoff & Bashir, 2015). However, less research exists that directly compares automated advice to human advice, especially in situations where automation replaces humans or for decisions with more subjective and less demonstrable consequences. Additionally, given that the thought of automation replacing humans concerns many people, it is surprising how little research exists on the psychological state of people who must work with the new machines that have replaced a human. The purpose of the study here is to investigate (1) the effects of task demonstrability on advice utilization from human and automated advisors and, (2) how perceptions of advisor and advice are affected by task type, advisor type, and advisor replacement. Finally, (3) we wish to see if different perceptions of the advisor are more important predictors of advice utilization in different decision-making tasks.

People frequently seek the advice of others when making important decisions in both professional and personal contexts. Employees in organizations frequently turn to peers or advice networks for input about problems (Agneessens & Wittek, 2012; MacGeorge & Van Swol, 2018), and individuals do the same for personal problems, such as college and career decisions (Carlson, 2014). There is a considerable and growing body of research on the factors that lead decision-makers to both seek and utilize advice once received. Our study here is focused on how advice utilization, evaluations of advisors, and decision-maker feelings differ depending on if the advisor is human or automated, and how task may moderate these differences. In this section we begin by conceptualizing differences in task demonstrability. Next, we develop hypotheses as we review relevant findings and theory from advice and decision-making research.

Task Demonstrability

A continuum of decision-making tasks anchored by intellective and judgmental tasks was explicated by Laughlin and Ellis (1986). Although used for group decision-making research, this

distinction has become important in interpersonal advice research (Bonaccio & Dalal, 2006; Van Swol, 2011). Intellective tasks are distinguished by having a demonstrable answer that all parties can understand. For example, an algebra problem has a correct answer and any advice provided to a decision-maker suggesting a correct answer is demonstrably correct or incorrect provided the advice receiver shares the same conceptual system with the advisor. On the other hand, judgmental tasks involve uncertain future states or personal opinions and there is no demonstrably correct advice. Examples of such tasks include deciding if an artistic performance warrants an award or what level of pain inflicted by a medical treatment is acceptable given the expected benefit. In addition to many group decision-making studies (e.g., Laughlin & Ellis, 1986; Stasser & Stewart, 1992), several interpersonal advice studies have varied decision-making tasks on the intellective-judgmental continuum (Logg, 2017; Tzioti, Wierenga, & Van Osselaer, 2014; Van Swol, 2011).

The difference between intellective and judgmental decisions is related to the difference between objective and subjective decisions. Scholars in morality and ethics have long debated the tension that exists between objective judgment (based on factual reasoning) and subjective judgement (based on value judgments) when making moral decisions (Cortese, 1990; Crowley, 1968). When decision-makers receive advice on problems with ethical consequences, it can heighten the salience of accountability and blame sharing (Bonaccio & Dalal, 2006; Grossman, 2009). Additionally, when decisions have moral consequences, decision-makers may utilize advice to engage in "moral distancing," which is an attempt to absolve oneself of responsibility for a suboptimal or morally wrong outcome (Cummings, 2006; Grossman, 2009). Groups also behave differently when decisions involve moral consequences such as employing the use of different decision processes than when decisions involve more use of facts. For example, jury

judgments about the guilt of a defendant tend to be based on factual reasoning, whereas less demonstrable judgments about the appropriate punishment for a guilty party are based on norms and values (Costanzo & Costanzo, 1992; Kaplan & Miller, 1987). Additionally, different decision rules such as unanimous decision versus majority decision produce differing effects on group member satisfaction: less demonstrable decisions decided via majority rule are less satisfying than when produced under unanimous rule (Kaplan, Schaefer, & Zinkiewicz, 1994; Tindale, Kameda, & Hinsz, 2003). This may be because a person who is on the losing side of a majority with a demonstrable task can still be correct and feel secure in that knowledge. However, being on the losing side of a low demonstrability task that involves value judgments may cause the perception that others disapprove of one's own values and beliefs, and these feelings of rejection hurt (Gausel, Leach, Vignoles, & Brown, 2012; Wittenbaum, Shulman, & Braz, 2010).

Advisor Expertise

A decision-making task's position on the demonstrability continuum between intellective and judgmental clearly has the potential to affect advice processes such as advice utilization and perceptions of the advisor. Different types of tasks require different skills and advisor abilities in order to deliver effective advice. One of the strongest and most robust effects in advice research is that perceptions of advisor expertise directly affect rates of advice utilization (Sniezek & Van Swol, 2001; Van Swol & Sniezek, 2005). Decision-makers have such a propensity to evaluate expertise that it does not always have to be overtly established; decision-makers will infer advisor expertise from cues such as age, experience, confidence, or seniority (Feng & MacGeorge, 2006; Van Swol & Sniezek, 2005). An advisor's level of expertise is often perceived in relation to that of the perceived expertise of the decision-maker themselves, meaning expertise is more likely to influence decision-makers who believe themselves to be novices and discounted

by decision-makers who are highly confident in their own abilities (Harvey & Fischer, 1997; Sniezek & Buckley, 1995). The effect of expertise is strong enough that greater or lesser expertise that is only implied by status relationships can be a confound in organizational field studies (Agneessens & Wittek, 2012), whereas many laboratory studies avoid this by controlling status relationships (e.g., MacGeorge, Guntzviller, Hanasono, & Feng, 2013).

The expertise of an advisor is tied to the domain and decision-making context – after all, a decision-maker would not trust even the most experienced and respected basketball player's advice over that of a doctor when faced with a medical problem. Studies of parental advice to emerging adults, for example, found that the type of problem affects the likelihood of decision-makers to view the parent as a competent advisor and utilize advice (Carlson, 2014). Such findings are in line with the concept of "expert power" which is only possessed when people have skills and knowledge that are relevant to the problem (Bonaccio & Dalal, 2006). Therefore, even though expertise is an important variable in advice research, it is inherently tied to the type of decision task for which advice is sought. In the research presented here, we manipulate both the advisor type (human/automation) and the demonstrability of the decision-making task. Decades of research on technology acceptance, human-automation trust, and the growing field of automation ethics provides some insight as to how perceptions of automation attributes differ from the assessment of humans. These differing perceptions may, in turn, interact with type of task to moderate advice-utilization.

Perceptions of Automation and Perceptions of Humans

The comparison of human versus automated advice in tasks of varying demonstrability, especially those involving moral decisions, introduces an interesting question about the match between advisor characteristics and the context of a decision. The majority of automated advice

research has used highly demonstrable tasks as experimental stimuli, such as what the next number is in a mathematical sequence (de Visser et al., 2016) or if military equipment is present in an aerial surveillance photo (Rice & Geels, 2010). There exists little research to guide our assumptions about utilization of automated advisor advice on less demonstrable tasks, but some scholars have investigated perceptions of various forms of automation (i.e., robots) for varying roles in society (Coeckelbergh, 2010; Katz & Halpern, 2014; Takayama, Ju, & Nass, 2008). Robots are often perceived as being more suitable for roles that do not require emotion or sensitive communication, but more suitable for roles that require memorization and unselfish service-orientation (Takayama et al., 2008). However, this perception of suitability can be affected by automation features such as the anthropomorphism or social features of the robot (Katz & Halpern, 2014). The more social cues that are provided by a robot sometimes lead to perceptions that it is closer to humans in abilities, although being "too human" can lead to a backlash effect like the well-known uncanny valley (Burleigh, Schoenherr, & Lacroix, 2013).

Because automation humanness seems to matter, it is logical to assume that decision-makers do have fundamentally different perceptions of human and automated advisors. Several theoretical perspectives in human-automation trust literature highlight the importance of the different expectations between humans and automation (Dzindolet et al., 2003; Madhavan & Wiegmann, 2007b). For example, automation is expected to be high performing but invariant, whereas humans (even if not as high performing) are adaptable and are expected to learn from their mistakes. Additionally, mistakes are less expected from automation in general because decision-makers do not see automation as susceptible to biases and emotions that plague human judgement (Merritt et al., 2015). On the other hand, people know that other people are not perfect, humans make mistakes. Effects of these expectations include a bias towards utilizing

automation more than human advice on certain tasks and less propensity to abandon human advisors when they deliver unreliable advice quality (Dzindolet et al., 2003; Madhavan & Wiegmann, 2007b). Several authors in human-robot trust have investigated questions of emotion more deeply, but there is little evidence that humans consider automation to possess the attributes needed to understand human emotion (which is different than simply possessing emotion). For example, even when some participants in experiments have perceived a robot to have thoughts and desires, the same participants do not believe that the robot could become their intimate friend or that the robot could provide them comfort in a time of sadness (Kahn et al., 2012, 2006).

If human and automated advisors are perceived as having differing fundamental attributes, it will affect the perception of either advisor's capabilities. It follows that when an advisor is not assessed to have capabilities required for a certain task, the assessments of the advisor that are dependent on the task will be affected as well. As established earlier, advisor expertise is one of these advisor attributes that is assessed in relation to the decision-making task. In this study, we do manipulate the task via task demonstrability. Because less demonstrable tasks require value judgments that, in turn, are tied to emotion and subjective evaluation (Horberg, Oveis, & Keltner, 2011), we believe that the underlying assumption that automation lacks emotion will lead to lower assessment of advisor expertise in less demonstrable tasks.

Hypothesis 1a: Automated advisors will be perceived as having less expertise in less demonstrable decision tasks than more demonstrable tasks.

Perceptions of advisor expertise also affect perceptions of the advice itself. If advice comes from non-expert sources, it is perceived as less useful. This effect is both logical and established in past advice research (Bonaccio & Dalal, 2006; Sniezek & Van Swol, 2001), so

much so that even implied expertise, such as having to pay for an advisor's advice, is found to increase perceptions of usefulness (Yaniv, 2004).

Hypothesis 1b: Advice from automated advisors will be perceived as being less useful in less demonstrable than more demonstrable decision tasks.

Advice may also be perceived as less appropriate if a decision-maker feels the advisor does not possess the expertise needed. On decision-tasks of low demonstrability, such as handling a relationship conflict, advice from close friends may be perceived as more appropriate than advice from acquaintances because only a close friend could understand the consequences and feelings of the decision-maker (Feng & MacGeorge, 2006; Feng & Magen, 2016). Although our task below is not designed to be personal, because it is low demonstrability it is more likely to elicit an emotional response from the decision-maker. An inability to understand this emotional response will therefore lead to lower assessed advice appropriateness.

Hypothesis 1c: Advice from automation advisors will be perceived as being less appropriate in less demonstrable decision tasks.

Task, Advisor, Utilization, and Perception

A large number of theories that describe human social behavior recognize that a person's perception and beliefs do not always lead to corresponding behavior (Ajzen, 1991; Gilbert, Fiske, & Lindzey, 1998). Disconnects between beliefs and behavior are fundamental to understanding well established phenomena like cognitive dissonance and judgmental biases in self-evaluation (Amos Tversky & Kahneman, 1974). For example, a decision-maker may receive advice from a source they consider to have high expertise, but they may feel the desire to only use their own judgement and discount the high quality advice (Van Swol, Paik, & Prahl, 2018). On the other hand, a source that is considered to have low expertise may be utilized if the

decision-maker feels there is a risk of hurting the feelings of the advisor (Mayer et al., 1995; Yaniv & Kleinberger, 2000). In the experiment below, we attempt to control factors like task that moderate the relationship between perceptions of the advice and advisor expertise, therefore we expect to see the effects of hypotheses 1a-1c reflected in advice utilization rates as well.

Hypothesis 2: Automated advice will be utilized less than human advice, in general, when the decision is less demonstrable.

Previous research that has manipulated task demonstrability in interpersonal advice suggests that there are other perceptions of advisor characteristics beyond expertise that are of interest, especially as tasks become less demonstrable. An important advisor characteristic in judgmental tasks is the degree to which the decision-maker believes the advisor shares their thought process and values. For example, when giving advice on what movie would be enjoyable – decision-makers preferred advice from advisors they felt were more similar to them (Bonhard, Harries, McCarthy, & Sasse, 2006; Van Swol, 2011). The same effects of advisor similarity did not hold when the task was a highly demonstrable math task (Van Swol, 2011). When making decisions on matters of taste, both advisor demographic and behavioral similarity drive influence on decision-makers (Yaniv, Choshen-Hillel, & Milyavsky, 2011). This is consistent with a large amount of social psychological research indicating that people like to cooperate and seek help from people who they perceive to share values and beliefs (Berscheid & Reis, 1998). Similarly, value congruence is also a fundamental part of many theoretical models of trust (Mayer et al., 1995). Like advisor expertise, the effect of advisor similarity may depend on the domain and context of the problem being faced – there must be a match between values and the problem. For example, an employee may utilize the advice of a coworker when dealing with an organizational problem – where both decision-maker and advisor value the success of the organization – but not

on a romantic relationship problem where the decision-maker and advisor have differing values. Because the research below manipulates task demonstrability, we expect to observe perceptions of advisor similarity being more predictive of advice-utilization in less demonstrable tasks.

Hypothesis 3a: Perceptions of advisor thought process and value similarity will be more associated with advice utilization in less demonstrable decision tasks.

We expect to find other perceptions of the advisor, specifically those related to advisor expertise, to predict advice utilization more when tasks are more demonstrable. As reviewed above, a large amount of automated advice literature finds a bias to utilize automated advice more and this is tied to an expectation of higher performance and reliability. This effect, while robust, appears to be limited to tasks that are high in demonstrability such as signal detection or deciphering a numerical sequence.

Hypothesis 3b: Perceptions of advisor expertise and reliability will be more associated with advice utilization in more demonstrable than less demonstrable decision tasks.

Our final hypothesis is deductively reasoned from hypotheses 2 and 3 presented above. Together, our reasoning for these hypothesis leads to our expectations that (1) human advice will be used more than automated advice on less demonstrable decision tasks, and (2) value similarity will predict advice utilization on less demonstrable decision tasks. If human advice is to be utilized more, therefore, perceptions of advisor similarity should be higher for humans in less demonstrable decision tasks than in more demonstrable decision tasks. Using the same logic, we predict that perceptions of advisor expertise and reliability will be higher for automated advisors in more demonstrable decision tasks. Our assumptions about advisor expertise are essentially covered in hypothesis 1a, however, reliability represents a different dimension of advisor ability that is more closely related to concepts like predictability (Barber, 1983) or consistency (Butler,

1991). The research reviewed above highlights the importance of considering reliability as a separate construct from expertise because even if the performance expectation of a human and automated judge is equal, the human is expected to be fallible (Madhavan & Wiegmann, 2007b).

Hypothesis 4a: Human advisors will be perceived as having more thought process and value similarity with decision-makers in less demonstrable decision tasks.

Hypothesis 4b: Automated advisors will be perceived as more reliable in more demonstrable decision-tasks.

Advisor Replacement

In the research study below, we manipulate advisor type, task type, and also advisor replacement. There is essentially no interpersonal advice or automated advice research that has been conducted specifically to test the effects of advisor replacement, but because the perception of humans and automation relies on different underlying assumptions, perceptual effects related to the comparison of two stimuli may be applicable to guide our expectations. We first should consider that the decision to utilize advice is highly dependent on the decision-maker's comparison of an advisor to themselves (Bonaccio & Dalal, 2006). Expertise of the advisor, for example, is evaluated relative to the self-efficacy and confidence of the decision-maker (Sniezek & Van Swol, 2001). Other perceptions such as value and thought similarity also inherently involve comparison. When one advisor is replaced by another type, this may also elicit a comparison of the two advisors that makes the perceived attributes of both more salient. Such an effect would fit into existing literature on contrast effects (Palmer & Gore, 2014).

Contrast effects are widely studied in perception research and describe the process by which exposure to one target of evaluation can change the evaluation of targets presented subsequently. For example, unattractive faces are rated as more unattractive if the evaluator is

shown an attractive face before the unattractive one (Wedell, Parducci, & Geiselman, 1987). Evaluation of the credibility of a politician can be affected positively or negatively if the evaluator thinks about a different politician first (Schwarz & Bles, 1992). Contrast effects extend to more global evaluations of target individuals, such as job performance (Palmer & Gore, 2014). We are not aware of any literature to suggest that contrast effects will not extend to the evaluation of an advisor. If a contrast effect is found, we expect it to result in our hypothesized effects of advisor type becoming stronger. If a decision-maker is presented with a new automated advisor after gaining experience with a human advisor, it may result in an even stronger perception of invariance and exaggerate expectations of high performance. The opposite will occur for human advisors that replace automation. To be clear, we only manipulate advisor replacement, we do not replace one task with one another; our hypothesis below therefore only is intended to cover effects that are driven by advisor perception. Because we expect a link between advisor perception and utilization, we expect the contrast effect to be reflected for utilization as well. For brevity, we summarize these effects in the below hypothesis.

Hypothesis 5a: Automated advisors will be evaluated as less expert, useful, appropriate, and similar in less demonstrable decision tasks when they replace a human advisor compared to if they replace another automated advisor.

Hypothesis 5b: Human advisors will be evaluated as more expert, useful, appropriate, and similar in less demonstrable decision tasks when they replace an automated advisor compared to if they replace another human advisor.

Hypothesis 5c: Automated advisors will be evaluated as more expert, useful, appropriate, and similar in more demonstrable decision tasks when they replace a human advisor compared to if they replace another automated advisor.

Hypothesis 5d: Human advisors will be evaluated as less expert, useful, appropriate, and similar in more demonstrable decision tasks when they replace an automated advisor compared to if they replace another human advisor.

Hypothesis 5e: In more demonstrable tasks, when an automated advisor replaces a human advisor, the automated advice will be utilized more than in conditions when an automated advisor replaces another automated advisor.

Hypothesis 5f: In less demonstrable tasks, when a human advisor replaces an automated advisor, human advice will be utilized more than in conditions when a human advisor replaces another human advisor.

Decision-Maker Emotions

Our earlier discussion of emotions primarily discussed a decision-maker's perception that an advisor possesses emotions or at least the capability to understand emotions. But perceptual processes themselves are affected by emotions, and interpersonal advice research has shown decision-maker emotions to have substantial effects on perceptions of the advisor, advice, and utilization of advice (MacGeorge et al., 2013; Tost, Gino, & Larrick, 2012). In interpersonal advice research using demonstrable tasks, researchers have manipulated decision-maker emotions, finding that the induction of other-directed negative emotions (i.e., anger, frustration) resulted in less advice utilization, while other-directed positive emotions (i.e., love, gratitude) resulted in more utilization (Gino & Schweitzer, 2008). Such effects were found in research using less demonstrable tasks (de Hooge, Verlegh, & Tzioti, 2014). However, for self-directed negative emotions (i.e., anxiety, shame), participants are more likely to utilize advice, presumably because emotions like anxiety or shame lower decision-maker's confidence and, hence, increase their tendency to utilize advice (Gino, Brooks, & Schweitzer, 2012).

Research on the effects of decision-maker emotions and automated advice is less conclusive about the effects of emotions on advice-utilization. This is largely because advisor anthropomorphism can have such a strong effect on emotion, and there are large differences in anthropomorphism for automation. For example, consider that the comparison of research using a humanoid robot to that using a text-based forecasting support system is difficult to compare (Hancock et al., 2011; Hoff & Bashir, 2015). Literature that is specifically focused on the acceptance of technology does provide evidence that good experiences with automation results in positive emotions and the expectations of more positive emotions in future interactions with technology, with the opposite effect for negative emotions (Venkatesh, Morris, Davis, & Davis, 2003). Applied to advice utilization, it would imply a similar effect of emotion on utilization intent, but because we do not manipulate decision-maker emotion in our study, we are only able to predict potential effects that result from our manipulations of interacting with different advisor types and on tasks of differing demonstrability. Thus, we are left with insufficient evidence to propose hypotheses related to decision-maker emotions, but our study design is ideally suited to begin investigating what emotions may be produced by interacting with human versus automated advisors.

RQ1: Are decision-maker emotions affected by task and advisor type?

We take special interest in the emotion of anxiety because it has been found to produce a unique effect on utilization. Unlike other-directed negative emotions, anxiety can lower decision-maker confidence, leading to increased utilization of advice and an inability to discern good advice from bad advice (de Hooge et al., 2014; Gino et al., 2012). Our experimental manipulation of task could cause more anxiety due to one decision being perceived as harder to make and requiring more compassion (see pilot results below).

RQ2: Do decision-makers feel more anxiety as a result of task or advisor type?

We also take a special interest in emotions that imply a perception of socialization with the advisor. There is conflicting evidence regarding to what degree humans perceive they socialize automation, with some research suggesting that humans socialize with automation to the extent that human-automation interaction is essentially a form of interpersonal interaction (Nass & Moon, 2000). Other research, such as that on robot suitability for job roles, suggests that socialization may occur but to a limited extent (Savela, Turja, & Oksanen, 2017). Reciprocity is one emotion that may be unique to humans because it implies thankfulness for another human's generosity and benevolence. Some scholars assert that benevolence requires perception of emotion in the target of benevolence (Livnat Yuval, 2004) and that automation is typically perceived to lack this (Gefen et al., 2003; Kahn et al., 2006).

RQ3: Are decision-maker feelings of reciprocity towards the advisor affected by task and advisor type?

Finally, we examine attributions of fault for mistakes. Our interest stems first because of the differing expectation between human and automated advisors outlined above. If humans are expected to be fallible, it may result in decision-makers generally finding less fault in human advisor's mistakes. Making a mistake as a human is not something that automatically means they have performed in an unacceptable manner. The opposite may occur for automation because automation is expected to be reliable, even perfect. On the other hand, there is conflicting evidence regarding the question of humans perceiving intention and free will in automation (Friedman, 1995; Kahn et al., 2012; Shen, 2011), and it is possible that decision makers will consider free will as a prerequisite to placing fault, because how can you fault something if it had

no choice? Thus, there is not clear research of whether humans or automation will be faulted more for a mistake.

The second reason we are interested in fault is more related to our task manipulation. It is possible that decision-makers will fault automation to a greater degree than human advisors because fault is related to blame. Placing blame on a human may hurt their feelings, and even a decision maker who is experiencing negative emotions resulting from a bad decision does not want to hurt another human's feelings. But an automated advisor has no feelings that can be hurt, making it easier to place blame. However, robot-ethics research suggests that this tendency, if true, may change if a decision has moral consequences. Our low demonstrability decision task in this experiment has consequences that result in the loss of human life, and though the measurement of what is perceived as "moral" is complicated (Bandura, 1969; Jordan, Leliveld, & Tenbrunsel, 2015), it is not unreasonable to assume that decision-makers could sense moral implications in this type of decision. Some research shows many humans have a discomfort with placing blame on automation for making decisions with moral implications because many people do not perceive automation to possess moral accountability. For example, the most common reason given for not allowing military drones to autonomously select and fire upon targets is because it introduced problems with moral accountability for mistakes (Moon, Danielson, & Loos, 2012). Yet other research has shown that some people do hold some forms of automation morally accountable (Malle & Scheutz, 2015), but it depends on the form of automation. A social robot was seen as morally wrong for malfunctioning, while the same participants said they would not hold a vending machine accountable for a malfunction (Kahn et al., 2012). In sum, our experimental manipulations offer a unique opportunity to investigate attributions of fault both between advisor and task demonstrability conditions.

RQ4: Is the attribution of fault for mistakes affected by task and advisor type?

Method

Design Considerations

Our literature review revealed the need for careful design. For example, research on how the perceived automation suitability for societal roles is affected by anthropomorphism has important design implications. First, we do not anthropomorphize the automation in order to provide the cleanest manipulation of advisor type. second, we limit the social aspects of the advice exchange process; there is no direct interaction with either advisor type and advice is delivered in a simple text format. Additional design implications are based on above review of advisor expertise – we avoid creating implied expertise by clearly introducing the human and automated advisors as having equivalent expertise (see study 3 method, for advisor introductions). We also precisely control the accuracy of advice to rule out the confound of an advisor actually being better at a decision-making task. Our design therefore is optimized to discover differing assumptions that people have about the attributes of automation versus humans on tasks of different demonstrability.

Participants

Participants were recruited through Amazon's Mechanical Turk (MTurk) service and were required to be United States citizens over the age of 18. Seven percent of participants were between the ages of 18-25, 75% between the ages of 25-55, the remaining 18% were over the age of 55. Participants were 60% male and 40% female; two participants did not report their sex. Approximately once per hour over the course of one week, a group of eight survey participation slots were released to the MTurk worker portal for survey respondents to complete the survey. A random number generator determined assignment to conditions for the group of eight responses:

task (low demonstrability/high demonstrability), first advisor (human/automated), and advisor replacement (similar advisor/different advisor). We released the surveys in blocks of eight so that we could constantly update worker qualifications to exclude participants from completing the survey twice.

A total of 746 participants completed the study. In the high demonstrability task *n*=356, there were *n*=82 participants in the automated advisor replaced by automation (ArA), *n*=81 in human advisor replaced by human (HrH), *n*=90 in automated advisor replaced by human advisor (ArH), and *n*=85 in human advisor replaced by automated advisor (HrA). After the data was cleaned for incomplete or fraudulent responses or failure to answer several attention check questions, the responses for analysis were *n*=77 (A), *n*=76 (H), *n*=82 (ArH), *n*=73 (HrA), (*N*=308). In the low demonstrability task *n*=390, there were *n*=84 participants in the automated advisor replaced by automation (ArA), *n*=85 in human advisor replaced by human (HrH), *n*=112 in automated advisor replaced by human advisor (ArH), and *n*=119 in human advisor replaced by automated advisor (HrA). After the data was cleaned, the responses for analysis were *n*=80 (A), *n*=73 (H), *n*=106 (ArH), *n*=108 (HrA), (*N*=367). All 20 graphs and 20 forecasting scenarios were randomly assigned into four sequences using a random number generator, the sequences were then randomly assigned within each advisor condition.

Task

A judgmental forecasting task was used so results could easily be related to the large amount of previous research on forecasting support systems (e.g., Fildes, Goodwin, & Lawrence, 2006; Önkal et al., 2009). Participants completed 20 forecasting tasks consisting of a graph of past data relating to a variety of issues. Each graph presented two lines representing the percentages of two items used or event occurrences, and the two lines symmetrically trended in

opposition to one another. In the high task demonstrability condition, the forecasting tasks were related to hospital operating room management. An example of a forecast to make would be "There are two sizes of latex gloves, what percentage of gloves ordered next month should be for hand sizes XS-M versus L-XXL?" The graph would display two lines representing the percentage of all gloves that were consumed in either size, monthly for the past several years. See the Appendix for screenshots of the task. In the low task demonstrability condition, the same graphs were displayed, but the scenarios dealt with humanitarian relief planning. For example, "The below graph displays deaths due to starvation in the two largest refugee camps in Uganda over the past several years. How should we split the percentage of our next food shipment to Uganda between the two camps?" In the high demonstrability condition, all forecasting tasks remained specific to the operating room unit of the hospital. In the low demonstrability condition, they remained specific to humanitarian relief planning. We picked these domains partially to control for participants having personal intuition for what the outcomes would be. Previous forecasting studies, for example, have often used students enrolled in economics classes to forecast stock prices. This may lead to overconfidence and egotistic discounting of advice as the participants consider themselves experts, or the participants have specific knowledge about the shortcomings of forecasting support systems (Goodwin, Fildes, Lawrence, & Nikolopoulos, 2007; Önkal et al., 2009). Additionally, recipients may underutilize advice when they perceive high self-efficacy in the domain (Lawrence, Goodwin, O'Connor, & Önkal, 2006), and our choice of forecasting tasks minimizes this risk.

Procedure

Participants signed up for the study via the MTurk portal. After digitally signing a consent form, participants read a short introduction about the types of tasks they would be given.

The manipulation of advisor type was simple and similar to past studies (Dzindolet et al., 2002; Madhavan & Wiegmann, 2007a; Önkal et al., 2009). Participants were either told at the opening that the advice would come from an algorithmic software program, or an experienced surgeon at the hospital, Logan (in the low demonstrability condition, the advisor worked for the United Nations Refugee Agency). Each advisor was introduced to the participant with a short paragraph introducing them as expert advisors, we did this to replicate real world scenarios in which advisors are unlikely to be sought unless they are experienced. After the first 10 trials, the participants filled out a survey assessing their advisor and then were told that they would be getting a new advisor for the remaining 10 forecasts. Advisor descriptions were as followed in the high demonstrability condition (followed by the low demonstrability condition):

> Your advisor today is an algorithmic computer program called OptiLytics. OptiLytics is a software program used by the Gain Healthcare System to help with forecasting. The statistical models in OptiLytics have been built using 10 years of past Gain Healthcare data, as well as some data from the Center for Disease Control in the United States.
>
> Your advisor today is: Logan Girard. Logan is a medical doctor who has been working for Gain Healthcare for 10 years doing operating room management and hospital operations. Prior to joining Gain, Logan gained experience in healthcare management with the Center for Disease Control in the United States.
>
> Your adviser today is an algorithmic computer program called ReliefLytics. ReliefLytics is a computer program used by the UNHCR to help with forecasting. The statistical models in ReliefLytics have been built using 10 years of past UN data, as well as some data from the Center for Disease Control in the United States.

> Your adviser today is Logan Girard. Logan is a medical doctor who has been working for the UNHCR for 10 years doing camp management and emergency relief. Prior to joining the UNHCR, Logan gained experience managing medical crises in developing nations while working for the Center for Disease Control in the United States.

When advisors were replaced midway through the survey, the introduction text was preceded by "Due to time constraints, we were not able to get OptiLytics/Logan's advice for every forecast. Therefore, you will have a new advisor to help you on the remaining tasks. Your new advisor is…" In the conditions where a human advisor was replaced by another human advisor or an automated advisor with another automated advisor, the names of the new advisors were OperationAid/Cameron, and the corresponding advisor introductions were nearly identical to those above.

Participants were compensated at the federal minimum wage ($7.15/hr) rate assuming the survey would take 45 minutes. In order to incentivize accuracy, participants were informed that they could win a $100 Amazon giftcard based on their performance if they finished all of the forecasting tasks with the lowest mean percentage error among all participants. No performance record or implementation history was provided or implied for either advisor because past performance and the perceived reasons for system implementation may cause participants to start the experiment with biased expectations (Dietvorst et al., 2015; Madhavan & Wiegmann, 2007a).

After viewing the graph of past data, participants entered an initial forecast (text entry). Then the screen advanced to an "advice" screen in which the advice of either the human or algorithmic advisor was presented. The advice was presented in a "point forecast" format (i.e., an exact number). Participants then had the opportunity to consider the advice and enter a revised forecast on the screen.

After each task was complete, participants received feedback on the accuracy of their forecast. A screen appeared that showed the absolute percentage error of the participant's revised forecast as well as the advice forecast (this allowed them to directly compare their performance to the advisor). Furthermore, the participant's mean absolute percentage error across all previous revised forecasts was computed and displayed on the feedback screen in a very large red font. Participants were reminded after several trials on the feedback screen that the participant with the lowest mean absolute percentage error would win a $100 gift card. Thus, it is expected that the participants would feel some degree of frustration/gratitude if the advice caused them to decrease/improve their forecast accuracy. Finally, in the high demonstrability condition, the participants percentage error was multiplied by 1000 and presented as (for a 1% error): “This forecasting error is estimated to have cost the hospital $1000.” In the low demonstrability condition, the percentage error was multiplied by 100 and displayed as: “This forecasting error is estimated to have resulted in 100 deaths.” This was to reinforce that the decisions had either financial consequences or consequences resulting in adversity to humans.

Participants were walked through the process of making a forecast and entering their confidence during one “warm up” forecast. Directions appeared on the screen reminding participants of the proper format for entering forecasts and also directing their attention to where and when the advice would appear. After the warm up, participants proceeded to the first of 10 forecasting trials. The survey was programmed such that the advice was always the exact percentage better (or worse) than the participants forecast on each trial, regardless of which graph or forecasting scenario was randomly presented. This avoided a problem common in repeated measures studies in that some participants simply, through luck or skill, perform

particularly good or bad on early measures. In our study, every participant experienced the exact same accuracy of advice on each trial.

The first 10 trials were performed to set the stage for the second group of 10 trials where our research interest in advisor replacement lies. Advice on trials 1-5, 7, 9, and 10 was better than the participants' forecast. But, to avoid the advisors always being better than participants, which could artificially inflate quality ratings and also reduce ecological validity, on the 6th and 8th trials we gave advice that was worse than the participants' forecast. Thus, any utilization of the advice would make their forecast worse (and this was clear on the feedback screen that the advice hurt accuracy). Thus, utilization on the first 6 trials occurs when the participant is experiencing reliable and consistently good advice. Trials 7-10 occur when the advice is not always reliably good. The first block of 10 trials was performed to set the stage for the second block of trials where our research interest in advisor replacement lies.

Measures

To measure advice utilization, a "SHIFT" variable used in previous forecasting studies was used to directly compare human versus automated advice utilization. We chose to use this measure first to maximize commonality with previous forecasting research that used it as a measure of trust (Önkal et al., 2009) and also because the shift variable is a variant of the "weight of advice" (WOA) variable commonly used in advice research (Bonaccio & Dalal, 2006; Harvey & Fischer, 1997). SHIFT is computed via the following equation:

$$\frac{\text{Judge revised forecast} - \text{Judge initial forecast}}{\text{Advisor forecast} - \text{Judge initial forecast}}$$

After shift values were calculated, z-scores were calculated. Any shift score that was more than 4 standard deviations from the mean was checked for obvious type-o's (i.e., 661

instead of 66.1). After type-o's were fixed, the shift scores were then truncated to the theoretical range of 1 (full advice utilization) and 0 (no advice utilization). This truncation is similar to past research (e.g., de Hooge et al., 2014; Harvey & Fischer, 1997). We feel this was needed because occasionally a participant would overshoot their adjustment in the direction of the advice resulting in a shift score over 1 (or alternatively, move away from the advice to a negative shift score). Consistent with past research using shift scores, such occurrences were less than 5% of the data, but such scores can bias mean estimates of advice utilization, especially outlier scores. Furthermore, this operationalization of advice utilization is a continuous measure with full advice utilization on one end and no utilization on the other; we do not believe that moving away from an advisor's estimate (i.e., initial forecast is 50, advice is 52; forecast is revised to 49) constitutes something beyond "no" advice utilization.

A questionnaire was administered after the first set of ten trials and after the final ten trials with the replacement advisor. The questionnaire measures were identical and measured perceptions of advice quality and advisor quality on a sematic differential scale. Additionally, Likert survey questions measured emotions when receiving advice, trust of advisor, similarity (value, social norm, and thought process) to advisor, and perceptions of advisor effort. The survey was constructed from items from previous advice literature (MacGeorge et al., 2013). Positive emotions were measured with four Likert questions on a 1 (*not at all*) to 5 (*extremely*) scale for four positive emotions: Appreciative, Happy, Grateful, Thankful. The negative emotion scale was composed of Mad, Frustrated, Annoyed, Irritated. The four positive and negative emotion questions produced sufficient reliability (positive: $\alpha = 0.940$, negative: $\alpha = 0.943$), and the mean was used as an index of positive/negative emotion. Eight sematic differential questions were used to measure advice quality (e.g., thoughtful, careless); the eight items achieved

sufficient reliability ($\alpha = 0.811$) and is hereon presented as an index of advice quality. Finally, in order to keep the survey a reasonable length, a pair of questions was asked to assess feelings of reciprocity to the advisor (i.e., "I feel like I owe something to my advisor for their help").

Results

Manipulation Checks

To confirm that participants perceived the tasks differently, we conducted independent samples *t* tests on our manipulation check questions between tasks. Participants indicated their agreement with questions on a 7-point scale, for example, "The decisions I made were more about human life than money," and "The decisions I made were more about money than human life." Participants perceived the humanitarian task as being significantly more about human life than money, $t(669) = 22.980$, $p < 0.001$, while management tasks were perceived more about money than human life, $t(669) = 9.113$, $p < 0.001$. The humanitarian task was also perceived significantly more as having "no right answers," $t(668) = 2.963$, $p = 0.003$; and being more relevant to all humans than management tasks, $t(668) = 4.162$, $p < 0.001$. But, there were no significant differences on perception of the task being more relevant to the participant personally $t(669) = 0.083$, $p = 0.934$. Finally, participants in the humanitarian condition indicated more agreement with the statement that "someone must be a compassionate person to make the decisions on a task like this," significantly more than participants in the management condition, $t(667) = 3.022$, $p = 0.003$. Means and standard deviations are displayed in table 1.

Table 1: Manipulation checks independent t tests

Statement	Humanitarian Task	Management Task	Mean Difference	*p*	*d***
Task is more about Human Life	*M* = 5.573 *SD* = 1.381	*M* = 2.938 *SD* = 1.554	2.634	0.000*	1.845
Task is more about Money	*M* = 2.781 *SD* = 1.283	*M* = 4.008 *SD* = 2.190	1.226	0.000*	1.064
Task has no right answers	*M* = 2.641 *SD* = 1.269	*M* = 2.349 *SD* = 1.284	0.293	0.004	0.251
Task requires compassion	*M* = 3.424 *SD* = 1.761	*M* = 3.058 *SD* = 1.363	0.365	0.002*	0.334
Task relevant to all humans	*M* = 3.831 *SD* = 2.061	*M* = 3.265 *SD* = 1.446	0.567	0.000*	0.301
Task relevant to me personally	*M* = 2.961 *SD* = 1.060	*M* = 2.954 *SD* = 1.155	0.007	0.960	0.005

* Signifies Levine's test for equality of variances violated, equal variances not assumed test statistics used.
** Cohen's d effect size (maximum likelihood estimator).

Hypotheses

Hypothesis 1 stated that automated advice would be perceived as (1a) less expert, (1b) less useful and (1c) less appropriate than human advice in the humanitarian than the management task. We conducted two-way ANOVAs for perceptions of advisor expertise, appropriateness of advice, and usefulness of advice in the first block (pre-replacement) with task type and advisor type as independent variables. There was no significant interaction between task type and advisor type for expertise of advisor $F(1,669) = 0.062$, $p = 0.803$, $d = 0.062$, appropriateness of advice, $F(1,658) = 0.025$, $p = 0.873$, $d = 0.058$, or usefulness of advice, $F(1,658) = 0.003$, $p = 0.998$, $d = 0.001$. Follow up univariate tests indicated a main effect of advisor only: regardless of task,

human advice was always perceived as being more expert $F(1,669) = 20.681$, $p < 0.001$, $d = 0.356$, more appropriate $F(1,658) = 53.803$, $p < 0.001$, $d = 0.570$, and more useful $F(1,658) = 62.350$, $p < 0.001$, $d = 0.616$, in the first block of trials. Because the main effect of advisor type was significant in the survey results from the first block of trials, we could not treat every participant as coming from the same baseline condition when completing the second survey. Thus, we created a difference score by subtracting scores from the first survey from the second– a positive score on a survey index thus indicates the participant rated the advisor/advice higher after the second block of trials compared to the first. We then conducted a 2 x 2 x 2 ANOVA with task type, advisor type, and replacement type (similar or different) as factors. Results indicated no significant interaction between the three factors and ratings of advice appropriateness, $F(1,659) = 1.297$, $p = 0.255$, $d = 0.086$; but there was a significant interaction for ratings of advice usefulness, $F(1,662) = 5.829$, $p = 0.016$, $d = 0.189$, and advisor expertise $F(1,662) = 3.940$, $p = 0.048$, $d = 0.155$. To understand these interactions, we conducted two-way ANOVAs comparing the effect of replacement and advisor type on advisor expertise and ratings of advice usefulness within each task. In the humanitarian task there was a significant interaction between replacement and advisor type for advisor expertise, $F(1,365) = 13.156$, $p < 0.001$, $d = 0.288$, but this interaction was not significant ($F(1,305) = 0.058$, $p = 0.810$, $d = 0.005$) in the management task condition. The analyses were similar for the interaction between advisor and replacement type on advice usefulness in the humanitarian condition, $F(1,362) = 8.205$, $p = 0.004$, $d = 0.0224$, but this interaction was not present ($F(1,300) = 0.605$, $p = 0.437$, $d = 0.051$) in the management task condition. An inspection of the means makes these interactions clear: automated advisors replacing human advisors in the humanitarian task produced the largest decrease in evaluations of advisor expertise ($M = -0.439$, $SD = 0.714$) and advice usefulness (*M*

= -0.317, *SD* = 0.627), whilst human advisors replacing automated advisors produced the largest increase in ratings of expertise (*M* = 0.038, *SD* = 0.818) and advice usefulness (*M* = 0.063, *SD* = 0.766). Thus, we find partial support for hypothesis 1a and 1b, human advice is perceived as more expert and more useful than automated advice in humanitarian decision making tasks, but only when one advisor type has replaced the other. A table detailing the three factor ANOVAs for H1a-c, H4a-b, and our RQs can be found in Table 2; graphs for measures with a significant three-way interaction are displayed in figure 2.2-5 (Appendix II).

Hypothesis 2 stated that in the less demonstrable task (humanitarian), automated advice would be used less, in general, than human advice. Because all participants started the first 10 trials with a human or automated advisor, we first conducted a 2 (humanitarian/management task) x 2 (automated/human advisor) ANOVA with average advice utilization as the dependent variable. There was no significant interaction between advisor and task type $F(1,671) = 3.136$, $p = 0.077$, $\eta p^2 = 0.005$. There was a main effect of task ($F(1,671) = 8.775$, $p = 0.002$, $\eta p^2 = 0.013$) that indicated advice was used significantly more in the humanitarian (*M* = 0.561, *SD* = 0.191) than management task (*M* = 0.521, *SD* = 0.188), see figure 1. Follow up one-way ANOVAs within each task with advisor type (human/automation) as the independent variable found no significant difference in advice utilization for each advisor in the humanitarian task condition $F(1,365) = 0.204$, $p = 0.652$, $\eta p^2 = 0.001$, but in the management task, human advisors (*M* = 0.543, *SD* = 0.182) were used significantly more than automated advisors (*M* = 0.500, *SD* = 0.197), $F(1,306) = 3.833$, $p = 0.051$, $\eta p^2 = 0.012$. When we analyzed each advisor type individually with task type as the independent variable, results indicated that automated advice was utilized significantly more in the humanitarian task condition (*M* = 0.558, *SD* = 0.188) than in the management task condition (*M* = 0.516, *SD* = 0.205), $F(1,528) = 6.023$, $p = 0.014$, $\eta p^2 =$

0.013. We do not find the same effect for human advice utilization, which is not significantly different between the humanitarian (M = 0.549, SD = 0.197) and management task condition (M = 0.523, SD = 0.183), $F(1,520) = 2.396$, $p = 0.122$, $\eta p^2 = 0.002$, see fig. 2.1 (Appendix II) for a graph of important utilization results.

To properly analyze the second block of trials, we had to account for similar or different advisor replacement as well as advisor type and task. We computed a variable composed of average advice utilization on the second block of trials and then conducted a 2(human/automated advisor) x 2(similar/different advisor replacement) x 2(humanitarian/management task) ANOVA but found no significant interaction between task, advisor and replacement type, $F(1,667) = 0.656$, $p = 0.418$. To replicate our analysis on the first block of trials we conducted a follow up 2(humanitarian/management task) x 2(automated/human advisor) ANOVA but we did not observe a significant interaction, $F(1,667) = 0.320$, $p = 0.858$, or observe a main effect of task ($p = 0.133$) or advisor $p = 0.602$). As a final analysis, because we did have two conditions where participants received advice from automation or humans the entire time, we thought it could be informative to analyze utilization across all 20 trials for these participants only. Our analysis showed no significant interaction between advisor type and task type, $F(1,302) = 1.001$, $p = 0.318$, $\eta p^2 = 0.003$. In sum, hypothesis 2 is not supported, we find a significant result in regard to advice utilization between tasks and advisors across the first 10 trials, but it is in the opposite direction as stated by hypothesis 2.

Hypothesis 3a stated that perceptions of advisor thought process similarity and value similarity would have a stronger effect on advice utilization in the humanitarian task compared to the management task. A multiple regression model was tested to investigate whether the association between thought process similarity and value similarity depends on the task. After

centering advisor similarity index scores and computing similarity-by-task interaction terms (Aiken & West, 1991), the two predictors and the interaction were entered into a simultaneous regression model. Because each participant rated an advisor twice, we conducted separate regressions for each block of trials and subsequent questionnaire. In the first block of trials, results indicated no significant interaction between thought process similarity and task for utilization of advice (b = -0.001, $t(662)$ = 0.013, β = -0.005, p = 0.933). This can be interpreted as meaning there was no significant difference in the association between thought process similarity and utilization between the two task conditions. We did not discover an interaction between thought process similarity and task in the second block of trials either (b = -0.005, $t(660)$ = 1.011, β = -0.19, p = 0.733). We used a similar regression test to investigate whether the association between perceptions of advisor value similarity and advice utilization differs between tasks. Our results indicated no significant interaction between advisor value similarity and task in either the first (b = 0.001, $t(660)$ = 0.290, β = 0.004, p = 0.939) or second(b = 0.001, $t(656)$ = 0.233, β = 0.003, p = 0.962) block of trials. Additionally, neither of the models we tested revealed a main effect for perceptions of advisor thought or value similarity across both tasks; hypothesis 3a is unsupported

Hypothesis 3b stated that perceptions of advisor expertise and reliability would have a stronger effect on advice utilization in the management task compared to the humanitarian task. We used similar regression tests as used for hypothesis 3a. In the first block of trials, results indicated that higher evaluation of advisor expertise (b = 0.044, $t(650)$ = 4.063, β = 0.189, p = .001) was associated with greater advice utilization. The interaction between expertise and task was significant (b = -0.041, $t(650)$ = 2.334, β = -0.126, p = 0.025, f^2 = 0.101), suggesting that the effect of expertise on utilization was different in each task. Simple slopes for the association

between expertise and utilization were calculated within each task. Simple slope tests revealed a significant association between expertise and utilization in the humanitarian task (b = 0.044, $t(346)$ = 2.334, β = -0.177, p = 0.001, f^2 = 0.250), but not in the management task (b = 0.004, $t(304)$ = 0.801, β = 0.16, p = 0.781). We did not discover a similar interaction between expertise and task in the second block of trials (b = 0.005, $t(652)$ = 0.711, β = 0.13, p = 0.807), but there was a main effect for expertise, b = 0.026, $t(671)$ = 2.653, β = 0.102, p = 0.008, f^2 = 0.111, suggesting that expertise was significantly associated with utilization to an equal degree in both tasks. We used a similar regression test to investigate the association between perceptions of advisor reliability and advice utilization. Our results indicated no significant interaction between advisor reliability and utilization in either the first (b = -0.024, $t(651)$ = 1.554, β = -0.081, p = 0.148) or second (b = 0.020, $t(651)$ = 1.334, β = 0.61, p = 0.277) block of trials. Overall, hypothesis 3b is unsupported, we do find a significant relationship between perceptions of advisor expertise and advice utilization but only in the humanitarian condition, the opposite task condition as posited by hypothesis 3b.

Hypothesis 4a stated that perceptions of human/automated advisor thought process and value similarity would be greater/less in the humanitarian compared to the management task. For brevity, we used similar analyses to hypothesis 1, but with perceptions of value and thought process similarity. In the first block of trials there was no significant interaction between advisor and task type for perceptions of thought process similarity, $F(1,662)$ = 0.109, p = 0.741, or value similarity, $F(1,656)$ = 0.256, p = 0.613. Tests on the second block of trials revealed a significant three-way interaction of task, advisor, and replacement type on perceptions of thought process similarity $F(1,645)$ = 7.067, p = 0.008, d = 0.208; but not on perceptions of value similarity $F(1,634)$ = 1.236, p = 0.267. Follow up two-way ANOVAs indicated a significant interaction

between advisor and replacement type in the humanitarian task condition $F(1,352) = 38.038$, $p < 0.001$ $d = 0.672$, but in the management task condition this interaction was only marginally significant, $F(1,297) = 3.556$, $p = 0.060$ $d = 0.217$. Similar to ratings of advice usefulness (H1b), a human advisor replacing an automated advisor resulted in an increase in perceptions of thought process similarity ($M = 0.464$, $SD = 1.282$), whereas an automated advisor replacing a human resulted in the largest decrease ($M = -1.156$, $SD = 1.298$). In sum, we find partial support for hypothesis 4a, decision-makers do perceive more thought process similarity with human advisors compared to automated advisors in humanitarian decision-making scenarios, but only when one advisor type has replaced the other.

Hypothesis 4b stated that automated advisors will be perceived as more reliable in more demonstrable decision tasks than in less demonstrable decision tasks. We used a similar analysis to hypothesis 4a. For the first block of trials there was no significant interaction between advisor and task type for perceptions of advisor reliability, $F(1,664) = 0.799$ $p = 0.372$. Additionally, there was no significant interaction in the second block of trials between task, advisor, and replacement type and perceptions of advisor reliability, $F(1,652) = 2.842$ $p = 0.092$, $d = 0.132$. Hypothesis 4b is unsupported.

Table 2: Difference between rating of 1st advisor and second advisor, three-way ANOVAs and follow up.

Measure	Interaction	F[1]	p	ηp^2
Expertise (H1a)[2]	**Task*Replacement*2nd Advisor**	**3.954**	**0.047**	**0.006**
- In Humanitarian Task	**Replacement*2nd Advisor**	**13.966**	**0.001**	**0.037**
- In Management Task	Replacement*2nd Advisor	0.100	0.752	0.001
Usefulness (H1b)	**Task*Replacement*2nd Advisor**	**5.494**	**0.019**	**0.008**
- In Humanitarian Task	**Replacement*2nd Advisor**	**7.786**	**0.006**	**0.021**
- In Management Task	Replacement*2nd Advisor	0.565	0.453	0.002
Appropriateness (H1c)	Task*Replacement*2nd Advisor	1.381	0.24	0.002

Thought Process Similarity (H4a)	**Task*Replacement*2nd Advisor**	**6.817**	**0.009**	**0.011**
- In Humanitarian Task	**Replacement*2nd Advisor**	**38.369**	**0.001**	**0.099**
- In Management Task	**Replacement*2nd Advisor**	**3.931**	**0.048**	**0.013**
Value Similarity (H4a)	Task*Replacement*2nd Advisor	1.202	0.273	0.002
Reliability (H4b)	Task*Replacement*2nd Advisor	2.842	0.092	0.004
Positive Emotions (RQ1)	Task*Replacement*2nd Advisor	2.528	0.112	0.004
Negative Emotions (RQ1)	**Task*Replacement*2nd Advisor**	**3.888**	**0.049**	**0.006**
- In Humanitarian Task	**Replacement*2nd Advisor**	**4.114**	**0.043**	**0.011**
- In Management Task	Replacement*2nd Advisor	0.673	0.413	0.002
Anxiety (RQ2)	Task*Replacement*2nd Advisor	0.360	0.549	0.001
Reciprocity (RQ3)	Task*Replacement*2nd Advisor	0.976	0.324	0.001
Faulting the Advisor (RQ4)	**Task*Replacement*2nd Advisor**	**9.659**	**0.002**	**0.014**
- In Humanitarian Task	**Replacement*2nd Advisor**	**12.434**	**0.001**	**0.034**
- In Management Task	Replacement*2nd Advisor	0.814	0.368	0.003

[1] Error *df* for three-way tests ranged between 650-667, for two-way tests 306-358.
[2] Significant three-way interaction graphs in figure 2.2-5 (Appendix II)

Hypothesis 5a – 5d suggested that the effects posited by hypotheses above (with the exception of 3a and 3b) would be affected by advisor replacement, making the observed effects stronger. Our results above effectively answered these questions. We find partial support for hypothesis 5a, 5b, 5c & 5d (see results for H1a, H4a) that referred to perceptions advisor expertise, advice usefulness, advice appropriateness, and perceived advisor thought process similarity. Automated advisors that replace human advisors are rated as having less expertise, less useful advice, and having less similar thought processes to the decision-maker when they (automated advisors) replace human advisors compared to when they replace other automated advisors in the humanitarian task, but this interaction between advisor type and replacement effect is not present in the management task. Similarly, human advisors that replace automation are rated as having more expertise, advice usefulness, and similar thought processes to the

decision-maker when they (human advisors) replace automation as opposed to replacing another human; again, this effect is present in humanitarian but not management tasks. We do not find that ratings of appropriateness are significantly affected by advisor replacement. Hypotheses 5e and 5f suggested there would be an effect of advisor replacement type on utilization as well, but there was no significant interaction between task, advisor, and replacement type on advice utilization (see results for H2), hypotheses 5e and 5f are unsupported.

With regards to RQ1 (emotional reactions), RQ2 (anxiety), RQ3 (feelings of reciprocity), and RQ4 (faulting the advisor for mistakes), our analyses above suggested that the most interesting investigation would be in the difference between the rating of both advisors. We summarize the results of both exploratory paired samples *t*-tests and tests of three-way interactions between task, advisor, and replacement type in Table 2. Although these tests have not been corrected for type I error rate, we find them sufficient to answer our research questions that both decision making task and advisor replacement type affect the emotions of decision-makers and perceptions of the advisor. Of particular note are the two items that also produced significant three-way interactions between task, advisor, and replacement type: feelings of negative emotion ($F(1,656) = 3.792$, $p = 0.052$), and feelings faulting an advisor for mistakes, ($F(1,634) = 3.843$, $p = 0.050$). We conducted follow up two-way test within each task (advisor x replacement type). In the humanitarian task, there was a significant interaction between advisor and replacement for both negative emotions ($F(1,345) = 4.114$, $p = 0.044$), and faulting the advisor ($F(1,349) = 12.434$, $p < 0.001$). The interactions were not significant in the management condition (negative emotions, $p = 0.413$; fault $p = 0.368$). Figures 2a & 2b help in interpreting the interactions, in the humanitarian task, a human replacing a similar human advisor produced the largest increase in negative emotions but a human replacing an automated advisor resulted in the only decrease. For

faulting advisors, all conditions resulted in decreased fault placed on the second advisor, but humans replacing automation resulted in the largest decrease.

Table 2: Summary of paired samples t-tests for RQ1-4. M_{Diff} is mean difference of second advisor rating – first advisor rating (negative value = more endorsement of survey item for second advisor compared to first advisor).

Item	1	Humanitarian Task				Management Task			
		M_{Diff}	*t*	*Df*	*p*	M_{Diff}	*t*	*df*	*p*
Positive Emotion	ArA	-0.09	1.05	76	0.30	**-0.09**	**2.78**	**74**	**0.01**
	ArH	0.09	1.18	102	0.24	-0.03	0.32	80	0.75
	HrA	**-0.34**	**5.13**	**107**	**0.00**	-0.16	1.56	71	0.12
	HrH	-0.15	1.41	71	0.16	-0.03	0.30	73	0.76
Negative Emotion*	ArA	0.01	0.12	78	0.90	**0.14**	**2.67**	**74**	**0.01**
	ArH	-0.05	0.78	105	0.44	-0.07	0.81	82	0.42
	HrA	**0.10**	**1.91**	**105**	**0.06**	0.02	0.27	70	0.79
	HrH	0.14	1.31	70	0.19	-0.09	1.08	74	0.29
Feelings of Anxiety	ArA	0.12	1.63	78	0.11	0.06	0.75	74	0.46
	ArH	0.01	0.13	104	0.90	-0.02	0.28	81	0.78
	HrA	**0.13**	**2.45**	**103**	**0.02**	-0.01	0.10	70	0.92
	HrH	0.03	0.39	69	0.70	-0.06	0.87	74	0.39
Feelings of Reciprocity	ArA	**0.26**	**2.03**	**76**	**0.05**	0.08	0.70	74	0.49
	ArH	**0.74**	**5.02**	**102**	**0.00**	**0.89**	**5.08**	**80**	**0.00**
	HrA	**-0.96**	**7.62**	**105**	**0.00**	**-0.86**	**4.22**	**71**	**0.00**
	HrH	-0.12	0.95	71	0.35	0.20	1.36	75	0.18
I could fault my advisor for mistakes*	ArA	0.19	1.43	78	0.16	0.13	1.12	76	0.27
	ArH	-0.13	1.05	105	0.30	-0.08	0.58	82	0.57
	HrA	**0.70**	**4.82**	**106**	**0.00**	**0.35**	**2.16**	**72**	**0.03**
	HrH	**0.40**	**2.45**	**71**	**0.02**	-0.07	0.48	75	0.63

1: Advisor/Replacement Type: ArA = Automated Advisor replaced by Automated Advisor (similar replacement condition). ArH = Automated replaced by Human Advisor (different replacement condition), etc.
*Advisor Type x Replacement Type x Task Type three-way interaction is significant at the 0.05 level.

Discussion

Technological innovation is leading to the increased prevalence of automated advice in personal and professional life for decisions of varying demonstrability. Additionally, automated advisors are increasingly replacing human advisors. Our results shed light on the effects of task type and advisor replacement as it relates to human versus automated advisors. To summarize,

our findings show effects relating to the perception of an advisor as well as actual advice utilization, although both sets of effects are not always related. We found support for our first hypotheses which suggested human advice would be perceived as more expert and useful than automated advice for a humanitarian relief planning decision than a management decision. However, this effect was only significant when an automated advisor had replaced a human advisor and vice-versa. Our fourth hypothesis suggested that advisors would also differ across tasks on decision maker's perceptions of advisor similarity (humans perceived as more similar in humanitarian tasks) and reliability (automation perceived as more reliable in management tasks). We found that human advisors were perceived as having more thought process similarity (to the decision-maker) but, again, only when the human advisor had replaced an automated advisor and vice-versa. We did not find significant effects of task demonstrability or advisor replacement on perceptions of advice appropriateness, perceptions of advisor value similarity, or reliability. Our second hypothesis hypothesized that human advice would be used more on humanitarian than management tasks and the opposite pattern would occur with automated advice. However, the only significant result we found suggested that, if anything, the pattern was the opposite with human advice utilized significantly more in the first half of the management task but not the humanitarian task. There were no significant effect once the decision-makers had their advisors replaced. With so many potential repeated measures, individual trials, and manipulations (task, advisor, replacement), we certainly could have analyzed the data in so many ways that we might find something, but the truth is that with nonsignificant omnibus tests and main effects, doing so would be an exercise in making something out of nothing. Overall, there are not large effects of task and advisor type on utilization.

The plethora of significant effects found regarding advisor and advice perception, but not utilization, foretold our results to our third hypothesis which suggested different links between perception and utilization in each task. Results showed that for the most part, neither ratings of advisor expertise, reliability, or similarity were better predictors of advice utilization in either task. We did find that perceptions of advisor expertise were more predictive of utilization in the humanitarian than management task, but it was the opposite of what was hypothesized. Our research questions were intended to investigate decision-maker emotions and attributions, and our results show some interesting patterns, especially regarding the experience of negative emotions when one advisor type replaces a different advisor type. Our study was the first experimental study to start investigating the behavior and psychology of people who must work with automation that has replaced a fellow human. As a whole, our results foretell this future as being an uneasy and unhappy place for the humans who remain, even if that psychological state is not detectable in their behavior.

Perception and Behavior

Perceptions, beliefs, and attitudes are not always predictive of behavior. This principle is central to many of the most well researched theories in fields such as persuasion and trust (Fishbein & Ajzen, 1975; Hovland, Janis, & Kelley, 1953; Mayer et al., 1995). There are a number of potential moderators that may explain why the manipulation of advisor type, replacement, and task type affected perceptions of advisors more than actual utilization behavior. One explanation regards the difficulty and unfamiliarity of the task. Interpersonal advice research finds that decision-makers seek and utilize advice more when they perceive tasks to be more difficult (Gino & Moore, 2007). This effect is even more established in automated advice research where task difficulty and decision-maker workload are key concerns due to the high

workload in real-world environments in which automation is often introduced (Parasuraman et al., 2000). Reliance on automation tend to increase when decision-maker workload increases presumably because the decision-maker does not have time to either monitor the automation, or in a decision-making scenario, think through decision alternatives (Hoff & Bashir, 2015). Although perceived task difficultly and cognitive load are different, they point to the same mechanism driving advice utilization, which is that the decision-maker feels they do not have the ability to do make a better decision than the advisor. If decision-makers in our experiment perceived the actual act of forecasting as something they were not able to do well, it may have driven the utilization of advice regardless of perception of the advisor or advice. In other words, although one advisor was perceived more positively, participants may have still perceived either advisor as more informed than themselves. An interesting manipulation for future study is to select easier tasks or tasks on which decision-makers perceive themselves to have expertise. People who believe they are experts are more prone to overconfidence and advice-discounting in general (Sniezek & Van Swol, 2001; Yaniv & Kleinberger, 2000), and thus, there would be a higher bar towards advice utilization. There is also evidence that in professional contexts, the use of automation may pose some ego threat to decision-makers who feel if the task can be done by a computer, it implies their judgment and their livelihood is not unique and valuable (Goodwin et al., 2013; Naweed et al., 2013). In sum, our forecast tasks were designed to minimize decision-maker self-efficacy as much as possible given that they were only simple graphs and a highly unfamiliar context, but decision makers with moderate self-efficacy and especially high expertise would likely react to advisor replacement differently. Such a manipulation in future research may be able to find a more direct connection between utilization and evaluation of automated advice.

Another potential explanation relating to expertise is the possibility that differing ratings of expertise between advisors in task conditions was due to decision-makers feeling that human advisors had more expertise than automation in understanding the consequences of a decision (human lives), but not in actually comprehending the forecasting data and producing an optimal forecast. Our survey did not discern between advisor expertise referring to the ability to make a forecast as opposed to ability to understand the larger decision context such as decision consequences. This is an interesting direction for future study in research comparing human versus automated advice because it may uncover further underlying assumptions that decision-makers have about automated and human advisors that are specific to different parts of a decision process. Decision-making researchers have generated a wide variety of models to describe the human decision making process, but essentially all make distinctions between stages in the decision process such as information acquisition, information processing, and judgment and decision selection (Einhorn & Hogarth, 1981; Parasuraman et al., 2000). In our experiment, the graph stimuli remained exactly the same, but the numbers clearly meant something different (lives versus dollars). Thus, evaluation of the numbers themselves may involve a different cognitive process (i.e., information processing) than evaluating their meaning, which is a more judgmental process. Utilization of the advice may have reflected a decision maker's assessment of advisor's ability to perform one aspect of the decision-process, but perceptions of the advisor's expertise and usefulness may reflect an assessment of a different process such a judgement. Expectations of can automation play a critical role in predicting detrimental behaviors such as over or undereliance on automation (Dzindolet et al., 2003; Lyons & Stokes, 2012), and a better understanding of what aspects of the decision-making process these expectations refer to may lead to better automation design and integration.

A deeper understanding of the psychology of decision-makers in both tasks may also be gained by considering the strong effects of life/death decision framing found in a large amount of framing effects research (Tversky & Kahneman, 1981). Kühberger's (1998) meta-analysis found that the Asian Disease problem (and related variants) consistently produced the largest framing effects. In the original Asian Disease problem, decision-makers are asked to select alternatives based on the potential loss (or saving) of lives. Although our experiment here lacks a manipulation of gain and loss frames, there may be a greater sense of risk present simply because there are lives on the line, and human life is a more subjective reference point than something like money that is lost in framing scenarios that involve gambling (Kahneman & Tversky, 1979; Kühberger, 1998). Such reasoning also highlights the role of demonstrability because it may be that more value-based judgments are inherently seen as riskier. Risk itself is a key concept in advice and other related literature such as trust (Mayer et al., 1995) because it must be present in order for a decision-maker to truly have a choice of using the advice or not. Thus, future studies looking at both human and automated advice should consider task manipulations that are not only demonstrability based but represent a qualitatively different risk perception. The consideration of framing effects research also points to the benefits of future advice studies that can use designs and manipulations common in framing research such as gain/loss frames, and single versus multiple risky decisions.

Emotions

Another potential moderator between perception and behavior is decision-maker emotions. We examined both positive/negative emotion and anxiety due to past research showing emotion's effect on decision-maker perceptions and utilization (de Hooge et al., 2014; Gino et al., 2012). Our results showed a significant effect of task on negative emotions, but perhaps the

more interesting result is that for positive emotions there was only one condition that produced and increase in positive emotions: when a human advisor replaced an automated advisor on humanitarian tasks. In the same task, automation replacing a human produced the largest decrease in positive emotions (see table 2 for mean differences). This condition (automation replacing a human) is also notable because it is the only condition that produced a significant increase in decision-maker anxiety between the first and second advisor (within task and advisor paired *t*-test). These results should be interpreted with caution because they did not result in significant omnibus effects, but there is a clear direction implied – decision-makers are not feeling positive emotions when automation replaces a human advisor when making a less demonstrable decision. In our study and interpersonal advice research in general, it is unclear how emotions are tied to utilization behavior. For example, Gino et al. (2008) found that inducing anxiety led to more advice utilization, but (de Hooge et al., 2014) found that negative emotions resulted in lower perceived expertise of an advisor and lower utilization. Our results show that emotion is an important area for future research, especially because the emotional reaction to receiving advice seems to differ between human and automated advisors.

It is also informative to consider research conducted on overall perception of automation versus humans, (Jian et al., 2000) attempt to develop a trust in automation scale suggested that there may be subtle differences in the way that humans perceive trust and distrust in other humans versus in automation. Imagining a human being as untrustworthy may be a more unpleasant experience than assessing automation as untrustworthy. This would be in line with the general tendency of humans to want to trust one another (Levine, 2014) and believe that other humans have benevolent intentions towards them (Waal, 2009). However, we do not extend these same feelings towards automation. Thus, assessment of human advisors and advice may always

be biased such that humans are rated more positively. This explanation would be consistent with our results showing that human advisors were always rated as more expert, giving more useful advice, and giving more appropriate advice regardless of task.

Advisor Similarity and Liking

Perceived advisor similarity is another set of findings that provide insight into the complicated relationship between humans and automation. Unlike earlier research that has found value similarity is a more important predictor of utilization on less demonstrable tasks, we find that neither value nor thought process similarity was strongly associated with advice utilization in on our low demonstrability (humanitarian) task. However, advisor thought process similarity ratings for human advisors increased more when they replaced automated advisors. Although we do not know if decision-makers are consciously comparing one advisor to the other, a large amount of contrast effect research suggests this process happens unconsciously (Palmer & Gore, 2014). Perceived similarity does generally result in more liking (Strauss et al., 2010) and the implication of this is not only that humans may like human advisors that replace automation, but that humans do not like automation that replaces other humans. One potential manipulation in the future that could help uncover a liking effect would be to provide decision-makers with a choice of who to receive advice from. If decision-makers were given a choice of working with an automated or human advisor, it would be interesting to see which advisor participants would "like to" receive advice from. Together with a task demonstrability manipulation, such an experiment could reveal some affective processes behind the effects witnessed in our research. This is an important area for future research; there is conflicting survey evidence about how much the average people like the idea of automation replacing humans (Savela et al., 2017), and some field research suggests that automation is sometimes welcomed as a replacement to

humans (Wasen, 2010) or desired not to replace humans (Kristoffersson, Coradeschi, Loutfi, & Severinson-Eklundh, 2011; Naweed et al., 2013). However, research has not examined how people react when the human advisor they have is actually replaced with automation.

Human advisors may be liked more in general because our results show that decision-makers feel more reciprocity (i.e., "I owe something to my advisor") towards human advisors. This is quite a remarkable result if one considers that our manipulation of human and automated advisor was very minimal in this experiment. There was no interaction with either advisor nor was there any conversational wording added to the advice; it was simply delivered as a number. In conjunction with our results regarding advisor perception above, this result has important implications for human-automation trust and interpersonal advice theories. Interpersonal communication theories have been used many times to study human-automation interaction (Burgoon et al., 2016b), and there are some researchers that assert humans socialize automation to such an extent that human-automation communication can be considered a special case of interpersonal communication (Nass & Moon, 2000). However, our results suggest that even our small manipulation with no social interaction leads to very different assessments of a social feeling like reciprocity. Small manipulations in anthropomorphism consistently produce effects in other human-automation trust research (de Visser et al., 2016), but this research frequently lacks an actual human control condition. It is hard to predict if the differences seen in our experiment would be exaggerated or attenuated if we added more social features to the advice or had face to face interaction with the human advisor.

Research Implications

Our results have real-world implications. If a human is replaced by automation in the workplace, the interpersonal advice process clearly is a more social process than the human

advisor in this experiment. Yet, our experiment revealed this perception of a social process is substantially different for human versus automated advisors despite the advice being hardly social at all thus. In settings outside the lab, there are likely to be several differences between human and automated advice that would act as confounds if not controlled in a laboratory setting. Our experiment removed the confounds introduced by actual real-world social relationships between humans that exist in the workplace, and thus it is a very conservative comparison of humans and automation. When humans with real relationships are replaced by automation, perhaps these elements of social interaction are not "replaced," but actually "lost" instead. Humans are social creatures and the feeling that someone is helping you is a good one. There could be serious long-term consequences to the lost positive emotions that come from social interaction, everything from organizational commitment (Fisher, 2010) to productivity (Oswald, Proto, & Sgroi, 2015) is at risk when employees are not happy at work. Gaining a more thorough understanding of what is lost, socially and emotionally, when a human colleague is replaced by automation is a critical research area for scholars moving forward. Our results with a very conservation manipulation of human and automation highlight the importance of this future research agenda.

Additional real-world implications of our study are numerous. It is clear that humans do not like it when automation replaces a human advisor, even a human advisor who is zero-acquaintance and only imagined. Furthermore, our results suggest that decision-makers really do not like it when this replacement occurs on a task that is less demonstrable and may have more serious, moral consequences. But the negative feelings experienced when automation replaces a human do not necessarily mean that the automated advisor will be used less than the human advisor. If anything, our utilization results suggested automated advisors were used more in the

humanitarian task, the same task that produced the most negative evaluations of the automated advisor when it replaced a human. Understanding how the manipulation of advisor characteristics, situational context, decision-maker self-efficacy, and advice accuracy affect this complicated relationship is important if automation is to be effectively introduced into new contexts, like medicine, where machines are advisors on less demonstrable decisions. Our research shows this process will be complicated, and most importantly it shows that it is not only the humans who are replaced that will be unhappy; the people who must work with these new machines may not be happy either.

Chapter V: Summary

Automation replacing humans in a variety of job and societal roles is now an everyday topic of conversation in public, legal and political discourse. The prospect of humans being replaced has captured the imagination of the public and has a remarkable ability to bring out unbridled optimism in some, and pessimism from others. My research agenda in these three studies is largely shaped by a desire to start understanding what separates interpersonal relationships from human-automation relationships. I used the burgeoning field of interpersonal advice to guide the methodology and drew upon many concepts from the fields of trust, decision-making and social cognition. My results show that the relationship between humans and automation is complicated, but three overarching themes emerge from this work: (1) humans perceive automation differently than other humans, leading to differing expectations, (2) when different advisor types replace one another, contrast effects may lead to differing perceptions, and (3) even when humans use new automation that replaces a fellow human, it may be hiding the fact that in reality, they are not happy. My results have serious implications for future study and theoretical development in interpersonal advice and automated advice.

A recent review suggested that automation is quickly becoming accepted into job roles that are thought to highly involve on the emotions of care and benevolence (Savela et al., 2017). For example, robots are being welcomed as companions for elderly people, and surprisingly, the biggest advocates are often the elderly people themselves (Jen & Hung, 2010). One of the most laborious but delicate jobs in elderly care is bathing, and new robotic bathtubs have been the subject of several research studies (Beedholm, Frederiksen, Frederiksen, & Lomborg, 2015; Moon et al., 2012). Contrary to what some believe, a majority of people from clients to caregivers to average citizens view this replacement of human labor positively, despite the very

close interpersonal encounter that it replaces. There is evidence that automation also is found acceptable for roles in education and even in something thought to be artistic like dance performance (Wallis, Popat, McKinney, Bryden, & Hogg, 2010). However, many studies that investigate the mere perception of automation run into confounds such as the automation being seen as a form of entertainment (as in the dance study) or being perceived as only suitable for low level labor tasks. In the education study for example, automation was not perceived as ever being capable of teaching in the arts or social sciences (Savela et al., 2017). Additionally, rarely do studies study the actual possibility of humans being replaced by automation, but there is some field research showing that people do not desire for automation to replace human workers (Wasen, 2010).

The healthcare field has been at the forefront of automation adoption into the workplace. Everything from simple database technology to robotic surgeons and the artificially intelligent "Doctor Watson" have made their way into the current healthcare landscape (Doyle-Lindrud, 2015). Field studies from real hospitals and clinics reveal a myriad of factors that can lead to the acceptance or rejection of technology such as training, ease of use and prior experiences with technology (Chismar & Wiley-Patton, 2002). Fundamental to these moderators of technological adoption are the expectations that a person has about what the technology can do. Expectations in real life are obviously affected by an innumerable number of cues that a decision-maker perceives, but in all of the three experiments above, I carefully controlled cues that could affect the perception of the advisor, the strong results go to show that the differing expectations people have of automation versus humans, in absence of any other information, are strong and robust.

The results overall compliment prior research comparing humans and automation as well as contribute to theoretical development in interpersonal advice, human-automation trust, and

comparative theories. Below I will concisely outline the theoretical contributions of the three major themes coming out of this work. Throughout, I will identify key areas for future study.

Theoretical Contributions

Human-Automation Advice

The results from study 1 and 2 deepen the understanding of the differing expectations that a decision-maker may have about an automated advisor compared to a human advisor. The results are consistent with the predictions of the perfect automation schema model (Madhavan & Wiegmann, 2007b) and the framework for automation use developed by (Dzindolet, Beck, Pierce, & Dawe, 2001). These models first suggest that a decision-maker expects automation to generally perform better than a human. My results in studies 2 & 3 support this proposition, with the results from study 2 (within the management task only) suggesting that this effect can be stronger when automation replaces a human. The second effect predicted by the perfect automation schema is that automation will be utilized less (compared to a human) after the issuance of bad advice. The results in study 1 and 2 also support this proposition. In study two I see an interaction between advisor and replacement as well. When automated advisors replaced human advisors, the automation seemed to suffer even more from a decrease in advice utilization compared to if the automated advisor did not replace a human. On the flip side, humans that replaced automation suffered less of a decrease in advice utilization than humans that did not replace automation. Because of the unique focus within the management task (building off study 1), and the focus on theoretically important phases of advice utilization, study 2 in particular makes a major contribution to human-automation trust literature. Not only does study 2 replicate past research, but I also that the replacement of a human advisor with an automated advisor can exaggerate the effects.

Study 3 also makes contributions to human-automation trust theory. Prominent human-automation trust theories consistently highlight the importance of the task that the automation is to perform that will predict trust of the automation. Such propositions are in line with validated theories of technology acceptance as well (Hancock et al., 2011; Hoff & Bashir, 2015; Venkatesh et al., 2003). To my knowledge, study 3 is the first human-automation advice study to conceptualize variations in task on the intellective-judgmental continuum of demonstrability and the results show that such a manipulation can affect utilization behavior. Although it was not discussed at length in the study above due to the demands of academic publishing and space constraints, my results when looking at individual (critical) phases of advice utilization in the humanitarian task produced almost no significant results; this is clearly different than in the management task where I found significant results between advisor types. These non-significant results tell their own story because it they suggest that the task may attenuate the biases predicted by the perfect automation schema. Alternatively, if task matters, then perhaps there are task manipulations that will exaggerate the biases as well. For future studies, there are many different ways to manipulate tasks, but study 3 here shows that a manipulation conceptualized in terms of demonstrability has the potential to produce effects. Future study with task demonstrability manipulations will contribute greatly to theoretical development as well. Some of the most comprehensive human-automation trust models fail to conceptualize task related factors in a parsimonious way (e.g., Hoff & Bashir, 2015) and the classification of tasks along a set of consistent criteria such as task demonstrability would aid in making these theories more efficient and testable.

The results also contribute to current human-automation interaction theories that are larger in scope than simply use or acceptance. By comparing advisor types with clean

manipulations to control confounds, all three of my studies show that the differing expectations of humans compared to automation also leads to differing perceptions of those advisors. Automation researchers in the field of robotics are particularly interested in the impression formation process for social robots and if interpersonal impression formation theories are applicable. My results do not answer this question with a definitive yes or no, but I do show that impression formation for human and automated advisors may start from a different baseline condition – or in other words, decision-makers evaluate human and automated advisors on different underlying attributes they are believed to have. For example, I kept descriptions of the advisor expertise and experience constant in study 2 & 3, yet the actual perceived expertise of each advisor type was different, and I saw effects of task demonstrability on these perceptions of advisor expertise. A keen observer might note that perceptions of expertise were likely driven by advice accuracy – but I held that perfectly constant as well – the only explanatory mechanism I have is assumptions that decision-makers have about automation and humans drive differing perceptions. Unfortunately, these underlying assumptions are unmeasured in my experiments, but this does not mean they are unobserved. Using existing theory and past research we can use my results to infer what these underlying assumptions may be.

Perceptions of Automation

The utilization results, due to their replication of past effects, are likely a result of the underlying assumption that automation is high performing but invariant, whereas human performance is more variable (Dzindolet et al., 2003). Such underlying assumptions (in the absence of differentiating diagnostic information) would explain a greater use of automation until it no longer functions as expected. An unexpected error, due to automation's assumed invariance, is likely to be interpreted by decision-makers as a guarantee that the automation will

fail again in the future, leading to increased suspicion and less utilization. Humans on the other hand, can recognize their mistakes and improve. The findings regarding advisor perceptions have greater potential to uncover new underlying assumptions about automated advisors versus human advisors because they reflect evaluations of the advisor that are not evident in the utilization data. Additionally, task manipulations appeared to change these differences in perception.

First, I start with perceptions that were consistent across both tasks, most notably perceived advisor expertise which was essentially always perceived as being higher in humans than in automated advisors. There are a number of potential explanations. First, it is possible that humans have a natural tendency to want to evaluate other humans positively because giving a negative assessment of another person (even zero acquaintance) is unpleasant (Jian et al., 2000). But, giving a nonhuman a negative assessment may be less unpleasant and this naturally biases evaluations of humans to be better (Madhavan & Wiegmann, 2007b). Second, it may be that both of the tasks were ones that humans were perceived as being more expert at, and there are other tasks I could have chosen that would have resulted in higher ratings of automated advisor expertise, perhaps a task like signal detection that has found a bias towards automation in the past (Dzindolet et al., 2002). A third explanation – and one that contributes to current human-automation trust theory – is that expertise is not thought of as a very "automated" quality. It may be that expertise is something that humans use to evaluate other humans, but the concept of expertise has connotative meanings that are thought to be inapplicable to automation.

To explore this idea further I refer back to the well-accepted model of trust from Mayer et al., (1995) In this model, the "expertise" of an advisor would fall under the concept of ability, which is one of the three primary dimensions that trustees (advisors) are evaluated on. Of the many aspects of ability, a key concept is experience (Butler, 1991). Becoming experienced in a

task implies that (1), the advisor has performed the task before, but critically (2) the advisor has *learned* from their past experiences. This critical aspect of experience, and therefore expertise, may be something that automation is perceived as lacking. Such a proposition would be supported by the underlying assumptions assumed by the perfect automation schema, as the concept of invariance implies an inability to change. The concept of learning therefore, as an advisor attribute, has the potential to explain differing perceptions of advisors and differing utilization behaviors as well. I think this concept should be tested and if it does prove to inform a key underlying assumption that differs between humans and automation it should be incorporated into future theory. With the advance of artificial intelligence and "machine learning," the perception of learning as an advisor attribute is becoming a more critical concept to study every day.

Interpersonal Advice Theory

The research presented here also makes clear contributions to interpersonal advice theory. Some authors in comparing humans to automation have remarked that studying how humans perceive automation can teach us new aspects of the human-human perceptions process as well (Waytz, Heafner, & Epley, 2014), I find that to be the case especially as it relates to advice utilization. Advice Response Theory (MacGeorge et al., 2013) is one of the primary frameworks around which the interpersonal advice process has been studied in the last decade (e.g., Feng & MacGeorge, 2010; Feng & Magen, 2016), but this research has frequently only measured intentions to implement advice and not actual advice utilization. Sometimes, this is a needed design compromise in order to study actual problems and advice situations that are relevant to participants at the time of study. My results show that though perceptions of the advisor and advice quality may change, these changes may not be reflected in actual utilization behavior. To

be fair, no advice researchers have ever asserted that intentions could be assumed to lead to utilization, but my research shows how tenuous this link is even if it assumed there is some relationship.

A second contribution of the above research to interpersonal advice theory comes from the discovery of contrast effects in evaluating advisors. The comparison of one advisor to another is implied to occur but to my knowledge has not be tested in the framework of contrast effects. My results show that the sequence of advisors matter, especially as it relates to advisor perception. To summarize, future interpersonal advice theory development will benefit from the research presented here as it furthers the understanding of utilization intent versus behavior and contrast effects in advisor perception.

Expectancy Violation Theory

Expectancy Violations Theory (EVT) was originally developed to specifically describe nonverbal communication, but since its proposal, it has been used to study a wide range of communication phenomena such as verbal communication and media use (Bartholow, Fabiani, Gratton, & Bettencourt, 2001; Burgoon & Jones, 1976; Palmgren & Rayburn, 1982). Many of My hypotheses in study 2 were based on the application of EVT to the perfection automation schema. This application proved useful and made my results easier to understand. Before reviewing my contributions to EVT in studying automation versus humans, I will review some important aspects of EVT that should be tested in automated advice research.

EVT offers a number of core concepts that allow for a more detailed comparison of the expectations of automation versus humans. First of all, EVT proposes that violations have a valence (Burgoon, 2015). Essentially all human-automation trust literature considers expectancy violations to be negative. Perhaps this is natural because it is easy to observe cases when

automation malfunctions or does not perform to the high level expected. But, it is possible that automation could perform in a way that is better than the decision-maker expected. Examples would include automated advisors that provide more information that they are expected to, and the provision of this information is believed to increase decision quality. A second component of EVT that may be useful going forward is the proposition that expectancy violations produce differing levels of arousal, and the level of arousal predicts how heavily the violation will affect impression formation as well as future communication intentions. While interpersonal advice research has long acknowledged that negative advisor behaviors (in trust research: "trust violations") can have a larger impact on future advice-seeking behavior, acknowledgement of the degree to which a violation occurs is missing from some prominent human-automation trust theories (Hoff & Bashir, 2015). The prospect that some violations may be weighted more heavily in perception of advisors is particularly interesting when the manipulation of task demonstrability is brought into consideration. It is possible that certain violations (e.g., accuracy) are perceived as "worse" than other violations by decision-makers, and this may differ based on task demonstrability. For example, if the task is highly demonstrable, therefore causing high perceptions of advisor suitability and high-performance expectations for automated advisors, are violations of the performance expectation more arousing than if the task were low in demonstrability? EVT proposes that this would be the case, and future experiments to uncover the differing dynamics of violation arousal and violation valence would be informative for understanding the different expectations of automation and humans.

The research above also contributes to EVT, especially the recent development of EVT as a theory to predict impression formation with automated virtual agents. It is here that I see my curious results related to the faulting of advisors being especially informative. Being willing to

fault an advisor for mistakes has a number of potential meanings, one that fascinates me is that faulting an advisor involves a perception that they *should* have done something, but the advisor failed to do it. If a decision-maker believes that an advisor should do something, it follows that the decision-maker had certain expectations of the advisor in the first place. Thus, the contributions to expectancy violation theory are first in terms of measurement. Perceptions of fault may reveal different aspects of the expectations that a communicator had in an interaction and EVT researchers in the future should consider measuring this construct. My second contribution to EVT are the results regarding perceptions of similarity. Recent EVT research on virtual agents frequently uses references to virtual agents being anthropomorphized so that they are "similar" to humans. The general thought is that the more similar an agent is to a human, the better that existing EVT research in interpersonal interaction will predict tendencies with virtual agents, but this similarity is only conceptualized in terms of similar social cues. My results above show that the perception of similarity is not complete unless it is measured in terms of thought process and value similarity as well.

Contrast Effects

The research presented above has straightforward implications for the continued study of contrast effects in advice research. The most important finding regarding contrast effects is that they appear to be real and a factor in advice utilization, advisor perception, and decision-maker emotions. It cannot be overstated that a contrast effect was the most consistent finding throughout studies 2 and 3. This discovery has serious real-world implications for areas that automation is replacing humans in the workplace. Primarily, the results show that the large amount of research done on human-automation trust is probably not capturing the strength of the effects discovered unless the experiments test a condition where humans are replaced by

automation (provided that is what the automation is designed to do). In no way does this invalidate past research, but it clearly points out that experimental designs thus far have been missing this important consideration. Fortunately, some research on integrating robots into human work teams has begun to test this manipulation (Carpenter, 2013), and I urge all scholars studying human-automation trust to begin considering a human replaced by automation condition as a standard manipulation in experiments going forward.

Mortal Versus Machine Theory III

When study 1 was completed, I engaged in a long process of comparing human automation trust literature with interpersonal trust literature to develop a theoretical model that directly describes the differences between human and automation trust. The resulting theory, Mortal versus Machine (MvM version I & II) was an entertaining exercise in theory development but I basically wrote it off as too complex to test. However, as I reviewed my results above I felt that some of the predictions made in MvM II were useful for understanding my results. I think a valuable way to finish this research project is by briefly situating my results in MvM theory and adding a new construct that will be of use in its future development.

MvM III General Structure

What follows is an extremely brief overview of the constructs in MvM theory. The details of the conceptualizations can be found in (Prahl, 2015) which will be attached in a reduced form in the appendix. The general structure of MvM III follows from the structure of the Integrated Model of Organizational Trust (IMOT) (Mayer et al., 1995), and the concepts of ability, value congruence and benevolence largely follow from the concepts of ability, integrity and benevolence in IMOT. Put simply, advisors are evaluated on these three criteria. Ability refers to possessing the skills and competence needed to provide good advice. Value congruence is the

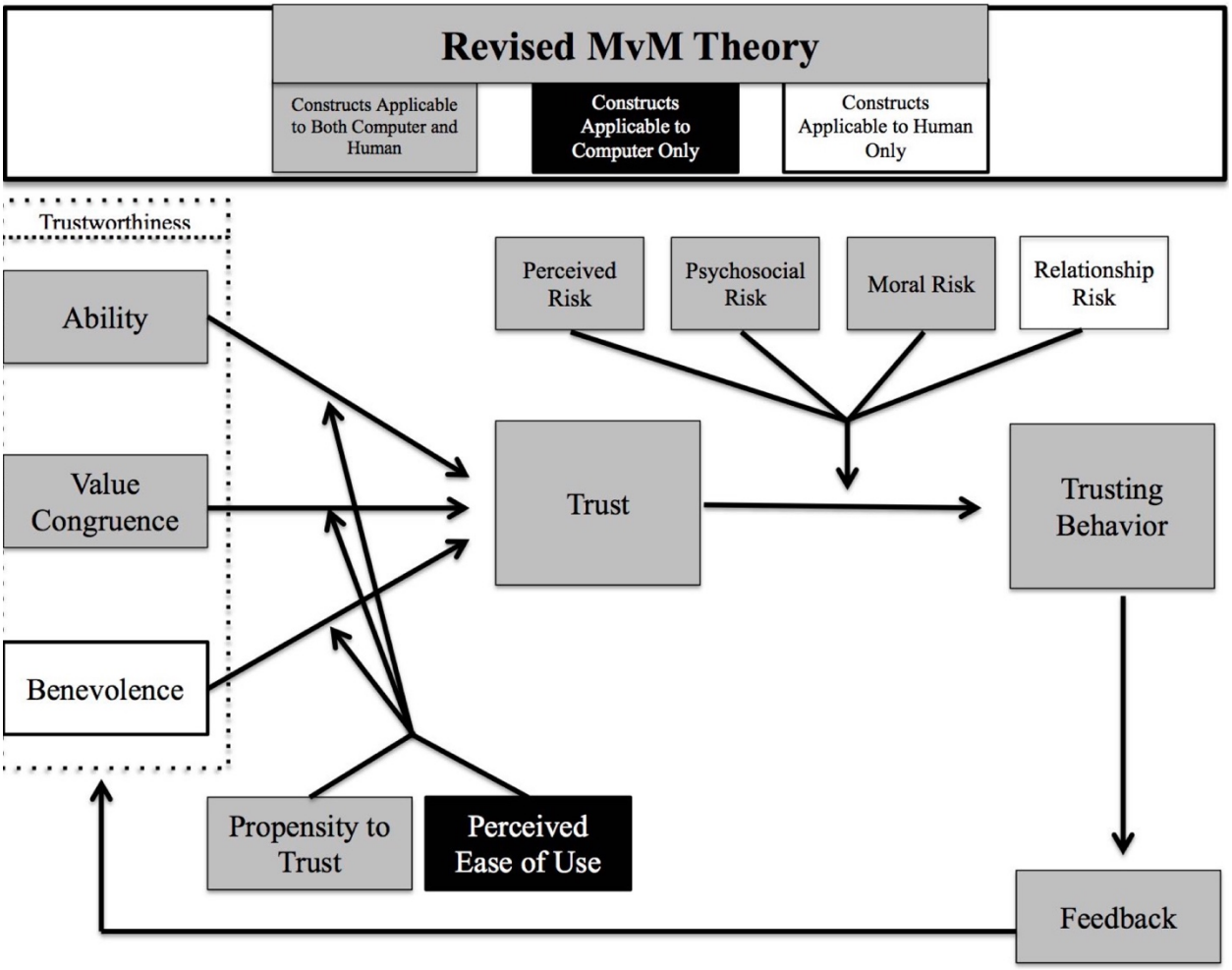
perception that the advisor shares the same value structure as the decision-maker (similar to value similarity as measured above). Finally, benevolence refers to the perception that the advisor has good intentions towards the decision-maker. These three dimensions form an overall perception of trustworthiness that leads to the mental state of trust (a willingness to take risk).

A number of potential moderators affect this relationship between perception and trust. First, for automated systems, they may be difficult to use. While the perceptions of them may not change because of this, the desire to use them may because they are seen as not worth the trouble. Additionally, consistent with the original IMOT model, individuals have an overall propensity to trust – this moderator therefore accounts for people with similar assessments of an advisor but differing trust levels. MvM also acknowledges that there are moderators between trust and trusting behavior. It is in this area that I see the research in study 2 & 3 becoming extremely relevant. Tasks differ in the amount of risk that they present to a decision-maker, and this risk can be of various kinds. In MvM there is perceived risk which is a general category for decision aspects such as financial risk. Accepting an advisor's advice, though, is somewhat an admission that one does not think they are good enough to make the decision alone. This may pose a threat to one's ego and this is conceived as psychosocial risk. Moral risk is also included to account for decisions that have ethical risks. Finally, relationship risk deals with the reciprocal nature of trust and advice. In particular, not using an advisor's advice may harm your relationship with them as they could have their feelings hurt by the apparent discounting of their advice.

The results from study 2 & 3 fit nicely into these conceptualizations of risk. First, the reciprocal nature of advice (and how it differs between human and automated advisors) is clear in study 3. Indeed, as stated in MvM, relationship risk seems to be far more present in

interpersonal advice than automated advice. The presence of differing emotions also hints at different psychological processes that may be occurring when accepting advice from either advisor type; such emotional risks could fit into the psychosocial risk construct. Perhaps most interesting is the idea of moral risk. It is likely that the low demonstrability manipulation (humanitarian decisions) were perceived as having moral consequences. However, as nice as these effects may fit with risk, MvM is left with a major gap: risk is proposed as a moderator between perception and behavior, yet many of the significant results showed an effect of task on perceptions of the advisor. MvM would benefit from the identification and addition of a measurable construct to understand how task affects advisor perception. Because what follow is not included in the original MvM manuscript, I go into more detail with the conceptualization of the advisor suitability construct.

MvM II Model (Prahl, 2015)



A New Element: Advisor Suitability to Task Demands

A core part of my research rationale is that automation is being used to inform decisions that are much less demonstrable than people typically may associate with automation. Advising a medical treatment, for example, has many more considerations (i.e., pain) than just mathematical probability of success, but automation is widely in use for such purposes (Doyle-Lindrud, 2015). Nearly every trust theory (both human and automated) and advice theory cited in this document consider evaluations of an advisor to be context specific. But this process of comparing the

perception of an advisor to the context/demands of a decision is left unexplained. This is understandable because it seems like common sense that perceptions of an advisor would certainly be dependent on the situation context. But such an abstract notion is hardly testable as an effect on other model constructs. I think that MvM (and all advice theories) may benefit from the actual comparison process between task context and advisor perception being better understood. Below, I will use a recent model in impression formation literature to create the construct of perceived advisor suitability. My goal is not only to explore the perceptions of suitability but also to turn it into a testable construct.

Warmth and competence have been proposed by a number of scholars as the two core tenants of social perception and they are central to the Stereotype Content Model, which proposes that impression formation of groups of people form along both dimensions (Fiske, C, Glick, & Xu, 2002). It may seem out of place to use concepts from a group theory to describe human versus automated advice, but it fits well with my experimental results because my manipulations do little to individualize each advisor; we can assume that decision-makers in the above experiments were making assumptions about the attributes of humans and automation generally. The most concise and straightforward understanding of the warmth dimension comes from (Leach, Ellemers, & Barreto, 2007) study that broke the warmth dimension into moral and social assessments. Social warmth was associated with the words, “likeable,” “friendly,” and “warm” and moral warmth with, “honest,” “sincere,” and “trustworthy.” Competence on the other hand was associated with, “intelligent,” “competent,” and “skilled.” When measurements of the warmth and competence dimensions are compared to trust models (like those I have used extensively) parallels emerge between the competence dimension and the ability dimension of trustworthiness (Mayer et al., 1995); and between warmth and the benevolence dimension of

trustworthiness. However, in trust theories, these assessments are highly context specific when compared to the Stereotype Content Model's use of the concepts to describe perception in a wider range of contexts.

Is it possible to conceive of decisions as being more suited to advisors that, as a group, are perceived differently in warmth and competence? Given that warmth appears to be described more in terms of emotion, and competence in terms of ability, these could be seen a core element of what makes an advisor suitable for an advisor role. To return to the conceptualization of decisions as demonstrable, a commonly cited example of the most demonstrable (intellective) tasks it math problems. Does making a decision about the answer to a math problem require warmth? It seems unlikely, emotion will not change the answer to a math problem. On the other hand, solving a math problem requires intelligence and competence. However, what if the decision is about the right balance between cost and safety when deciding what hours school crossing guards will work? Automation may be able to process the numbers and have the competence to produce a statistical forecast, but it lacks the emotion to understand the consequences of a suboptimal decision. With children's safety at risk, such a decision may feel like it requires warmth from a decision-maker or advisor. The analogy on the task demonstrability continuum is that the decision requires value judgments, making it a low demonstrable (judgmental) task.

Conceptual overlap between task and advisor perception

Concept from Stereotype Activation Model	
Warmth	Competence
Similar Concept from Trust and Advice Theories (e.g., Mayer et al., 1995)	
Benevolence, Integrity	Ability, Experience
Similar Concept from Task Demonstrability Model	

Judgmental (Low Demonstrability)	Intellective (High Demonstrability)

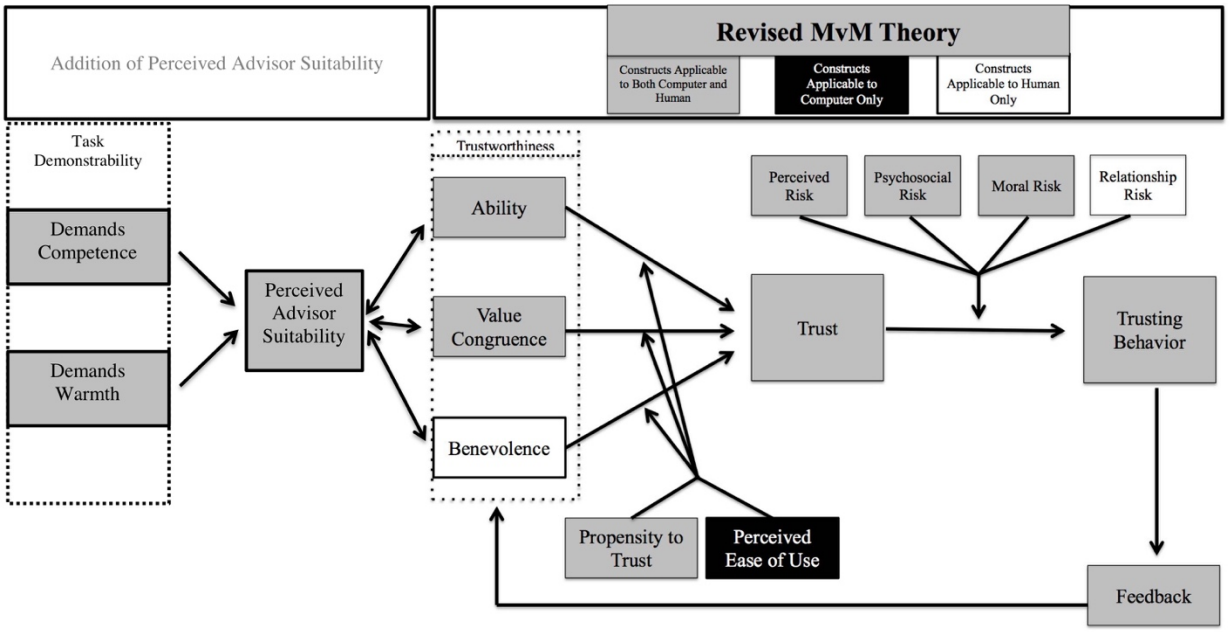
The central reason I have introduced the concepts of warmth and competence is because they may be able to describe the matching between perceptions of advisors and task demonstrability. Warmth and competence are conceptually similar to well established dimensions of trustworthiness and the demonstrability of a task clearly has differing demands for warmth and competence: high demonstrability tasks demand competence and as tasks become less demonstrable, they demand more warmth. Task demonstrability, however, was never conceptualized as a system by which humans form perceptions of other people. The concepts of warmth and competence – firmly established in impression formation literature – therefore allow for a conceptual bridge between task demands and how an advisor is perceived to meet those demands.

If the concepts of warmth and competence are to be useful for predicting perceived advisor suitability, I should first return to the results present in my research presented here. In my pilot study, I had one question that asked about the requirement of compassion for the forecasting task; there was a significant difference between perceptions of the tasks with the humanitarian task being perceived as requiring more compassion. This is a very clear indicator that the task was perceived as requiring more warmth than the management task. But automated advisors are likely not perceived as possessing as much warmth (if any) as human advisors. In fact, there is considerable efforts being made in human-automation interaction and robotics research to ascertain if automation is perceived as possessing emotion, or many of the other aspects of the warmth concept such as benevolence and honesty (Kahn et al., 2006; Savela et al., 2017). As of today, there is not a clear empirical answer to this question, but a growing number of scholars are

suggesting that the social responses to automation seen in early human-automation interaction research were only social and not indicative of perceiving the automation as "human" like some authors thought at the time (de Visser et al., 2016; Nass, Moon, Fogg, Reeves, & Dryer, 1995). The research above supports the view that automation is perceived as having less of (or not possessing) the warmth that was demanded by the humanitarian task. This may have led to the generally less positive evaluations of automation and automated advice in the humanitarian task compared to human advisors. Additionally, this may have led to effects on decision maker emotional reactions to receiving advice, which were less positive when receiving automated advice on the humanitarian task; it does not feel good to get advice from an advisor you feel is ill-well suited to the task. Structurally, I place perceived advisor suitability in between evaluations of the task and evaluations of the advisor. Both of these perceptions come together and are "matched" in the process of perceiving advisor suitability. To avoid having too many unidirectional arrows, I think it is appropriate to assume that the task itself is not affected by the perception of the advisor, but perceptions of the advisor are affected by perceived suitability.

To summarize, I propose the concept of perceived advisor suitability to describe the process in which a decision-maker combines their perception of the task demands and their perception of the advisor. This matching process can be thought of to occur on the two broad dimensions of warmth and competence as these concepts have clear analogs in task demonstrability and impression formation literature. The interpersonal trust and human-automation trust theories that were used to formulate many of the hypotheses in my study would benefit from a more detailed and testable construct to describe the comparison of perceptions of advisor attributes and task demands; the concept of perceived advisor suitability may accomplish this goal.

MvM III Model



Conclusion

The three research studies presented here tell a compelling story about human's complicated relationship with automation and how it differs from interpersonal relationships. Understanding how human communication changes when we communicate with automation is an important question that will only grow in importance with continued technological advance. The research here is especially important because it directly compares human advice to automated advice, studies the replacement of human advisors with automated advisors, and investigates manipulations of decision-making tasks. A large amount of human labor will undoubtedly be replaced by automation in the coming decades, my research suggests that those humans who remain will likely utilize their new automated companions, but that does not mean the automation is liked or that humans are happy about it.

Afifi, & Burgoon, J. K. (2006). The impact of violations on uncertainty and the consequences for attractiveness. *Human Communication Research*, *26*(2), 203–233. https://doi.org/10.1111/j.1468-2958.2000.tb00756.x

Agneessens, F., & Wittek, R. (2012). Where do intra-organizational advice relations come from? The role of informal status and social capital in social exchange. *Social Networks*, *34*(3), 333–345. https://doi.org/10.1016/j.socnet.2011.04.002

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Alexander, J. C. (1995). Refining the Degree of Earnings Surprise: A Comparison of Statistical and Analysts' Forecasts. *Financial Review*, *30*(3), 469–506. https://doi.org/10.1111/j.1540-6288.1995.tb00842.x

Alvarado-Valencia, J. A., & Barrero, L. H. (2014). Reliance, trust and heuristics in judgmental forecasting. *Computers in Human Behavior*, *36*, 102–113. https://doi.org/10.1016/j.chb.2014.03.047

Bachman, G. F., & Guerrero, L. K. (2006). Relational quality and communicative responses following hurtful events in dating relationships: An expectancy violations analysis. *Journal of Social and Personal Relationships*, *23*(6), 943–963. https://doi.org/10.1177/0265407506070476

Bahner, J. E., Hüper, A.-D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies*, *66*(9), 688–699. https://doi.org/10.1016/j.ijhcs.2008.06.001

Bandura, A. (1969). Social learning of moral judgments. *Journal of Personality and Social Psychology*, *11*(3), 275–279. https://doi.org/10.1037/h0026998

Barber, B. (1983). *The Logic and Limits of Trust*. New Brunswick, NJ: Rutgers University Press.

Bartholow, B. D., Fabiani, M., Gratton, G., & Bettencourt, B. A. (2001). A Psychophysiological Examination of Cognitive Processing of and Affective Responses to Social Expectancy Violations. *Psychological Science*, *12*(3), 197–204. https://doi.org/10.1111/1467-9280.00336

Baumol, W. J., Blinder, A. S., & Wolff, E. N. (2005). *Downsizing in America: Reality, Causes, and Consequences*. New York: Russell Sage Foundation.

Beedholm, Frederiksen, Frederiksen, & Lomborg. (2015). Attitudes to a robot bathtub in Danish elder care: A hermeneutic interview study. *Nursing & Health Sciences*, *17*(3), 280–286. https://doi.org/10.1111/nhs.12184

Behringer, R., Sundareswaran, S., Gregory, B., Elsley, R., Addison, B., Guthmiller, W., … Bevly, D. (2004). The DARPA grand challenge - development of an autonomous vehicle. In *IEEE Intelligent Vehicles Symposium, 2004* (pp. 226–231). https://doi.org/10.1109/IVS.2004.1336386

Benedikt Frey, C., & Osborne, M. (2013). *The Future of Employment: How Susceptible are Jobs to Computerization?* Oxford, Reino Unido: Oxford Martin School.

Berscheid, E., & Reis, H. T. (1998). Attraction and close relationships. In Daniel T. Gilbert & S. T. Fiske (Eds.), *The Handbook of Social Psychology, Vols. 1-2, 4th ed.* (pp. 193–281). New York, NY, US: McGraw-Hill.

Billings, D. R., Schaefer, K. E., Chen, J. Y., Kocsis, V., Barrera, M., Cook, J., … Hancock, P. A. (2012). *Human-animal trust as an analog for human-robot trust: A review of current evidence*. DTIC Document.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

Bond, C. F., Omar, A., Pitre, U., Lashley, B. R., Skaggs, L. M., & Kirk, C. T. (1992). Fishy-looking liars: deception judgment from expectancy violation. *Journal of Personality and Social Psychology*, *63*(6), 969–977.

Bonhard, P., Harries, C., McCarthy, J., & Sasse, M. A. (2006). Accounting for Taste: Using Profile Similarity to Improve Recommender Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1057–1066). New York, NY, USA: ACM. https://doi.org/10.1145/1124772.1124930

Brooks, A. W., Gino, F., & Schweitzer, M. E. (2015). Smart People Ask for (My) Advice: Seeking Advice Boosts Perceptions of Competence. *Management Science*, *61*(6), 1421–1435. https://doi.org/10.1287/mnsc.2014.2054

Burgoon, J. K. (2015). Expectancy Violations Theory. *The International Encyclopedia of Interpersonal Communication*. https://doi.org/10.1002/9781118540190.wbeic102

Burgoon, J. K., Bonito, J. A., Lowry, P. B., Humpherys, S. L., Moody, G. D., Gaskin, J. E., & Giboney, J. S. (2016a). Application of Expectancy Violations Theory to communication with and judgments about embodied agents during a decision-making task. *International*

Journal of Human-Computer Studies, *91*, 24–36. https://doi.org/10.1016/j.ijhcs.2016.02.002

Burgoon, J. K., Bonito, J. A., Lowry, P. B., Humpherys, S. L., Moody, G. D., Gaskin, J. E., & Giboney, J. S. (2016b). Application of Expectancy Violations Theory to communication with and judgments about embodied agents during a decision-making task. *International Journal of Human-Computer Studies*, *91*, 24–36. https://doi.org/10.1016/j.ijhcs.2016.02.002

Burgoon, J. K., & Hale, J. L. (1988). Nonverbal expectancy violations: Model elaboration and application to immediacy behaviors. *Communication Monographs*, *55*(1), 58–79. https://doi.org/10.1080/03637758809376158

Burgoon, J. K., & Jones, S. (1976). TOWARD A THEORY OF PERSONAL SPACE EXPECTATIONS AND THEIR VIOLATIONS. *Human Communication Research*, *2*(2), 131–146. https://doi.org/10.1111/j.1468-2958.1976.tb00706.x

Burleigh, T. J., Schoenherr, J. R., & Lacroix, G. L. (2013). Does the uncanny valley exist? An empirical test of the relationship between eeriness and the human likeness of digitally created faces. *Computers in Human Behavior*, *29*(3), 759–771. https://doi.org/10.1016/j.chb.2012.11.021

Butler, J. K. (1991). Toward Understanding and Measuring Conditions of Trust: Evolution of a Conditions of Trust Inventory. *Journal of Management*, *17*(3), 643–663. https://doi.org/10.1177/014920639101700307

Carlson, C. L. (2014). Seeking Self-Sufficiency: Why Emerging Adult College Students Receive and Implement Parental Advice. *Emerging Adulthood*, *2*(4), 257–269. https://doi.org/10.1177/2167696814551785

Carpenter, J. (2013). Just doesn't look right: exploring the impact of humanoid robot integration into explosive ordnance disposal teams. In *Handbook of Research on Technoself: Identity in a Technological Society* (pp. 609–636). IGI Global.

Cha, A. (2015, June 27). Watson's next feat? Taking on cancer. *Washington Post*. Retrieved from http://www.washingtonpost.com/sf/national/2015/06/27/watsons-next-feat-taking-on-cancer/

Chismar, W. G., & Wiley-Patton, S. (2002). Test of the technology acceptance model for the internet in pediatrics. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, 155–159.

Coeckelbergh, M. (2010). Moral appearances: emotions, robots, and human morality. *Ethics and Information Technology*, *12*(3), 235–241. https://doi.org/10.1007/s10676-010-9221-y

Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *Journal of Applied Psychology*, *92*(4), 909–927. https://doi.org/10.1037/0021-9010.92.4.909

Cortese, A. J. P. (1990). *Ethnic Ethics: The Restructuring of Moral Theory*. New York, NY: SUNY Press.

Costanzo, M., & Costanzo, S. (1992). Jury decision making in the capital penalty phase. *Law and Human Behavior*, *16*(2), 185–201. https://doi.org/10.1007/BF01044797

Crowley, P. M. (1968). Effect of training upon objectivity of moral judgment in grade-school children. *Journal of Personality and Social Psychology*, *8*(3, Pt.1), 228–232. https://doi.org/10.1037/h0025576

Cummings, M. L. (2006). Automation and Accountability in Decision Support System Interface Design. *Journal of Technology Studies*, *32*(1), 23–31.

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571–582. https://doi.org/10.1037/0003-066X.34.7.571

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

de Hooge, I. E., Verlegh, P. W. J., & Tzioti, S. C. (2014). Emotions in Advice Taking: The Roles of Agency and Valence. *Journal of Behavioral Decision Making*, *27*(3), 246–258. https://doi.org/10.1002/bdm.1801

de Visser, E. J., Monfort, S. S., McKendrick, R., B, A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331–349. https://doi.org/10.1037/xap0000092

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

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

Doyle-Lindrud, S. (2015). Watson Will See You Now: A Supercomputer to Help Clinicians Make Informed Treatment Decisions. *Clinical Journal of Oncology Nursing*, *19*(1), 31–32. https://doi.org/10.1188/15.CJON.31-32

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*(3), 177–190. https://doi.org/10.1016/S0921-8890(02)00374-3

Dzindolet, M. T., Beck, H. P., Pierce, L. G., & Dawe, L. A. (2001). *A framework of automation use*. Army Research Lab - Aberdeen Proving Ground - Maryland.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The Perceived Utility of Human and Automated Aids in a Visual Detection Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *44*(1), 79–94. https://doi.org/10.1518/0018720024494856

Einhorn, H. J., & Hogarth, and R. M. (1981). Behavioral Decision Theory: Processes of Judgement and Choice. *Annual Review of Psychology*, *32*(1), 53–88. https://doi.org/10.1146/annurev.ps.32.020181.000413

Epley, N., & Gilovich, T. (2006). The Anchoring-and-Adjustment Heuristic: Why the Adjustments Are Insufficient. *Psychological Science*, *17*(4), 311–318. https://doi.org/10.1111/j.1467-9280.2006.01704.x

Feng, B., & Burleson, B. R. (2008). The Effects of Argument Explicitness on Responses to Advice in Supportive Interactions. *Communication Research*, *35*(6), 849–874. https://doi.org/10.1177/0093650208324274

Feng, B., & MacGeorge, E. L. (2006). Predicting receptiveness to advice: Characteristics of the problem, the advice-giver, and the recipient. *Southern Communication Journal*, *71*(1), 67–85.

Feng, B., & MacGeorge, E. L. (2010). The Influences of Message and Source Factors on Advice Outcomes. *Communication Research*, *37*(4), 553–575. https://doi.org/10.1177/0093650210368258

Feng, B., & Magen, E. (2016). Relationship closeness predicts unsolicited advice giving in supportive interactions. *Journal of Social and Personal Relationships*, *33*(6), 751–767. https://doi.org/10.1177/0265407515592262

Fildes, R., Goodwin, P., & Lawrence, M. (2006). The design features of forecasting support systems and their effectiveness. *Decision Support Systems*, *42*(1), 351–361. https://doi.org/10.1016/j.dss.2005.01.003

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: an introduction to theory and research*. Retrieved from http://trid.trb.org/view.aspx?id=1150648

Fisher. (2010). Happiness at Work. *International Journal of Management Reviews*, *12*(4), 384–412. https://doi.org/10.1111/j.1468-2370.2009.00270.x

Fiske, S. T., C, J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902. https://doi.org/10.1037/0022-3514.82.6.878

Ford, M. (2015). *Rise of the Robots: Technology and the Threat of a Jobless Future*. New York, NY: Basic Books.

Freytas-Tamura, K. de. (2018, March 19). What's Next for Humanity: Automation, New Morality and a 'Global Useless Class.' *The New York Times*. Retrieved from https://www.nytimes.com/2018/03/19/world/europe/yuval-noah-harari-future-tech.html

Friedman, B. (1995). "It's the Computer's Fault": Reasoning About Computers as Moral Agents. In *Conference Companion on Human Factors in Computing Systems* (pp. 226–227). New York, NY, US: ACM. https://doi.org/10.1145/223355.223537

Friedman, B., Kahn, P. H., & Hagman, J. (2003). Hardware Companions?: What Online AIBO Discussion Forums Reveal About the Human-robotic Relationship. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 273–280). New York, NY, US: ACM. https://doi.org/10.1145/642611.642660

Garattini, C., Raffle, J., Aisyah, D. N., Sartain, F., & Kozlakidis, Z. (2017). Big Data Analytics, Infectious Diseases and Associated Ethical Impacts. *Philosophy & Technology*, 1–17. https://doi.org/10.1007/s13347-017-0278-y

Gausel, N., Leach, C. W., Vignoles, V. L., & Brown, R. (2012). Defend or repair? Explaining responses to in-group moral failure by disentangling feelings of shame, rejection, and inferiority. *Journal of Personality and Social Psychology*, *102*(5), 941–960. https://doi.org/10.1037/a0027233

Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in Online Shopping: An Integrated Model. *MIS Quarterly*, *27*(1), 51–90.

Geiselman, E. E., Johnson, C. M., & Buck, D. R. (2013). Flight Deck Automation Invaluable Collaborator or Insidious Enabler? *Ergonomics in Design: The Quarterly of Human Factors Applications*, *21*(3), 22–26. https://doi.org/10.1177/1064804613491268

Gilbert, Daniel Todd, Fiske, S. T., & Lindzey, G. (Eds.). (1998). *The Handbook of Social Psychology*. Oxford University Press.

Gino, F., Brooks, A. W., & Schweitzer, M. E. (2012). Anxiety, advice, and the ability to discern: feeling anxious motivates individuals to seek and use advice. *Journal of Personality and Social Psychology*, *102*(3), 497–512. https://doi.org/10.1037/a0026413

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

Gino, F., & Schweitzer, M. E. (2008). Blinded by anger or feeling the love: how emotions influence advice taking. *The Journal of Applied Psychology*, *93*(5), 1165–1173. https://doi.org/10.1037/0021-9010.93.5.1165

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Goodwin, P., Fildes, R., Lawrence, M., & Nikolopoulos, K. (2007). The process of using a forecasting support system. *International Journal of Forecasting*, *23*(3), 391–404. https://doi.org/10.1016/j.ijforecast.2007.05.016

Goodwin, P., Gönül, M., & Önkal, D. (2013). Antecedents and effects of trust in forecasting advice. *International Journal of Forecasting*, *29*(2), 354–366. https://doi.org/10.1016/j.ijforecast.2012.08.001

Grossman, D. (2009). *On Killing: The Psychological Cost of Learning to Kill in War and Society* (Revised edition). New York: Back Bay Books.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychological Assessment*, *12*(1), 19–30.

Habib, K. (2017). *Automatic vehicle control systems* (No. PE 16-007) (p. 12). National Highway Traffic Safety Administration. Retrieved from https://static.nhtsa.gov/odi/inv/2016/INCLA-PE16007-7876.PDF

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., Visser, E. J. de, & Parasuraman, R. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Harvey, N., & Fischer, I. (1997). Taking Advice: Accepting Help, Improving Judgment, and Sharing Responsibility. *Organizational Behavior and Human Decision Processes*, *70*(2), 117–133. https://doi.org/10.1006/obhd.1997.2697

Hevelke, A., & Nida-Rümelin, J. (2014). Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis. *Science and Engineering Ethics*, *21*(3), 619–630. https://doi.org/10.1007/s11948-014-9565-5

Hoff, K. A., & Bashir, M. (2015). Trust in Automation Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Horberg, E. J., Oveis, C., & Keltner, D. (2011). Emotions as Moral Amplifiers: An Appraisal Tendency Approach to the Influences of Distinct Emotions upon Moral Judgment. *Emotion Review*, *3*(3), 237–244. https://doi.org/10.1177/1754073911402384

Houser, M. L. (2006). Expectancy Violations of Instructor Communication as Predictors of Motivation and Learning: A Comparison of Traditional and Nontraditional Students. *Communication Quarterly*, *54*(3), 331–349. https://doi.org/10.1080/01463370600878248

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion; psychological studies of opinion change* (Vol. xii). New Haven, CT, US: Yale University Press.

Inthorn, J., Tabacchi, M. E., & Seising, R. (2015). Having the Final Say: Machine Support of Ethical Decisions of Doctors. In S. P. V. Rysewyk & M. Pontier (Eds.), *Machine Medical Ethics* (pp. 181–206). Springer International Publishing. https://doi.org/10.1007/978-3-319-08108-3_12

Jen, W.-Y., & Hung, M.-C. (2010). An Empirical Study of Adopting Mobile Healthcare Service: The Family's Perspective on the Healthcare Needs of Their Elderly Members. *Telemedicine and E-Health*, *16*(1), 41–48. https://doi.org/10.1089/tmj.2009.0093

Jha, S. (2012). Punishing the Lemon: The Ethics of Actuarial Fairness. *Journal of the American College of Radiology*, *9*(12), 887–893. https://doi.org/10.1016/j.jacr.2012.09.012

Ji, M. (2017). ARE ROBOTS GOOD FIDUCIARIES? REGULATING ROBO-ADVISORS UNDER THE INVESTMENT ADVISERS ACT OF 1940. *Columbia Law Review*, *117*(6), 1543–1583.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Jones, G. R., & George, J. M. (1998). The Experience and Evolution of Trust: Implications for Cooperation and Teamwork. *Academy of Management Review*, *23*(3), 531–546. https://doi.org/10.5465/AMR.1998.926625

Jordan, J., Leliveld, M. C., & Tenbrunsel, A. E. (2015). The Moral Self-Image Scale: Measuring and Understanding the Malleability of the Moral Self. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01878

Kahn, P. H., Ishiguro, H., Friedman, B., & Kanda, T. (2006). What is a Human? - Toward Psychological Benchmarks in the Field of Human-Robot Interaction. In *The 15th IEEE International Symposium on Robot and Human Interactive Communication, 2006* (pp. 364–371). https://doi.org/10.1109/ROMAN.2006.314461

Kahn, P. H., Jr., Kanda, T., Ishiguro, H., Gill, B. T., Shen, S., Gary, H. E., & Ruckert, J. H. (2015). Will People Keep the Secret of a Humanoid Robot?: Psychological Intimacy in HRI. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 173–180). New York, NY, USA: ACM. https://doi.org/10.1145/2696454.2696486

Kahn, P. H., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., … Severson, R. L. (2012). Do people hold a humanoid robot morally accountable for the harm it causes? In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 33–40).

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–292.

Kaplan, M. F., & Miller, C. E. (1987). Group decision making and normative versus informational influence: Effects of type of issue and assigned decision rule. *Journal of Personality and Social Psychology*, *53*(2), 306–313. https://doi.org/10.1037/0022-3514.53.2.306

Kaplan, M. F., Schaefer, E. G., & Zinkiewicz, L. (1994). Member Preferences for Discussion Content in Anticipated Group Decisions: Effects of Type of issue and Group Interactive Goal. *Basic and Applied Social Psychology*, *15*(4), 489–508. https://doi.org/10.1207/s15324834basp1504_6

Katsikopoulos, K. V., Pachur, T., Machery, E., & Wallin, A. (2008). From Meehl to Fast and Frugal Heuristics (and Back) New Insights into How to Bridge the Clinical—Actuarial Divide. *Theory & Psychology*, *18*(4), 443–464. https://doi.org/10.1177/0959354308091824

Katz, J. E., & Halpern, D. (2014). Attitudes towards robots suitability for various jobs as affected robot appearance. *Behaviour & Information Technology*, *33*(9), 941–953. https://doi.org/10.1080/0144929X.2013.783115

King, A. C., Hekler, E. B., Castro, C. M., Buman, M. P., Marcus, B. H., Friedman, R. H., & Napolitano, M. A. (2014). Exercise advice by humans versus computers: maintenance effects at 18 months. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, *33*(2), 192–196. https://doi.org/10.1037/a0030646

Kristoffersson, A., Coradeschi, S., Loutfi, A., & Severinson-Eklundh, K. (2011). An Exploratory Study of Health Professionals' Attitudes about Robotic Telepresence Technology. *Journal of Technology in Human Services*, *29*(4), 263–283. https://doi.org/10.1080/15228835.2011.639509

Kühberger, A. (1998). The Influence of Framing on Risky Decisions: A Meta-analysis. *Organizational Behavior and Human Decision Processes*, *75*(1), 23–55. https://doi.org/10.1006/obhd.1998.2781

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, *22*(3), 177–189. https://doi.org/10.1016/0022-1031(86)90022-3

Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: the importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology*, *93*(2), 234–249. https://doi.org/10.1037/0022-3514.93.2.234

Lee, J., Bachrach, D. G., & Lewis, K. (2014). Social Network Ties, Transactive Memory, and Performance in Groups. *Organization Science*, *25*(3), 951–967. https://doi.org/10.1287/orsc.2013.0884

Lee, J. E., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lee, Y., Kozar, K. A., & Larsen, K. R. T. (2003). The Technology Acceptance Model: Past, Present, and Future. *Communications of the Association for Information Systems*, *12*, 752–780.

Lerch, F. J., Prietula, M. J., & Kulik, C. T. (1997). Expertise in Context. In P. J. Feltovich, K. M. Ford, & R. R. Hoffman (Eds.) (pp. 417–448). Cambridge, MA, USA: MIT Press. Retrieved from http://dl.acm.org/citation.cfm?id=278605.278625

Levine, T. R. (2014). Truth-Default Theory (TDT) A Theory of Human Deception and Deception Detection. *Journal of Language and Social Psychology*, *33*(4), 378–392.

Lewicki, R. J., Tomlinson, E. C., & Gillespie, N. (2006). Models of Interpersonal Trust Development: Theoretical Approaches, Empirical Evidence, and Future Directions. *Journal of Management*, *32*(6), 991–1022. https://doi.org/10.1177/0149206306294405

Li, J., Kizilcec, R., Bailenson, J., & Ju, W. (2016). Social robots and virtual agents as lecturers for video instruction. *Computers in Human Behavior*, *55*, *Part B*, 1222–1230. https://doi.org/10.1016/j.chb.2015.04.005

Livnat Yuval. (2004). On the Nature of Benevolence. *Journal of Social Philosophy*, *35*(2), 304–317. https://doi.org/10.1111/j.1467-9833.2004.00234.x

Logg, J. (2017). *Theory of Machine: When Do People Rely on Algorithms?* (SSRN Scholarly Paper No. ID 2941774). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=2941774

Lyons, J. B., & Stokes, C. K. (2012). Human-human reliance in the context of automation. *Human Factors*, *54*(1), 112–121.

MacGeorge, E. L. (2016). Advice: Expanding the communication paradigm. *Communication Yearbook*, *40*, 213–244.

MacGeorge, E. L., Guntzviller, L. M., Hanasono, L. K., & Feng, B. (2013). Testing Advice Response Theory in Interactions With Friends. *Communication Research*, *43*(2), 211–231. https://doi.org/10.1177/0093650213510938

MacGeorge, E. L., & Van Swol, L. M. (2018). *The Oxford Handbook of Advice*. Oxford, New York: Oxford University Press.

Madhavan, P., & Wiegmann, D. A. (2007a). Effects of Information Source, Pedigree, and Reliability on Operator Interaction With Decision Support Systems. *Human Factors: The*

Journal of the Human Factors and Ergonomics Society, *49*(5), 773–785. https://doi.org/10.1518/001872007X230154

Madhavan, P., & Wiegmann, D. A. (2007b). Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301. https://doi.org/10.1080/14639220500337708

Malin, J. L. (2013). Envisioning Watson As a Rapid-Learning System for Oncology. *Journal of Oncology Practice*, *9*(3), 155–157. https://doi.org/10.1200/JOP.2013.001021

Malle, B. F., & Scheutz, M. (2015). When will people regard robots as morally competent social partners? In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 486–491). https://doi.org/10.1109/ROMAN.2015.7333667

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model Of Organizational Trust. *Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.5465/AMR.1995.9508080335

Meehl, P. E. (1986). Causes and Effects of My Disturbing Little Book. *Journal of Personality Assessment*, *50*(3), 370–375. https://doi.org/10.1207/s15327752jpa5003_6

Mehta, R., Rice, S., & Winter, S. (2014). Examining the Relationship between Familiarity and Reliability of Automation in the Cockpit. *Collegiate Aviation Review*, *32*(2), 1–13.

Merritt, S. M. (2011). Affective Processes in Human–Automation Interactions. *Human Factors*, *53*(4), 356–370. https://doi.org/10.1177/0018720811411912

Merritt, S. M., Unnerstall, J. L., Lee, D., & Huber, K. (2015). Measuring Individual Differences in the Perfect Automation Schema. *Human Factors: The Journal of the Human Factors*

and Ergonomics Society, 0018720815581247. https://doi.org/10.1177/0018720815581247

Moon, Aj., Danielson, P., & Loos, H. F. M. V. der. (2012). Survey-Based Discussions on Morally Contentious Applications of Interactive Robotics. *International Journal of Social Robotics*, *4*(1), 77–96. https://doi.org/10.1007/s12369-011-0120-0

Moor, J. H. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, *21*(4), 18–21. https://doi.org/10.1109/MIS.2006.80

Murdoch, & Detsky. (2013). The inevitable application of big data to health care. *JAMA*, *309*(13), 1351–1352. https://doi.org/10.1001/jama.2013.393

Naastepad, C. W. M., & Mulder, J. M. (2018). Robots and us: towards an economics of the 'Good Life.' *Review of Social Economy*, *0*(0), 1–33. https://doi.org/10.1080/00346764.2018.1432884

Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, *56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Nass, C., Moon, Y., Fogg, B. J., Reeves, B., & Dryer, D. C. (1995). Can computer personalities be human personalities? *International Journal of Human-Computer Studies*, *43*(2), 223–239. https://doi.org/10.1006/ijhc.1995.1042

Naweed, D. A., Bearman, D. C., Dorrian, D. J., Rose, M. J., & Dawson, P. D. (2013). *Evaluation of Rail Technology: A Practical Human Factors Guide*. Ashgate Publishing, Ltd.

Noble, D. (2017). *Forces of Production: A Social History of Industrial Automation*. Routledge.

Önkal, D., Goodwin, P., Thomson, M., Gönül, M., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, *22*(4), 390–409. https://doi.org/10.1002/bdm.637

Oswald, A. J., Proto, E., & Sgroi, D. (2015). Happiness and Productivity. *Journal of Labor Economics*, *33*(4), 789–822. https://doi.org/10.1086/681096

Palmer, J. K., & Gore, J. S. (2014). A Theory of Contrast Effects in Performance Appraisal and Social Cognitive Judgments. *Psychological Studies*, *59*(4), 323–336. https://doi.org/10.1007/s12646-014-0282-6

Palmgren, P., & Rayburn, J. D. (1982). Gratifications sought and media exposure an expectancy value model. *Communication Research*, *9*(4), 561–580. https://doi.org/10.1177/009365082009004004

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354

Parasuraman, R., Visser, E. de, Wiese, E., & Madhavan, P. (2014). Human Trust in Other Humans, Automation, Robots, and Cognitive Agents Neural Correlates and Design Implications. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*(1), 340–344. https://doi.org/10.1177/1541931214581070

Parry, K., Cohen, M., & Bhattacharya, S. (2016). Rise of the Machines: A Critical Consideration of Automated Leadership Decision Making in Organizations. *Group & Organization Management*, *41*(5), 571–594. https://doi.org/10.1177/1059601116643442

Pop, V. L., Shrewsbury, A., & Durso, F. T. (2015). Individual Differences in the Calibration of Trust in Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(4), 545–556. https://doi.org/10.1177/0018720814564422

Prahl, A. (2015). *Mortal versus machine: Development of a model to differentiate human and automated trust*. University of Wisconsin-Madison, Madison, WI, USA.

Prahl, A., & Enright, R. (2017). Forgiving Computers: The Rise of Automation and Implications for Counseling. *Counseling and Values*, *62*(2), 144–158. https://doi.org/10.1002/cvj.12056

Pruitt, D. G., & Rubin, J. Z. (1986). *Social Conflict: Escalation, Stalemate, and Settlement*. New York, NY: Random House.

Rice, S., & Geels, K. (2010). Using System-Wide Trust Theory to Make Predictions About Dependence on Four Diagnostic Aids. *The Journal of General Psychology*, *137*(4), 362–375. https://doi.org/10.1080/00221309.2010.499397

Rozin, P., & Royzman, E. B. (2001). Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review*, *5*(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Savela, N., Turja, T., & Oksanen, A. (2017). Social Acceptance of Robots in Different Occupational Fields: A Systematic Literature Review. *International Journal of Social Robotics*, 1–10. https://doi.org/10.1007/s12369-017-0452-5

Schaefer, K. E. (2016). Measuring Trust in Human Robot Interactions: Development of the "Trust Perception Scale-HRI." In R. Mittu, D. Sofge, A. Wagner, & W. F. Lawless (Eds.), *Robust Intelligence and Trust in Autonomous Systems* (pp. 191–218). Springer US. https://doi.org/10.1007/978-1-4899-7668-0_10

Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An Integrative Model of Organizational Trust: Past, Present, and Future. *The Academy of Management Review*, *32*(2), 344–354. https://doi.org/10.2307/20159304

Schwarz, N., & Bles, H. (1992). Scandals and the Public's Trust in Politicians: Assimilation and Contrast Effects. *Personality and Social Psychology Bulletin*, *18*(5), 574–579. https://doi.org/10.1177/0146167292185007

Shen, S. (2011). The curious case of human-robot morality. In *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 249–250). https://doi.org/10.1145/1957656.1957755

Siddiqui, F. (2018, April 1). NTSB 'unhappy' with Tesla release of investigative information in fatal crash. *Washington Post*. Retrieved from https://www.washingtonpost.com/news/dr-gridlock/wp/2018/04/01/ntsb-unhappy-with-tesla-release-of-investigative-information-in-fatal-crash/

Sitkin, S. B., & Pablo, A. L. (1992). Reconceptualizing the Determinants of Risk Behavior. *Academy of Management Review*, *17*(1), 9–38. https://doi.org/10.5465/AMR.1992.4279564

Sitkin, S. B., & Roth, N. L. (1993). Explaining the Limited Effectiveness of Legalistic "Remedies" for Trust/Distrust. *Organization Science*, *4*(3), 367–392. https://doi.org/10.1287/orsc.4.3.367

Sniezek, J. A., & Buckley, T. (1995). Cueing and Cognitive Conflict in Judge-Advisor Decision Making. *Organizational Behavior and Human Decision Processes*, *62*(2), 159–174. https://doi.org/10.1006/obhd.1995.1040

Sniezek, J. A., & Van Swol, L. (2001). Trust, Confidence, and Expertise in a Judge-Advisor System. *Organizational Behavior and Human Decision Processes*, *84*(2), 288–307. https://doi.org/10.1006/obhd.2000.2926

Stasser, G., & Stewart, D. (1992). Discovery of hidden profiles by decision-making groups: Solving a problem versus making a judgment. *Journal of Personality and Social Psychology*, *63*(3), 426–434. https://doi.org/10.1037/0022-3514.63.3.426

Stone, M. (2012). The death of personal service: Will financial services customers who serve themselves do better than if they are served? *Journal of Database Marketing & Customer Strategy Management*, *19*(2), 107–119. https://doi.org/10.1057/dbm.2012.8

Strauss, J., Barrick, M., & Connerley, M. (2010). An investigation of personality similarity effects (relational and perceived) on peer and supervisor ratings and the role of familiarity and liking. *Journal of Occupational and Organizational Psychology*, *74*(5), 637–657. https://doi.org/10.1348/096317901167569

Strayed, C. (2012). *Wild: From Lost to Found on the Pacific Crest Trail*. Knopf Doubleday Publishing Group.

Sunil, J., & Thirgood, J. (2016). *Working Without a Net: Rethinking Canada's social policy in the new age of work* (Mowat Research No. No. 132). Mowat Centre.

Sutherland, S. C., Harteveld, C., & Young, M. E. (2016). Effects of the Advisor and Environment on Requesting and Complying With Automated Advice. *ACM Trans. Interact. Intell. Syst.*, *6*(4), 27:1–27:36. https://doi.org/10.1145/2905370

Takayama, L., Ju, W., & Nass, C. (2008). Beyond Dirty, Dangerous and Dull: What Everyday People Think Robots Should Do. In *Proceedings of the 3rd ACM/IEEE International*

Conference on Human Robot Interaction (pp. 25–32). New York, NY, USA: ACM. https://doi.org/10.1145/1349822.1349827

Tasselli, S., Kilduff, M., & Menges, J. I. (2015). The Microfoundations of Organizational Social Networks: A Review and an Agenda for Future Research. *Journal of Management*, *41*(5), 1361–1387. https://doi.org/10.1177/0149206315573996

Thompson, D. (2015, August). A World Without Work. *The Atlantic*. Retrieved from http://www.theatlantic.com/magazine/archive/2015/07/world-without-work/395294/

Tindale, R. S., Kameda, T., & Hinsz, V. B. (2003). Group decision making: Review and integration. In M. A. Hogg & J. Cooper (Eds.), *Sage handbook of social psychology* (pp. 381–403). London: Sage.

Tomlinson, E. C., & Mayer, R. C. (2009). The Role of Causal Attribution Dimensions in Trust Repair. *Academy of Management Review*, *34*(1), 85–104. https://doi.org/10.5465/AMR.2009.35713291

Tost, L. P., Gino, F., & Larrick, R. P. (2012). Power, competitiveness, and advice taking: Why the powerful don't listen. *Organizational Behavior and Human Decision Processes*, *117*(1), 53–65. https://doi.org/10.1016/j.obhdp.2011.10.001

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, *103*(3), 299–314. https://doi.org/10.1037/0033-2909.103.3.299

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science (New York, N.Y.)*, *211*(4481), 453–458.

Tversky, Amos, & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Tzioti, S. C., Wierenga, B., & Van Osselaer, S. M. J. (2014). The Effect of Intuitive Advice Justification on Advice Taking. *Journal of Behavioral Decision Making*, *27*(1), 66–77. https://doi.org/10.1002/bdm.1790

van Dierendonck, D., & Jacobs, G. (2012). Survivors and Victims, a Meta-analytical Review of Fairness and Organizational Commitment after Downsizing. *British Journal of Management*, *23*(1), 96–109. https://doi.org/10.1111/j.1467-8551.2010.00724.x

Van Swol, L. (2011). Forecasting another's enjoyment versus giving the right answer: Trust, shared values, task effects, and confidence in improving the acceptance of advice. *International Journal of Forecasting*, *27*(1), 103–120. https://doi.org/10.1016/j.ijforecast.2010.03.002

Van Swol, L., MacGeorge, E. L., & Prahl, A. (2015). The effects of advice solicitation, confidence, and expertise on advice utilization. Presented at the International Communication Association Conference, San Juan, Puerto Rico.

Van Swol, L., & Sniezek, J. A. (2005). Factors affecting the acceptance of expert advice. *British Journal of Social Psychology*, *44*(3), 443–461. https://doi.org/10.1348/014466604X17092

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, *27*(3), 425–478.

Voiklis, J., Kim, B., Cusimano, C., & Malle, B. F. (2016). Moral judgments of human vs. robot agents. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 775–780). https://doi.org/10.1109/ROMAN.2016.7745207

Waal, F. B. M. de. (2009). *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Harvard University Press.

Wærn, Y., & Ramberg, R. (1996). People's perception of human and computer advice. *Computers in Human Behavior*, *12*(1), 17–27. https://doi.org/10.1016/0747-5632(95)00016-X

Wallis, M., Popat, S., McKinney, J., Bryden, J., & Hogg, D. C. (2010). Embodied conversations: performance and the design of a robotic dancing partner. *Design Studies*, *31*(2), 99–117. https://doi.org/10.1016/j.destud.2009.09.001

Wasen, K. (2010). Replacement of Highly Educated Surgical Assistants by Robot Technology in Working Life: Paradigm Shift in the Service Sector. *International Journal of Social Robotics*, *2*(4), 431–438. https://doi.org/10.1007/s12369-010-0062-y

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

Wedell, D. H., Parducci, A., & Geiselman, R. E. (1987). A formal analysis of ratings of physical attractiveness: Successive contrast and simultaneous assimilation. *Journal of Experimental Social Psychology*, *23*(3), 230–249. https://doi.org/10.1016/0022-1031(87)90034-5

Wise, J., Rio, A., & Fedouach, M. (2011). What really happened aboard Air France 447. *Popular Mechanics*, *6*, 35–36.

Wittenbaum, G. M., Shulman, H. C., & Braz, M. E. (2010). Social Ostracism in Task Groups: The Effects of Group Composition. *Small Group Research*, *41*(3), 330–353. https://doi.org/10.1177/1046496410363914

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, *93*(1), 1–13. https://doi.org/10.1016/j.obhdp.2003.08.002

Yaniv, I., Choshen-Hillel, S., & Milyavsky, M. (2011). Receiving advice on matters of taste: Similarity, majority influence, and taste discrimination. *Organizational Behavior and Human Decision Processes*, *115*(1), 111–120. https://doi.org/10.1016/j.obhdp.2010.11.006

Yaniv, I., & Kleinberger, E. (2000). Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation. *Organizational Behavior and Human Decision Processes*, *83*(2), 260–281. https://doi.org/10.1006/obhd.2000.2909

Zaltman, G., & Moorman, C. (1988). The importance of personal trust in the use of research. *Journal of Advertising Research*, *28*(5), 16–24.

Appendix I

Supplemental Analyses

Advice Justifications

I did collect additional data with identical stimuli to Study 2, however the advice was accompanied with a justification (analytic vs. intuitive justification). The justifications was similar to short phrases used in previous research (Tzioti et al., 2014). To fully cross advice justification type with replacement/no replacement and context financial/moral would result in two 4x4 experiments, and such a large project is outside the scope of this research project. As a result, I considered the replacement (human/automation replaced by automation/human) scenarios to be of the most applied and theoretical interest because they have hardly been studied and they most closely reflect what is happening in the real world when automated advisors replace human advisors. Additionally, only the business decision-context (from study one) will be studied. I do this for two reasons: first, this decision-context is the easiest to relate to past research (Prahl & Van Swol, 2017; Tzioti et. al., 2014). Second, past research (e.g., Tzioti et. al. 2014) suggests that the task from study 1 is suited towards analytic justifications – thus we expect the greatest effect size in the more analytic decision contexts.

Advice Justifications

The following reasons (justifications) were randomly presented in the intuitive justification condition. These justifications preceded the numerical forecast advice (i.e., 42): (1) A gut feeling leads me to suggest, (2) Intuition leads me to forecast, (3) A strong feeling lead me to (4) Intuitive reasoning makes me suggest. The following reasons (justifications) were randomly presented in the analytic justification condition. These justifications preceded the

numerical forecast advice: (1) Data analysis suggests, (2) Some statistical reasoning leads to (3) The math says that it will be (4) The data trends point to.

Rationale

The message focused paradigm of interpersonal advice is primarily concerned with the content and characteristics of advice messages themselves (MacGeorge, 2016). Advice messages may vary in a variety of ways such as politeness, face threat, and feasibility (MacGeorge et al., 2013). One common characteristic of advice messages in the real world is that they are accompanied with a justification of why the advisor has proposed the action. For example, a manager may not simply advise a subordinate on what to do, but also offer a reason for it, "I have a lot of experience with this sort of situation, so I advise you…" The addition of a justification to persuasive messages has been found to increase compliance even when the justification is relatively meaningless (Langer, Blank, & Chanowitz, 1978), or in forecasting research, the justification only repeats what the decision-maker can already see (Goodwin et al., 2013). In interpersonal advice research, advice with more explanation of the potential efficacy or feasibility of the advised action leads to greater implementation intention by decision-makers (Feng & Burleson, 2008).

Yaniv & Kleinberger (2000) suggest that the reason advice is often underutilized (i.e., not weighted equally to the decision-makers own estimate) is because decision-makers are "egocentric." One cause of this egocentrism may be that decision-makers are able to understand the reasoning behind their own estimate, but not their advisors. This leads to the decision-makers having greater confidence in their own estimate as it seems more justified to them. However, when advisors provide more information it allows the decision-maker to better understand the

advice – this applies beyond interpersonal advice (e.g., Feng & Burleson, 2008) and is evident with automated advisors, too. For example, the provision of confidence intervals for a forecast that comes from an algorithmic forecasting system can increase utliziation of the advice (Adya & Edward, 2012; Goodwin et al., 2013). However, not all forms of additional information are necessarily helpful. In Goodwin et al. (2013), confidence intervals that indicated a high level of uncertainty in the advised forecast did not increase utilization of the advice, and the communication of too much uncertainty has the potential to decrease utlization (Yaniv, 1997). A direct way of giving the reasoning behind a decision is providing a *justification* for the advice. Recent research suggests that providing a justification for advice reduces the tendency to discount advice because it provides direct access to the reasoning of the advisor (Tzioti et al., 2014; Yaniv, 2004).

H1: Advice that is justified will be utilized more than advice that is not justified.

Advice justifications fall into many categories, but recent research has investigated two broad types of justifications, intuitive vs. analytical justifications (Tzioti et al., 2014). These two categories are related to other broad classifications of thought processes that follow a dual process model, where "system 1" thinking tends to be absent of systematic thought and heuristic in nature, and "system 2" requires more cognitive effort and time to carefully process information (Tversky & Kahneman, 1974). Recently, research has shown that the inclusion of either type of justification can increase the utilization of advice depending on the decision context (Tzioti et al., 2014). However, in the same study, intuitive advice delivered on an analytical task (the analysis of market trends) decreased the utilization of that advice. Similar to other interpersonal advice studies (Sniezek & Van Swol, 2001; Van Swol & Sniezek, 2005), this effect was moderated by advisor status, more senior and more experienced advisors were more

likely to have intuitive advice utilized. Because expertise is partially a function of the type of problem advised on, this suggests that the type of justification must be consistent with the expectations that the decision-maker has about the problem and the advisor.

The studies proposed below manipulate the decision context as well as the type of advisor. However, the above review of advice justifications suggests another manipulation: the inclusion of different types of advice justifications. Humans perceive the qualities of automated and human advisors differently. If automated advisors are perceived as better (or "perfect") for quantitative reasoning tasks; and humans better at tasks requiring qualitative/moral reasoning, this effect may be moderated by providing justifications for advice messages. Research suggests that advice messages will likely be utilized more if accompanied by any justification, but this research has been conducted almost entirely with interpersonal advice.

Receiving advice that is justified intuitively ("my gut tells me," "I have a strong feeling…") from an automated source seems peculiar. After all, intuitive justifications and gut feelings are things that require consciousness, something we do not believe automated advisors have (Hogarth, 2001; Kahn et al., 2006). Therefore, receiving intuitive advice from an automated advisor may constitute an expectation violation. Like poor advice received from automation, this expectation violation could result in a significant drop in the use of automated advice. Human advisors, on the other hand, are expected to be capable of both intuitive and analytical judgement. So, while a decision-maker may perceive certain types of human-provided justifications as more appropriate for a problem, the justification itself is not an expectation violation. We therefore expect the effect of advice justifications to have differing effects for humans and automation: machines are expected to be analytical, justifications that are consistent

with this expectation should increase advice utilization, while justifications that violate this expectation should decrease utilization.

H2a: Automated advice accompanied by an intuitive justification will be utilized less than human advice that is accompanied by an intuitive justification.

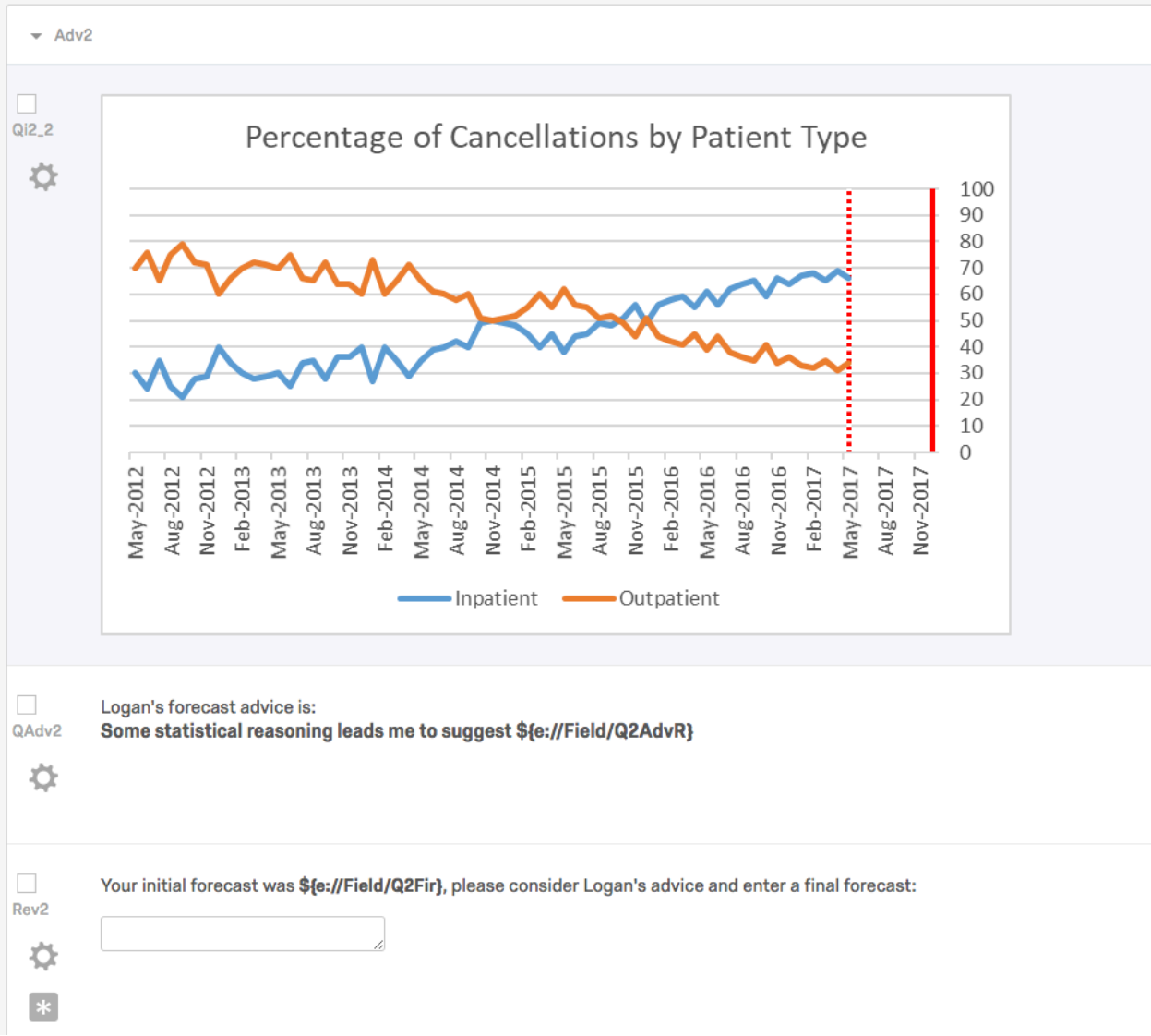
H2b: Automated advice accompanied by an analytical justification will be utilized more than human advice that is accompanied by an analytical justification.

Finally, there is not sufficient literature to predict the effects of advice justifications on the perfection automation schema. The initial bias towards automated advice has been observed only when "all things are equal" between advisors at the very start of an experiment. It is possible that the provision of justifications will cause this effect to disappear because when a human provides a justification, it provides a window into the rationalization of the advice – increasing utilization (Yaniv, 2004). However, machines do not really "think" or rationalize behavior, so the justification could be perceived as something the system is programmed to provide, but not actually a window into a thought process.

Regarding the punishment of mistakes, the justifications could attenuate the effect if the provision of intuitive justifications causes decision-makers to treat the automated advisors more like humans. But, we do not know if intuitive justifications alone are sufficient to elicit more human/social responses to automation. If justifications do change the bias towards automation advice and the increased punishment of mistakes, it will have important real-world implications for system designers.

RQ2: Does the inclusion of advice justifications moderate the effects of the perfect automation schema?

Figure A1: The advice/revision screen (includes justification).



Results

The measures were identical to studies 2 & 3 above. For comparison to advice that was not justified, we used the data collected in the management condition in study 2. Note that study 2 included similar advisor replacement conditions and different advisor replacement conditions, the justifications data comes only from different replacement conditions. Thus, we only compared data in the second block of trials in the different advisor replacement conditions.

Hypothesis 1 suggested that advice that was justified would be utilized more than advice that was not justified. I did not find support for this hypothesis. I conducted additional analyses to see if justified advice was used more than non-justified advice for each advisor type. I also did not find support for this hypothesis. These analyses (1-5) are summarized in Table 1.

Table 1: Hypothesis 1ANOVA Tests

Test	$F(1,382)$	p	ηp^2
1	0.021	0.885	0.000
2	0.308	0.579	0.001
3	0.044	0.835	0.000
4	1.237	1.237	0.002
5	0.001	0.988	0.000

1: Advice utilization across all 20 trials regardless of advisor, effect of justifications
2: Advice utilization across first 10 trials regardless of advisor, effect of justifications
3: Advice utilization across second 10 trials regardless of advisor, effect of justifications
4: Advice utilization across first 10 trials: advisor type (human/automated) * justifications interaction
5: Advice utilization across second 10 trials: advisor type (human/automated) * justifications interaction

Hypotheses 2a posited that human advice would be used more when accompanied with an intuitive justification compared to when it was accompanied with an analytical justification. Hypotheses 2b posited that automated advice would be used more when accompanied with an analytical justification compared to when it was accompanied with an intuitive justification. We did not find evidence to support either hypothesis, results are summarized in table 2.

Table 2: Hypothesis 2 ANOVA Tests

Test	F	df	p	ηp^2
1	1.187	(1,165)	0.278	0.007
2	0.001	(1,160)	0.982	0.000
3	0.004	(1,160)	0.722	0.001

4	0.491	(1,165)	0.484	0.003

1: Advice utilization across first 10 trials for automated advisor, effect of justification type
2: Advice utilization across first 10 trials for human advisor, effect of justification type
3: Advice utilization across second 10 trials for automated advisor, effect of justification type
4: Advice utilization across second 10 trials for human advisor, effect of justification type

Our research questions intended to investigate the perfect automation schema and if justification type may change the effects of it. For brevity, we conducted a series of ANOVA tests on a number of variables of interest given our results from study 1, 2 & 3. The variables of interest were the first three trials with each advisor and the difference between the mean of the first 6 trials (good advice period) and the last 4 trials (unreliable advice period). To summarize, we found no evidence that justifications or justification type changes the effects established in study 2. We found no significant interactions between advisor type and justification type; however, we did find a number of main effects for advisor type that largely replicate the results from study 2. Namely, in the second block of trials, automated advisors are utilized more on the initial trails but also suffer a greater decrease in advice utilization after delivering unreliable advice. Results are summarized in Tables 3 & 4.

Table 3: RQ ANOVA Tests

Test	*F*	*df*	*p*	ηp^2
First 3 Trials, First Block[A]	0.282	(1,325)	0.596	0.001
Decrease in Utilization in Unreliable Advice Period, First Block[A]	0.207	(1,325)	0.604	0.001
First 3 Trials, Second Block[A]	0.999	(1,325)	0.318	0.003
Decrease in Utilization in Unreliable Advice Period, Second Block[A]	0.938	(1,325)	0.333	0.003
First 3 Trials, Second Block[B]	**8.074**	**(1,325)**	**0.005**	**0.024**
Decrease in Utilization in Unreliable Advice Period, Second Block[B]	**6.826**	**(1,325)**	**0.009**	**0.021**

A: Interaction effect between advisor and justification type
B: Main effect of advisor type only

Table 4: Hypothesis RQ Descriptive

Test	*Automated Advisor Mean*	*Human Advisor Mean*
First 3 Trials, Second Block	0.533	0.452
Decrease in Utilization in Unreliable Advice Period, Second Block	-.1424	-.0731

Conclusion

Overall the results from adding justifications to the advice do not appear to change the effects on advice utilization established in studies 1, 2 & 3. It may be that our manipulation was not strong enough given that there was a great deal of visual stimuli (i.e., graphs) for participants to process in addition to the advice. The results do help validate the findings from the studies presented earlier and suggest that further study in the area of justifications is unlikely to be as informative as the other areas for future study outlined in the discussion section(s) and conclusion of the research presented in studies 1, 2 & 3.

Appendix II

Figures and Tables

Figure 1.1: Utilization between trials 6 & 7

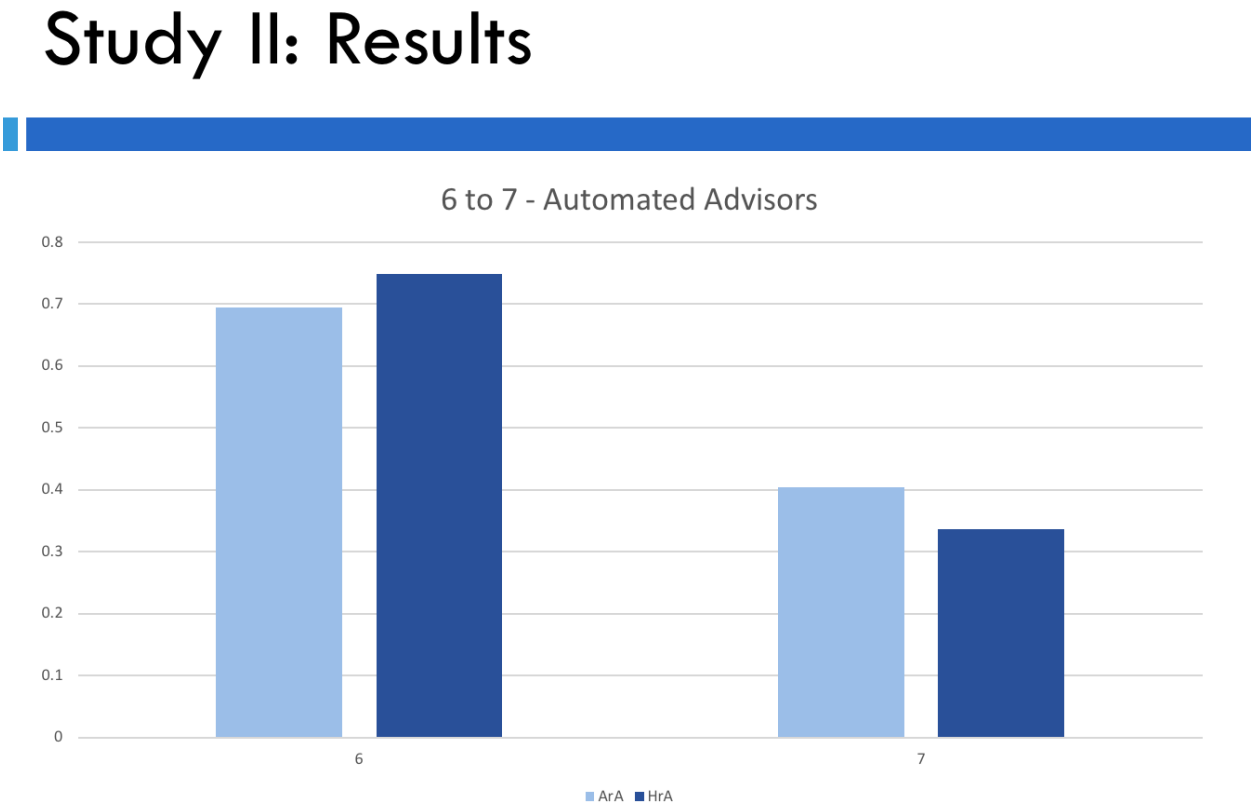


Figure 1.2: Utilization between trials 6 & 7

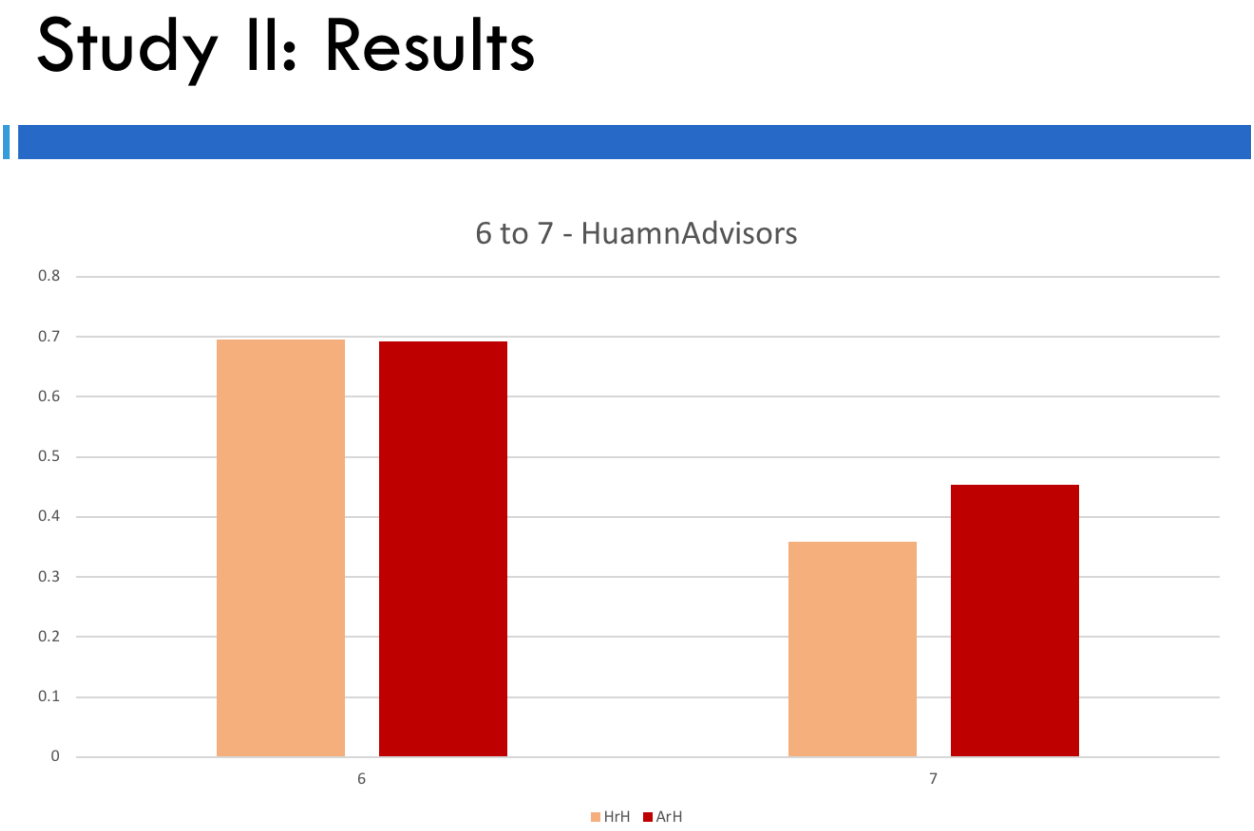
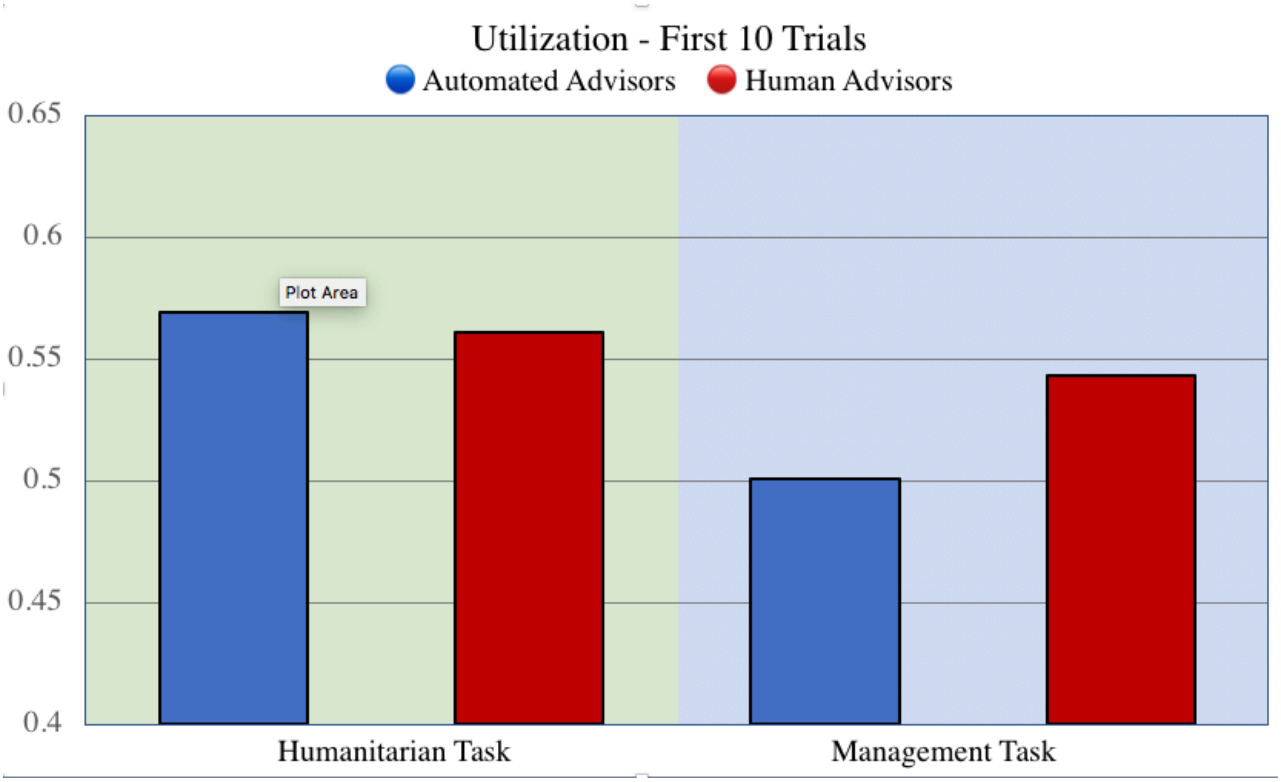
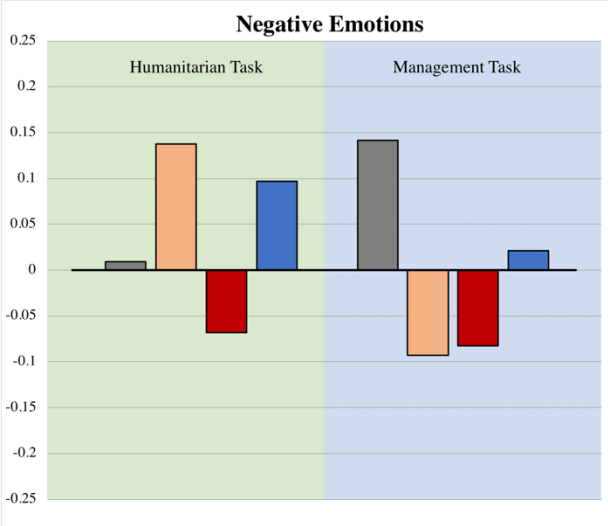
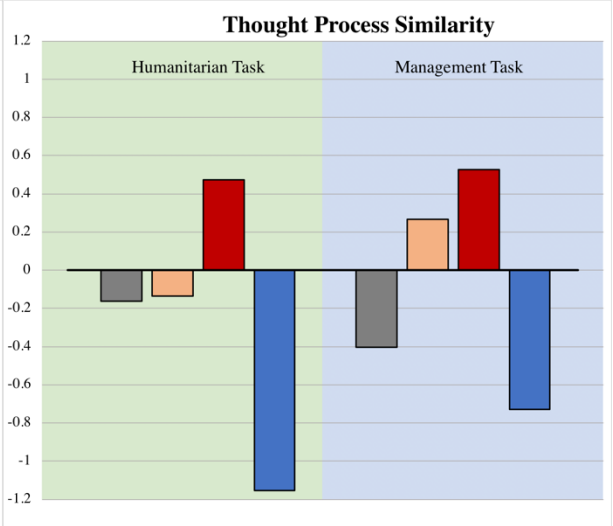


Figure 2.1: Utilization

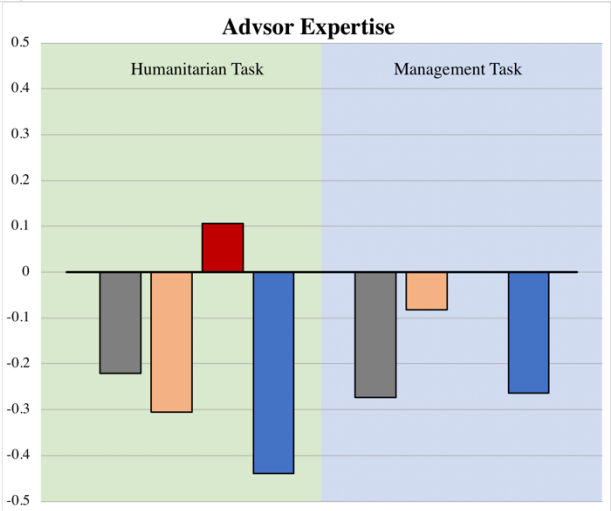






Figures 2.2-5: Graphing significant three-way effects. Difference between rating of first and second advisor. Negative values = lower rating for second advisor.

Automation Replaces Automation	**Human Replaces Human**
Human Replaces Automation	**Automation Replaces Human**



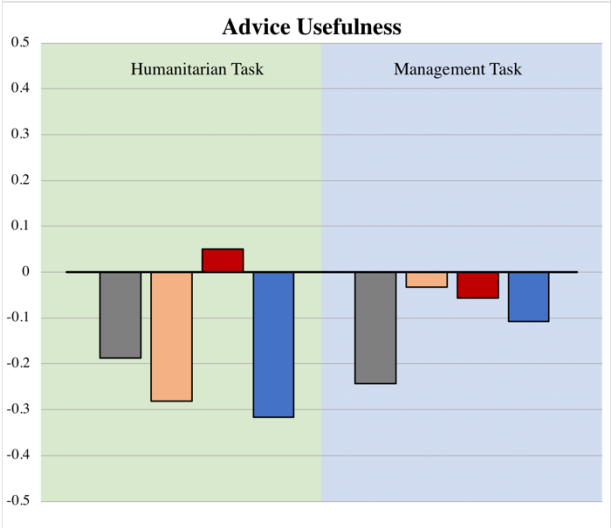
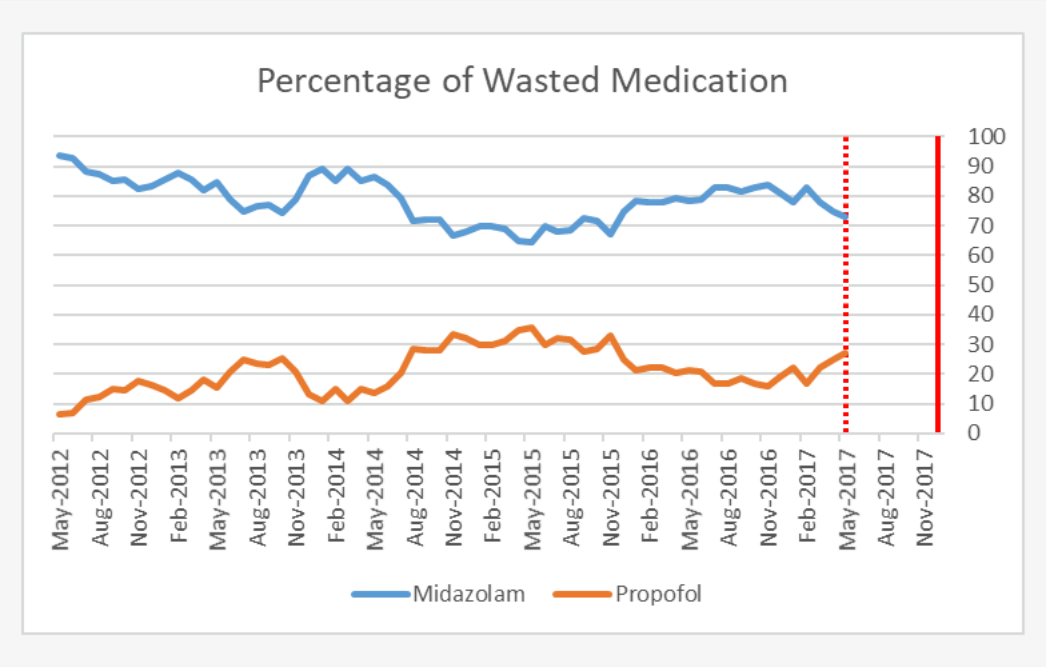


Figure 3.1 Screenshot of task

Propofol and Midazolam are common drugs used to sedate patients, but often too much is prepared for the surgery and not used, resulting in us having to throw away some very expensive materials! Management wants to order less of sedative drugs in December 2017 to reduce waste.



Please help us plan this reduction: what percent of wasted sedative drugs that you think would be from Midazolam (blue line) in December 2017?

Figure 3.2 Screenshot of feedback screen

The correct forecast was: 32.3

The advice was 33.3
Advice forecast percent error = **3.1%**

Your revised forecast was: 32
Your forecast percent error = **0.9%**

Your average percentage error across all forecasts is currently:

0.9%

This forecasting error costs the hospital: **$900**

Your combined errors thus far are estimated to have cost the hospital: **$900**