Paging Dr. JARVIS! Do people accept risk management advice from artificial intelligence in consequential decision contexts?

Connor Larkin[1], Caitlin Drummond[1], and Joseph Árvai[1,2]✉

[1]Erb Institute for Global Sustainable Enterprise
Stephen M. Ross School of Business, *and*
School for Environment and Sustainability
University of Michigan

[2]Decision Research
Eugene, OR

✉Corresponding Author
cplarkin@umich.edu

Abstract

Artificial intelligence (AI), a branch of computer science based upon algorithms that can analyze data and make decisions autonomously, is becoming increasingly prevalent in the technology that powers modern society. Relatively little research has examined how humans modify their judgments in response to their interactions with AI. Our research explores how people respond to different types of risk management advice received from AI *vs.* a human expert in two contexts where AI is commonly deployed: medicine and finance. Through online studies with representative samples of Americans, we first find that participants generally prefer to receive medical and financial risk management advice from humans over AI. In two follow-up studies, we presented participants with a hypothetical medical or financial risk and asked them to make an initial decision—to address the risk immediately or to wait for more information—and to rate their confidence in this decision. Next, participants were informed that either a human expert or AI had analyzed their case and recommended either immediate risk management action or a wait-and-see approach. Participant then made a final decision using the same response scale as before. We compared participants' initial and final decisions, examining the extent to which participants updated their decisions upon receiving their recommendation as a function of the recommendation itself and its source. We find that participants updated their decisions to a greater degree in response to recommendations from human experts as compared to AI, but the magnitude of this effect differed by context.

Key Words: artificial intelligence, decision-making, risk, medicine, finance

Acknowledgements: This research was supported by the U.S. National Science Foundation under award number SES 1728807 to Decision Research and the University of Michigan, and by the Erb Institute for Global Sustainable Enterprise at the University of Michigan.

1. Introduction

It was not long ago that artificial intelligence (AI) was largely confined to the realm of science fiction. Popular examples included the androids in the writing of Philip K. Dick; the defense computer, Skynet, which famously becomes "self-aware" in the *Terminator* films; and JARVIS (Just A Rather Very Intelligent System), Tony Stark's AI-based personal assistant in the Marvel Cinematic Universe. Today, AI is not only a reality, it has become a pervasive force—albeit largely in the background—in our daily lives. Advancements in machine learning and AI have led to the development and deployment of devices and software that can offer advice to human decision-makers across a wide range of contexts ranging from the mundane (e.g., music recommendations) to the consequential (e.g., medical diagnoses and treatment plans).

While specific definitions of AI can vary, AI is defined here as an advanced computer system that uses algorithms and statistical models to rapidly analyze large amounts of data in order to perform perceptual, cognitive, or conversational functions that are akin to speech or visual recognition, reasoning, and problem solving exhibited by humans (Longoni et al. 2019). We extend this definition of AI to include fully *autonomous* judgment and decision-making by a computer system; that is, the capability and freedom to make recommendations and choices without the need for regular human oversight.

As AI becomes even more intelligent and autonomous, it is poised to be among the most transformative technologies in human history. For example, people around the world already rely on AI—whether they recognize it, or not—to make recommendations about travel, entertainment, and romantic partners. Beyond these relatively mundane applications, AI is expanding rapidly to other, more consequential contexts. For example, AI is already capable of performing diagnostic healthcare functions—at scale—with a high level of accuracy (e.g., see Leachman and Merlino 2017) and at lower costs (e.g., see Esteva et al. 2017) when compared with humans working in a traditional healthcare setting. And, according to a 2017 analysis by the consulting firm Accenture[1], AI may save the healthcare sector $150 billion per year, as well as meet up to 20% of unmet healthcare demand in the U.S. alone, through increased technological efficiency.

Beyond healthcare, AI is making significant inroads in the domain of *FinTech* (finance and technology). Here, AI is being deployed at scale to analyze the risk tolerance and financial objectives of consumers in order to help them make goal-based financial decisions – or, in many instances, to make these decisions on their behalf (Jung et al. 2018). To this end, nearly $1 trillion of investment capital is already being

[1] See *Artificial Intelligence: Healthcare's New Nervous System* at https://www.accenture.com/_acnmedia/pdf-49/accenture-health-artificial-intelligence.pdf

managed by these so-called *robo-advisors*, or automated financial algorithms that allocate stocks within portfolios.

Overall, the technical barriers to widespread implementation of AI continue to fall more rapidly than many would have believed possible less than a decade ago. But, research into the willingness of individuals to accept advice from AI[2] has lagged behind technical developments in the field. Though not unexpected given the rapidity of AI deployment at scale, this gap could nevertheless have important implications for the development and rollout of AI technology. We make this claim because, regardless of technical achievements, adoption of AI will be delayed if individuals demonstrate an aversion to acceptance of algorithmic advice. There are countless examples of technologies—including nanotechnology (Siegrist et al. 2007), carbon capture and sequestration (L'Orange Seigo et al. 2014), and autonomous vehicles (Liu et al. 2019) among others—which have been developed to address health and environmental challenges, or to improve upon the societal *status-quo*, that have suffered from delayed or limited deployment because of elevated risk perceptions and a lack of public or consumer acceptance. In-depth exploration of the topic of human-AI interaction is therefore critical if we are to continue to harness the potential benefits of this technology.

Several past studies have examined the accuracy of *algorithmic* or *actuarial* decision-making, which describes the use of pre-set rules to analyze information and produce judgments. Generally speaking, this work has shown both that these types of algorithmic systems are more statistically accurate than individual "expert" judgments, and that experts tend to resist the use of such systems (Meehl 1954, Dawes et al. 1989). Possible sources of resistance to algorithmic decision-making include the misapplication of statistical probability to individual cases, and the cognitive availability bias of dramatic cases overshadowing less memorable but more likely outcomes (Dawes et al. 1989).

There is also the challenge of feeling like one is making a decision on their own when relying on advice from an algorithm. This feeling of isolation manifests because the social context associated with seeking a diagnosis or recommendation from a human expert is not present when interacting with a machine. In a healthcare context, for example, Promberger and Baron (2006) observed a preference for human versus algorithmic medical advice and concluded that participants did so because it created a sense of collaboration and shared responsibility regarding a treatment decision and its outcome.

---

[2] Previous research has examined the willingness of individuals to accept advice from a non-human system versus a human source, but has used inconsistent terminology, labeling the non-human system either as an "algorithm" (Dietvorst et al. 2015, Logg et al. 2019) or "AI" (Longoni et al. 2019). We choose to label the non-human system as "AI" in our research because we believe such a label is usually how algorithmic systems are and will be presented to the general public.

The rise to prominence of AI over the past five years has added to our understanding of algorithmic and actuarial decision-making with interesting and—at times—contradictory results. For example, Dietvorst et al. (2015) showed that participants appeared to demonstrate a resistance to algorithmic advice, even when human advice was demonstrably less accurate. Referring to this resistance as *algorithm aversion*, these results suggest that people appear to disproportionately 'punish' algorithms when they see such systems fail. That is, when observing both algorithmic and human-generated advice to be error-prone, and even in the knowledge that algorithmic advice is consistently less error-prone, participants lost confidence in AI systems more rapidly than they did in humans.

In contrast, recent research by Logg et al. (2019) on what they termed *algorithm appreciation* suggests a preference for advice from AI when compared to the same advice from a human. In this study, participants were first asked to make a prediction (e.g., forecasting the popularity of a song on the Billboard Hot 100 chart or forecasting the probability of a business meeting a production target), and then provided them with advice from either AI or a human source; participants could then choose to update their predictions based on the advice they received. Adopting the Judge Advisor System (JAS) paradigm (Sniezek and Buckley 1995), the degree to which participants updated their predictions as a function of the advice received was measured by comparing the participants' initial and final judgments. In contrast to prior work in AI, this study showed that participants were—on average—more receptive to advice from AI than they were to the same advice from a human across a variety of simple and relatively low-consequence forecasting judgments.

In today's reality, however, AI is being deployed to offer advice or to make decisions across a wide range of more immediately consequential contexts. As we note above, AI is being deployed to provide diagnostic advice to human and algorithmic decision-makers in the fields of healthcare (Yu et al. 2018), finance (Aw et al. 2019), environmental conservation (Gonzalez et al. 2016), weather forecasting (Ji et al. 2019), the day-to-day management of built infrastructure (Dounis 2010), and beyond. Thus, there is both an opportunity and a need to study the degree to which people will—or will not—use advice from AI to inform consequential judgments and decisions.

With this opportunity and need as backdrop, the research reported unfolded across two related studies[3] and was designed to address three related objectives:

First, in Study 1, we sought to examine to what extent people prefer to receive advice from AI *vs.* from a human expert in two consequential domains in which AI is commonly deployed: medicine and finance.

[3] Both studies were approved by the Health Sciences and Behavioral Sciences Institutional Review Board (protocol number HUM00162568 at the University of Michigan.

Based on the results of past literature pointing towards a generalized aversion to algorithmic advice, we hypothesized that people would generally prefer to receive advice from a human source relative to an algorithmic source.

Second, in Study 2, we explored the degree to which people would update their judgments in the direction of advice offered by AI or a human expert in two hypothetical but realistic risk management scenarios concerning medical treatments and financial planning. We hypothesized once again that participants would display a preference for advice from human experts, and we explored the degree to which contextual differences impact this preference.

Finally, the relationship between algorithm aversion (or appreciation) and a host of other variables (e.g., sociodemographic variables, value orientations, views about technology, etc.) has largely been unexplored. Thus, as part of Study 2, we conducted exploratory regression analyses to study the effect of these variables on the degree to which people were willing to update their judgments in the direction of advice offered by AI or a human expert.

2. Study 1: Stated Preferences for Advice from AI

2.1 Introduction

As we note above, researchers have been divided regarding people's willingness to accept advice from AI. Some studies (e.g., Longoni et al. 2019) have reported an aversion to accepting advice from AI out of concern that the algorithm will not take into account the unique characteristics or needs of the people seeking the advice. A substantial body of research in social psychology also points to the importance of self-uniqueness as a baseline condition of many individual judgments. For example, people routinely see themselves as unique in terms of their abilities, needs, values, and beliefs (Epstein 1990). In seeking and obtaining advice from a human expert, people are able to explain and contextualize their unique attributes in a way that is difficult to—at present—replicate with a computer. AI, on the other hand, is generally viewed by people as dehumanizing because computer systems, by the nature of their programming, are insensitive to this uniqueness and, as a result, aim for standardization at the expense of one's individuality (Haslam 2006).

Other research suggests that people may be averse to advice from AI because they are prone to quickly losing confidence in the technology after it has erred or given unhelpful advice. We know from research on risk perceptions that singular trust-destroying events loom larger in the mind than do an additive series of trust-building events (Slovic 1993). Recent research on AI suggests that this tendency to overreact to outcomes that may erode trust and confidence is even stronger when we consider the behavior of

autonomous computer systems *vs.* the same behavior by human providers of advice (Dietvorst et al. 2015).

However, other studies (e.g., Logg et al. 2019) suggest that people are apt to more readily and frequently rely on advice from AI when making judgments. This may especially be the case in contexts that are *algorithmically appropriate* in that people already rely on AI for advice, or in cases where AI is already being used by large numbers of people. A decision-maker's confidence in the advice received is also thought to matter. That is, in contexts where a decision-maker is not confident to proceed without external advice, they may be more willing to receive it from AI so long as the advice in question is of high quality.

Given these conflicting findings, Study 1 was designed to explore a relatively straightforward question: when considering in consequential risk management contexts—namely healthcare and finance—do people prefer to receive advice from AI or a human expert? In addition, if advice from AI or a human expert was given, would they express willingness to follow it, and would they be confident in it?

2.2 Methods

2.2.1 Participants

We recruited adults (over the age of 18) currently residing in the United States to participate in two variants of Study 1: a healthcare (Study 1A) and a finance (Study 1B) variant. Participants were recruited from online panels curated by Qualtrics® and were randomly drawn from a representative probability sample of active panel members. A power analysis conducted in G*Power (Faul et al. 2009) determined that, in order to obtain 80% power with the ability to detect a small effect size (Cohen's $d = 0.3$) with our planned paired-samples *t*-tests, we needed to recruit 92 participants for both Studies 1A and 1B. We oversampled for both of Studies 1A and 1B because we anticipated that several participants would fail our data quality assurances.

A total of 110 participants were recruited to participate in Study 1A and another 126 participants were recruited to participate in Study 1B. A total of 5 participants in Study 1A, and 21 participants in Study 1B, were removed from the final samples because they failed an instructed-choice attention check. Of the 105 participants that remained in Study 1A, 50.5% were female, the average age was 63.5 years (SD = 15.4), and 55.2% of participants had an education level corresponding with a bachelor's degree or higher. Of the 105 participants that remained in Study 1B, 48.6% were female, the average age was 50.6 years (SD = 15.5), and 60% of participants had an education level corresponding with a bachelor's degree or higher.

2.2.2 Design

After providing their informed consent, participants in Study 1A and 1B (see the Supplemental Materials section for the complete instruments) were first asked to read a preamble describing AI, which informed them that AI was defined in the study as an advanced computer system that could quickly analyze large amounts of data and makes recommendations or decisions without human input or supervision. Next, participants were asked to assume that in the near future, they will be faced with an important decision that could have a significant impact on their *personal health* (Study 1A) or their *financial health* (Study 1B). Participants in Study 1A and 1B were then asked to respond to the same five questions about receiving advice in their assigned context.

The first question asked participants if they would prefer to receive recommendations from a human expert or AI in their assigned context; responses were collected on 7-point Likert scales where 1 = *I strongly prefer a human expert* and 7 = *I strongly prefer AI* (midpoint = *I am indifferent*).

Question 2 asked participants about the degree to which they would follow a recommendation offered by a human expert, and Question 3 asked about the degree to which they would follow a recommendation offered by AI in their assigned context; responses to both questions were collected on 7-point Likert scales where 1 = *Definitely not* and 7 = *Definitely yes* (midpoint = *maybe*).

Question 4 asked how confident participants would be in a recommendation offered by a human expert, and Question 5 asked how confident participants would be in a recommendation offered by AI; responses to these questions were also collected on 7-point Likert scales where 1 = *Not at all confident* and 7 = *Very confident* (midpoint = *Somewhat confident*).

Before the close of the survey, participants reported their gender, age, education level, political ideology, and religiosity. Participants also answered the technological optimism and technological dependence subscales of the Technology Adoption Propensity Index (Ratchford and Barnhart 2012) and responded to the a series of questions aimed at ascertaining the degree to which they ascribe to egoistic, altruistic, and biospheric value orientations (de Groot and Steg 2007). Finally, participants were also asked to report the extent to which they agreed with the following two statements: "I generally trust medical professionals" and "I generally trust financial professionals." Responses to these questions were collected using 7-point Likert scales from *1 = Strongly agree* to 7 = *Strongly disagree* (midpoint = *Neither agree nor disagree)*.

2.2.3 Analysis

We compared responses to the five questions asked in Study 1A and 1B using paired-samples *t*-tests (for within-study comparisons) and Welch two-sample *t*-tests (for across-study comparisons).

2.3 Results and Discussion

In both the healthcare and finance contexts, participants indicated a strong preference for advice from human experts over AI (Table 1). Though participants in the finance context appeared to be less firm in their preference for advice from a human expert (*vs.* participants in the healthcare condition), this difference was not significant at a Bonferroni corrected alpha of 0.01 (adjusted for the 5 across-study comparisons in Table 1).

Within both the healthcare and finance contexts, participants reported being more likely to follow a recommendation from a human expert over a recommendation from AI, $t_{\text{Healthcare}} = 12.36$, $p < 0.001$; $t_{\text{Finance}} = 7.96$, $p < 0.001$. Across contexts, participants were more willing to follow a recommendation from a human expert for a healthcare decision than a financial decision, and equally likely to follow a recommendation from an AI in both contexts (Table 1).

Within both the healthcare and finance contexts, participants reported being more confident in a recommendation from a human expert when compared with a recommendation from AI, $t_{\text{Healthcare}} = 12.01$, $p < 0.001$; $t_{\text{Finance}} = 7.73$, $p < 0.001$. Across contexts, participants were more willing to follow a recommendation from a human expert for a healthcare decision than a financial decision, and equally likely to follow a recommendation from an AI in both contexts (Table 1).

Table 1. Summary data for Studies 1A and 1B.

	Healthcare		**Finance**			
	$\bar{x}^*$	SD	$\bar{x}^*$	SD	*t*	*p*
Q1. Preference (Human Expert *vs.* AI)	2.03	1.58	2.49	1.81	-1.99	0.048
Q2. Likely to Follow – Human Expert	5.77	1.35	5.26	1.37	2.75	0.007
Q3. Likely to Follow – AI	3.32	1.51	3.61	1.66	-1.31	0.19
Q4. Confidence in Human Expert	5.66	1.25	5.18	1.24	2.77	0.006
Q5. Confidence in AI	3.25	1.56	3.67	1.67	-1.88	0.062

*Preferences were elicited on a 7-point Likert scale where 1 = a strong preference for a human expert and 7 = a strong preference for AI.

Tables S1 and S2 of the Supplemental Materials report exploratory linear regressions predicting the five dependent measures of Table 1 separately for the two contexts, as a function of demographic covariates. Across contexts, participants with greater trust in professionals report being more willing to follow and have confidence in a recommendation from a human expert; participants with greater technological optimism and egoism report being more likely to follow and have confidence in a recommendation from an AI. Detailed results are in the Supplemental Materials.

These results are consistent with prior research (Meehl 1954, Dawes et al. 1989, Longoni et al. 2019) that points to algorithm aversion on the part of human decision-makers who may receive advice from a human expert or AI. Participants in both the healthcare and finance contexts indicated a strong preference for receiving recommendations from human experts over AI. Across contexts, however, the preference for a human expert appeared to be somewhat stronger in the healthcare context: participants reported being more likely to follow a recommendation from a human expert, and placing more confidence in a recommendation from a human expert, in the healthcare as compared to the finance context.

A limitation of Study 1 is that it, like other studies, focuses on the question of decision-makers' receptivity to getting advice from AI using a stated preference research approach. We know from several studies across different disciplines that attitudes and preferences expressed in the absence of the need to make a choice—i.e., revealed preference—are prone to potentially biased reporting by participants (Samuelson 1938, Fishbein and Ajzen 1975). Thus, Study 2 looked more deeply into the question of whether people would accept and follow advice from AI in the same two consequential risk management contexts: healthcare and finance.

3. Study 2: Revealed Preferences for Advice from AI

3.1 Introduction

In addition to the challenges posed by stated preference studies, prior research on algorithm aversion has focused largely on one's willingness to accept and follow advice from AI in rather inconsequential (e.g., forecasting the popularity of a song) or contextually distal (e.g., forecasting political events) situations. We, therefore, questioned whether the observed discrepancies regarding algorithm aversion and algorithm appreciation may be caused by the fact that, in many instances, people are not being asked about their willingness to accept advice from AI in situations that are likely to be both personally consequential and algorithmically appropriate (meaning that AI as a source of advice is already being deployed for these contexts in the real-world and at scale).

Thus, Study 2 was designed to place participants in either a healthcare or finance context in which they would receive advice from either an established human expert or AI. We hypothesized that participants would, as in Study 1, exhibit strong algorithm aversion. However, we also expected that in the more temporally distal, and existentially less threatening finance context, people would exhibit lower levels of algorithm aversion.

3.2 Methods

3.2.1 Participants

We recruited adults (over the age of 18) currently residing in the United States to participate in one of two experiment, examining decision-making in either a healthcare or a finance context. A power analysis conducted in G*Power indicated that in order to have 80% power to detect a small effect size of $f = 0.15$, for a planned 4-group, one-way ANOVA, we would need to recruit 492 participants per context. Anticipating that some participants would fail data quality assurance measures, we aimed to recruit 500 participants for each context. Participants were recruited from online panels curated by Qualtrics® and were randomly drawn from a representative probability sample of active panel members. Anticipating that some participants would fail our data quality assurances, we oversampled for each context, ultimately recruiting 600 participants for the healthcare context and 574 participants for the finance context. Eighty-three participants from the healthcare context and 64 participants from the finance context were excluded due to failing an instructed-choice attention check. Of the final sample of 517 participants in the healthcare context, 56.7% were female, the average age was 51.9 years (SD = 17.1) and 49.9% of participants had a bachelor's degree or higher level of education. Of the final sample of 508 participants in the finance context, 47.8% were female, the average age was 50.2 years (SD = 16.5) and 50.4% of participants had a bachelor's degree or higher level of education. One-way ANOVAs, treating each of the four conditions separately, revealed no significant differences in gender, age, and education across conditions (healthcare survey: all p's > 0.15; finance survey: all p's > 0.12).

3.2.2 Design

Within each of the two contexts, participants were randomly assigned to condition using a 2 (recommendation source: human expert vs. AI) by 2 (recommendation: immediate action vs. delayed action) between-subjects design.

After providing informed consent, all participants read the same preamble, informing them that they would be receiving a recommendation from either a human expert or AI, and defining AI as in Study 1. Then, participants were asked to imagine themselves in a risky and uncertain situation:

In the healthcare context, participants were told to imagine that they were diagnosed with a cancerous tumor that could metastasize and be fatal, or remain static and benign. The probability of metastasis was presented as uncertain. Participants were next informed that they could either decide—on their own—to have immediate surgery (accompanied by the risk of post-operative infection) or wait-and-see for one year (accompanied by the risk of cancer metastasis in the interim). See the Supplemental Materials section for the complete scenario.

In the finance context, participants were told to imagine that they owned a portfolio of investments dominated by companies that produce oil and gas, and were informed of the possibility that companies

that produce energy from renewables may soon outperform them. The timing of this inflection point in the market was presented as uncertain. Participants were next informed that they could either decide—on their own—to immediately rebalance their portfolios in the direction of companies that produce energy from renewables (accompanied by the risk of losing money of oil and gas companies continue to perform well) or wait-and-see for one year (accompanied by the risk of losing money if renewables companies surge in their profitability). See the Supplemental Materials section for the complete scenario.

In both contexts, participants were next asked to make an initial binary judgment—to take immediate action or to wait-and-see—and then rate their confidence in this judgment on a sliding scale from 50% (just guessing) to 100% (completely certain).

After making this initial judgment, participants in both contexts were randomly assigned to one of four treatments:

In the healthcare context, participants were informed that either their human physician or a medical AI (working without direct human supervision) had reviewed their case and recommended either immediate action or a wait-and-see approach for one year.

In the finance context, participants were informed that either their human financial advisor or a financial AI (working without direct human supervision) had reviewed their portfolio and recommended immediate action or a wait-and-see approach for one year.

Participants were then asked to make a final binary judgment—to take immediate action or to wait-and-see—and then rate their confidence in this decision on a sliding scale from 50% (just guessing) to 100% (completely certain).

After this final judgment was made, we asked participants to answer the same demographic and attitudinal questions posed at the end of Study 1.

3.2.3 Analysis

In order to analyze the judgments of participants as a function of the kind of advice they received, we constructed an dependent (updating) variable, U, which was the extent to which participants updated their initial judgments in the direction of the recommendation they received, based on prior research (Logg et al. 2019). To do so, we converted initial and final judgments (to delay or take immediate action) and their accompanying confidence rating into a 0 to 100 scale. Here, 0 represented complete confidence in the decision to delay action and 100 represented complete confidence in the decision to take immediate action; 50 represented all of the "just guessing" responses regardless of whether a participant chose to take immediate action, or to delay action.

Next, we treated the advice received from the human expert or AI as certain, corresponding to a scale rating of either 0 (complete confidence in the recommendation to delay action) or 100 (complete confidence in the recommendation to take immediate action). We then calculated the updating variable for each participant as follows:

$$U = \frac{\text{Recommendation} - \text{Final Judgment}}{\text{Recommendaton} - \text{Initial Judgment}}$$

For example, assume a participant initially selected immediate action in the healthcare condition but indicated 50% confidence in this judgment. Then, after receiving a recommendation from AI to take immediate action the participant once again selected immediate action but now with 75% confidence in this judgment. Thus, *U* for this participant in their assigned combination of condition and treatment was:

$$U = \frac{100 - 75}{100 - 50} = 0.5$$

This value of $U = 0.5$ was interpreted as the participant updating their judgment by 50% in the direction of the recommendation given.

Our study design (and the calculation of our dependent measure, U) also gave participants the option to update in the opposite direction of the advice. For example, a participant who decided to take immediate action with 75% confidence, was then provided with a recommendation to take immediate action, and then decided to wait-and-see with 75% confidence, would be updating in the *opposite* direction of the advice. Updating in the opposite direction of the advice corresponds with a negative value for U. In prior research (Logg et al. 2019), these negative values have been Winsorized and converted to a value of $U = 0$. We adopted the same approach in this study.

Updating in the opposite direction of the advice may occur due to participant inattention, difficulty recalling initial responses, an aversion to the advice, or a rethinking of one's initial stance. Initial analyses of our data revealed that 22% of participants ($n = 114$) in the healthcare context and 20% of participants ($n = 100$) in the finance context updated in the opposite direction of the advice. If these responses stemmed from aversion to the advice or rethinking of initial responses, then Winsorizing these responses by setting them equal to 0 would treat such responses as having placed no weight whatsoever on the advice. On the other hand, if such responses stemmed from inattentive participants, then excluding these inattentive responses would be appropriate as they provide no informational value about updating from advice. Since we could not tell the difference between those two explanations based on the data we collected, we report the results with Winsorized data below. However, we report the results *excluding*

responses with a negative updating variable, *U*, in the Supplemental Materials section. We discuss the differences between these data sets in the Supplemental Materials.

Additionally, in the healthcare context, 41 participants received a recommendation that was equivalent to their initial judgment, and were thus excluded from the data because of our inability to examine updating in response to the advice. Likewise, in the finance, 19 participants received a recommendation that was equivalent to their initial judgment; these participants were also excluded from our analysis.

We used a two-way ANOVA to examine the extent to which participants updated in the direction of the recommendation received as a function of the type of recommendation (delayed or immediate action), its source (human expert or AI), and their interaction. We also conducted exploratory linear regression to predict updating as a function of recommendation, source of recommendation, and the covariates and demographics.

3.3 Results and Discussion

In the healthcare context, a two-way ANOVA showed a significant main effect of recommendation source, such that individuals updated their judgments in the direction of a recommendation to a greater degree when it came from a human expert *vs.* AI ($F_{(1,472)} = 80.51$, $p < .001$, $\eta^2 = 0.15$). We observed no significant main effect of recommendation (e.g. to delay action or to take immediate action) ($F_{(1,472)} = 0.82$, $p > 0.05$, $\eta^2 = 0.001$), and no interaction between the recommendation and its source ($F_{(1,472)} = 0.11$, $p > 0.05$, $\eta^2 = 0.000$); see Table 2.

Table 2. Updating by condition.

		Healthcare			Finance				
Source	Recommendation	*n*	$\bar{x}$	SD	*n*	$\bar{x}$	SD	*t*	*p*
Human Expert	Immediate Action	108	0.45	0.36	128	0.29	0.31	3.87	<0.001
Human Expert	Delayed Action	126	0.44	0.35	123	0.28	0.31	3.75	<0.001
AI	Immediate Action	107	0.20	0.30	123	0.17	0.26	0.91	0.37
AI	Delayed Action	135	0.17	0.27	117	0.23	0.29	-1.64	0.10

In the finance context, we also found a significant main effect of recommendation source, such that individuals updated their judgments in the direction of a recommendation to a greater degree when it came from a human expert *vs.* AI ($F_{(1,487)} = 10.7$, $p < 0.001$, $\eta^2 = 0.02$). Once again, we observed no significant main effect of recommendation ($F_{(1,487)} = 0.9$, $p > 0.05$, $\eta^2 = 0.002$), and no interaction between the recommendation and its source ($F_{(1,487)} = 1.3$, $p > 0.05$, $\eta^2 = 0.001$); see Table 2.

Participants updated their initial judgment to a greater degree when receiving a recommendation from a human expert in the healthcare context than they did in the finance context; we observed this pattern for

both recommendations (immediate action or wait-and-see; Table 2). These differences were significant at a Bonferroni-corrected alpha of 0.0125. In contrast, participants updated their initial judgment to the same degree when receiving a recommendation from AI in both the healthcare and finance contexts (Table 2).

We conducted exploratory linear regressions predicting updating as a function of condition and covariates. Table 3 shows that, in the healthcare context, more religious participants and those with greater technological dependence updated less in response to the advice; those with greater trust in medical professionals and greater biospherism updated more. In the finance context, participants with greater technological optimism and those with greater trust in financial professionals updated more in response to the advice.

Table 3. Updating from the recommendation as a function of condition and covariates.

	Healthcare			Finance		
	β	95% CI	η^2	β	95% CI	η^2
Source (binary; 1 = AI)	-0.285***	-0.36, -0.21	0.17	-0.04	-0.11, 0.034	0.02
Recommendation (binary; 1 = immediate action)	0.02	-0.059, 0.099	0.00	0.017	-0.055, 0.089	0.00
Source x Recommendation	0.022	-0.09, 0.13	0.00	-0.073	-0.17, 0.03	0.00
Male	0.045	-0.014, 0.1	0.00	-0.033	-0.086, 0.02	0.00
Age	-0.0003	-0.002, 0.002	0.00	0.0001	-0.0017, 0.0019	0.00
Education	-0.019	-0.039, 0.0012	0.01	0.012	-0.007, 0.03	0.00
Political Liberalism	-0.012	-0.027, 0.0024	0.01	-0.01	-0.023, 0.003	0.01
Religiosity	-0.016*	-0.031, -0.002	0.01	-0.012	-0.025, 0.0015	0.02
Technological Optimism	0.026	-0.003, 0.057	0.01	0.038**	0.013, 0.064	0.00
Technological Dependence	-0.021*	-0.042, -0.0001	0.01	-0.005	-0.025, 0.015	0.02
Trust in Human Experts	0.060***	0.035, 0.085	0.05	0.028**	0.0087, 0.047	0.00
Egoism	0.001	-0.018, 0.021	0.00	-0.005	-0.023, 0.014	0.00
Biospherism	0.026*	0.001, 0.051	0.01	0.002	-0.023, 0.028	0.00
Altruism	-0.006	-0.036, 0.024	0.00	0.012	-0.017, 0.041	0.00
Constant	0.183	-0.062, 0.43		-0.049	-0.28, 0.19	
n	476			491		
R^2	0.238			0.086		
Adjusted R^2	0.215			0.059		
Residual Standard Error	0.305			0.286		
F	10.306*** (df= 14, 461)			3.213*** (df = 14, 476)		

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

4. General Discussion

Artificial intelligence (AI) is being used today to provide advice to decision-makers across a wide variety of contexts, but relatively little research has examined uptake of advice proffered by AI. What research has been conducted has found conflicting evidence as to whether humans prefer advice from human experts or AI (Dietvorst et al. 2015, Logg et al. 2019, Longoni et al. 2019). We investigated preferences for advice from human experts and AI in two contexts in which AI has been commonly and consequentially deployed: healthcare and finance. In Study 1, we found that participants generally preferred advice from a human expert as compared to AI in both contexts, and indicated a greater

willingness to follow and have confidence in advice from a human expert. However, participants more strongly preferred advice from a human expert in the healthcare context as compared to the finance context; there were no differences across context with respect to preferences for advice from AI.

In Study 2, we presented participants with a hypothetical medical or financial decision and examined the degree to which they updated their initial judgments after receiving advice from a human expert or an AI. In both scenarios, participants updated their judgments further in the direction of advice received when it came from a human expert as compared to an AI, regardless of the nature of advice (i.e., to take immediate action or to wait). Participants updated more in response to the human expert in the healthcare scenario as compared to the finance scenario; participants updated similarly in response to the AI in both scenarios. Across both studies, we find A) an overall preference for advice from human experts as compared to AI, B) a stronger preference for human experts in a healthcare vs a financial context, and C) similar evaluations of AI advice in a healthcare and financial context.

Overall, our findings reflect a preference for advice from human experts as compared to AI, with differences in advice-taking across healthcare and finance contexts. At first glance, our results are more consistent with previous findings of algorithm aversion (Dietvorst et al. 2015, Dietvorst et al. 2016, Longoni et al. 2019) than algorithm appreciation (Logg et al. 2019). However, these previous studies differed from our study, and each other, in several important ways that complicate our conclusions.

Previous findings of algorithm aversion (Dietvorst et al. 2015, Dietvorst et al. 2016) have typically asked participants to choose between their own judgments and those of an algorithm; in contrast, studies showing algorithm appreciation typically examine the weight placed on advice from either a human (another participant) or an AI advisor (Logg et al. 2019). Indeed, this latter study showed that algorithm appreciation decreases when choosing between the estimates of an AI and one's own estimates, as compared to when choosing between the estimates of an AI and another participant's estimates. We examined the weight placed on advice from either humans or AI, but unlike in Logg et al. (2019), the human advice in our study came from human experts rather than other participants.

Another important difference between our work and previous work is that tasks used in these previous studies—estimating a person's weight from a photograph, the popularity of a song, the attractiveness of a person from a written description, the success of an MBA applicant, the test scores of a student, the ranking of US states by number of departing airline passengers (Dietvorst et al. 2015, Dietvorst et al. 2016, Logg et al. 2019)—are about outcomes external to one's own, while our studies asked participants to consider hypothetical scenarios involving oneself. Our study is most similar to research by Longoni et al. (2019), who compared receptivity to human vs automated healthcare providers in the medical domain

by asking participants to consider hypothetical scenarios involving their own personal health, and who found a preference for human healthcare providers over automated providers. Our findings in the healthcare domain reflect those of Longoni et al. (2019), while our comparisons across the healthcare and finance domains suggest potentially important differences in preferences for algorithms vs. human experts across domains.

Longoni et al. (2019) suggest that a preference for human-generated advice in a medical context may stem from "uniqueness neglect", or the perceived inability of AI to properly account for the unique characteristics of individuals. This concept harkens back to the "broken leg" hypothesis (Meehl 1954), which points out that people may resist algorithmic decision-making because they believe that the algorithm may be unable to account for relevant features that lie outside of the decision-making domain. For example, an algorithm that predicts whether or not the individual will go to the cinema on a Friday night, based on past behavior of the individual, may have difficulty accounting for a sudden leg injury, whereas such a factor would be patently obvious to a human judge. Longoni et al. (2019), find that participants who perceive their case to be more unique are more likely to resist AI-generated medical advice, and finds that resistance is lower when the AI-generated medical advice is described as personalized, or is given as a supplement to a human doctor's advice.

In Study 2, we observed similar levels of updating from AI-generated advice across the healthcare and finance domains, which could suggest similar importance of "uniqueness neglect" across these domains as a driver of algorithm aversion. An individual's financial portfolios and goals might be perceived as unique in the same way that one's health status is. In contrast, we observe greater updating from human-generated advice in the healthcare versus the financial scenario. These differences may stem from differential amounts of trust in human doctors as opposed to financial providers; indeed, we find participants report significantly greater levels of trust in medical professionals than financial professionals ($\bar{x}$ = 5.40 vs. $\bar{x}$ = 4.42, $t = 11.6$, $p < 0.001$).

Taken together, these results suggest that future work on algorithm aversion or appreciation should take into account not only the factors influencing trust in or aversion to algorithms, but also the factors influencing the extent to which human expertise is trusted in the task domain, which may differ by domain. Algorithms may be more likely to be seen as viable alternatives to human judgment in domains where human judgment is less trusted or in which the conditions for human expertise are less likely to be met (Kahneman and Klein 2009).

Likewise, future experimental research may also present respondents with information that contextualizes AI programming. Algorithms require initial input and programming from humans; information about AI

that makes clear to users *who* was behind its creation (e.g., programming input from people with domain-specific expertise—i.e., medical or financial experts in the case of our research—who also understand the user's personal objectives) may counteract the apparent skepticism of AI that we observed in our study. Similarly, future work should further explore how willingness to trust AI in risk management decisions is related to other dimensions of (un)trustworthiness such as concerns about racial bias (e.g., Caliskan et al. 2017, Gianfrancesco et al. 2018).

Our findings are subject to several limitations. As mentioned, we observed a stronger preference for human experts in the healthcare context compared to the finance context, suggesting that preferences for advice from human experts vs. AI may be context-dependent. Future work should test the generalizability of these findings to additional contexts in which AI is deployed. In addition, we examined preferences and decision-making in hypothetical scenarios that may lack key context present in real-world decision-making involving AI advice. Future work should continue to examine acceptance of AI advice in real-world decision-making, and should expand to examine not only individual decision-makers but also policymakers and business leaders whose decisions impact many.

AI can perform tasks that are typically thought of as being within the exclusive domain of human experts—such as medical diagnoses and treatment recommendations, and personal financial planning—with high accuracy and low costs (e.g., see Leachman and Merlino 2017), potentially leading to substantial gains in welfare. As AI continues to shape everyday life and decision-making, however, our research suggests it may have significant barriers to overcome before its advice is seen as a trustworthy as that of a human expert.

5. References

Aw, E. N. W., J. Jiang, and J. Q. Jiang. 2019. Rise of the machines: Factor investing with artificial neural networks and the cross–section of expected stock returns. The Journal of Investing **29**:6.

Caliskan, A., J. J. Bryson, and A. Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. Science **356**:183.

Dawes, R. M., D. Faust, and P. E. Meehl. 1989. Clinical versus actuarial judgment. Science **243**:1668.

de Groot, J. I. M., and L. Steg. 2007. Value orientations to explain beliefs related to environmental significant behavior: How to measure egoistic, altruistic, and biospheric value orientations. Environment and Behavior **40**:330-354.

Dietvorst, B. J., J. P. Simmons, and C. Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General **144**:114–126.

Dietvorst, B. J., J. P. Simmons, and C. Massey. 2016. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Management Science **64**:1155-1170.

Dounis, A. I. 2010. Artificial intelligence for energy conservation in buildings. Advances in Building Energy Research **4**:267-299.

Epstein, S. 1990. Cognitive-experiential self-theory. Pages 165-192 *in* L. A. Pervin, editor. Handbook of Personality: Theory and Research. Guilford Press, New York, NY.

Esteva, A., B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**:115-118.

Faul, F., E. Erdfelder, A. Buchner, and A.-G. Lang. 2009. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods **41**:1149-1160.

Fishbein, M., and I. Ajzen. 1975. Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research. Addison-Wesley, Reading, MA.

Gianfrancesco, M. A., S. Tamang, J. Yazdany, and G. Schmajuk. 2018. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. JAMA internal medicine **178**:1544-1547.

Gonzalez, L. F., G. A. Montes, E. Puig, S. Johnson, K. Mengersen, and K. J. Gaston. 2016. Unmanned aerial vehicles (UAVs) and artificial intelligence revolutionizing wildlife monitoring and conservation. . Sensors **16**:97.

Haslam, N. 2006. Dehumanization: An integrative review. Personality and Social Psychology Review **10**:252-264.

Ji, L., Z. Wang, M. Chen, S. Fan, Y. Wang, and Z. Shen. 2019. How much can ai techniques improve surface air temperature forecast? A report from ai challenger 2018 global weather forecast contest. Journal of Meteorological Research **33**:989-992.

Jung, D., V. Dorner, C. Weinhardt, and H. Pusmaz. 2018. Designing a robo-advisor for risk-averse, low-budget consumers. Electronic Markets **28**:367-380.

Kahneman, D., and G. Klein. 2009. Conditions for intuitive expertise: A failure to disagree. American Psychologist **64**:515-526.

L'Orange Seigo, S., J. Arvai, S. Dohle, and M. Siegrist. 2014. Predictors of risk and benefit perception of carbon capture and storage (CCS) in regions with different stages of deployment. International Journal of Greenhouse Gas Control **25**:23-32.

Leachman, S., and G. Merlino. 2017. Medicine: The final frontier in cancer diagnosis. Nature **542**.

Liu, P., R. Yang, and Z. Xu. 2019. Public Acceptance of Fully Automated Driving: Effects of Social Trust and Risk/Benefit Perceptions. Risk Analysis **39**:326-341.

Logg, J. M., J. A. Minson, and D. A. Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes **151**:90-103.

Longoni, C., A. Bonezzi, and C. K. Morewedge. 2019. Resistance to medical artificial intelligence. Journal of Consumer Research **46**:629-650.

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

Promberger, M., and J. Baron. 2006. Do patients trust computers? Journal of Behavioral Decision Making **19**:455-468.

Ratchford, M., and M. Barnhart. 2012. Development and validation of the technology adoption propensity (TAP) index. Journal of Business Research **65**:1209-1215.

Samuelson, P. A. 1938. A note on the pure theory of consumer's behaviour. Econometrica **5**:61-71.

Siegrist, M., M.-E. Cousin, H. Kastenholz, and A. Wiek. 2007. Public acceptance of nanotechnology foods and food packaging: The influence of affect and trust. Appetite **49**:459-466.

Slovic, P. 1993. Perceived risk, trust, and democracy. Risk Analysis **13**:675-683.

Sniezek, J. A., and T. Buckley. 1995. Cueing and cognitive conflict in Judge-Advisor decision making. Organizational Behavior and Human Decision Processes **62**:159–174.

Yu, K.-H., A. L. Beam, and I. S. Kohane. 2018. Artificial intelligence in healthcare. Nature Biomedical Engineering **2**:719-731.

Supplemental Materials

Study 2: Supplementary Methods and Results:

Twenty-two percent of participants ($N = 114$) in the healthcare context and 20% of participants ($N = 100$) in the financial context) updated in the opposite direction of the advice. In the main text, we report the results with Winsorized data; here, we report the results excluding responses with negative updating values.

In the healthcare context, omitting rather than Winsorizing data corresponding to updating in the opposite direction of the advice left us with 402 participants. Of these participants, 44.8% were male, mean age was 51.9 ($SD = 17.2$), and 50% had a Bachelor's degree or higher.

In the finance context, omitting rather than Winsorizing data corresponding to updating in the opposite direction of the advice left us with 402 participants. Of these participants, 50.4% were male, mean age was 51.2 ($SD = 16.1$), and 52.3% had a Bachelor's degree or higher.

In the healthcare context, we found that individuals update more when they receive advice from a human source relative to AI ($F(1,398) = 62.9$, $p < .001$). We also found that individuals update more when they receive advice to take immediate action relative to advice to delay action ($F(1,398) = 4.2$, $p = 0.04$). We did not find a significant interaction between the advice and the source of advice ($F(1,398) = 0.8$, $p = 0.36$).

In the financial context, we also found that individuals update more when they receive advice from a human source relative to AI ($F(1,405) = 5.8$, $p = 0.02$). We did not find a significant difference in updating to take immediate action versus delaying action ($F(1,405) = 0.15$, $p = 0.7$), and we also did not find a significant interaction between the advice and the source of advice ($F(1,405) = 0.2$, $p = 0.66$).

Table S1. Predictors of Preferences for Advice from Humans vs. AI in Healthcare Context (Study 1A)

	Preference for Human Expert (vs AI)	Willingness to Follow Advice from a Human Expert	Willingness to Follow Advice from an AI	Confidence in Advice from a Human Expert	Confidence in Advice from an AI
Male	-0.009	-0.284	0.024	0.013	-0.023
Age	-0.005	0.035***	0.002	0.018*	0.0004
Education	-0.03	0.006	0.05	0.078	0.067
Political Liberalism	-0.068	-0.036	-0.031	0.015	0.009
Religiosity	-0.174*	-0.012	-0.131	0.023	-0.1
Technological Optimism	0.340*	0.087	0.319*	-0.036	0.307
Technological Dependence	0.111	-0.09	0.06	-0.1	0.117
Trust in Health Professionals	-0.216	0.149	-0.097	0.325**	-0.063
Egoism	0.125	0.102	0.19	0.032	0.227*
Biospherism	0.269	-0.004	0.091	-0.165	0.13
Altruism	-0.282	0.178	-0.032	0.232	-0.129
Constant	2.413	1.742	1.491	2.15	0.963
N	105	105	105	105	105
R^2	0.209	0.255	0.14	0.266	0.157
Adjusted R^2	0.116	0.167	0.038	0.179	0.057
Residual Standard Error	1.49	1.229	1.48	1.137	1.51
$F(11, 93)$	2.235*	2.889**	1.376	3.063**	1.575

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

Table S2. Predictors of Preferences for Advice from Humans vs. AI in Finance Context (Study 1B)

	Preference for Human Expert (vs AI)	Willingness to Follow Advice from a Human Expert	Willingness to Follow Advice from an AI	Confidence in Advice from a Human Expert	Confidence in Advice from an AI
Male	0.455	-0.069	0.266	-0.135	0.264
Age	-0.022	0.009	-0.017	-0.008	-0.011
Education	-0.189	0.063	0.026	0.045	0.042
Political Liberalism	0.019	-0.08	-0.044	-0.056	-0.095
Religiosity	-0.093	-0.029	0.009	0.143*	0.045
Technological Optimism	0.234	-0.154	0.415*	-0.12	0.506**
Technological Dependence	-0.053	-0.021	-0.07	0.009	0.024
Trust in Financial Professionals	0.014	0.463***	-0.085	0.326**	-0.102
Egoism	0.260*	0.058	0.235*	0.063	0.223
Biospherism	0.224	0.101	0.1	-0.114	0.24
Altruism	-0.276	0.163	-0.018	0.301*	-0.084
Constant	2.99	2.102	1.692	2.861**	0.35
N	105	105	105	105	105
R^2	0.27	0.258	0.251	0.294	0.322
Adjusted R^2	0.184	0.17	0.162	0.211	0.242
Residual Standard Error	1.634	1.244	1.515	1.1	1.457
$F(11, 93)$	3.134**	2.939**	2.827**	3.525***	4.014***

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