

Authors are encouraged to submit new papers to INFORMS journals by means of a style file template, which includes the journal title. However, use of a template does not certify that the paper has been accepted for publication in the named journal. INFORMS journal templates are for the exclusive purpose of submitting to an INFORMS journal and should not be used to distribute the papers in print or online or to submit the papers to another publication.

Till Tech Do Us Part: Betrayal Aversion and its Role in Algorithm Use

(Authors' names blinded for peer review)

Failing to follow expert advice can have real and dangerous consequences. While any number of factors may lead a decision maker to refuse expert advice, the proliferation of algorithmic experts has further complexified the issue. One potential mechanism that restricts the acceptance of expert advice is betrayal aversion, or the strong dislike for the violation of trust norms. This study explores whether the introduction of expert algorithms in place of human experts can attenuate betrayal aversion and lead to higher overall rates of seeking expert advice. In other words, we ask: are decision makers averse to algorithmic betrayal? The answer to this question is uncertain *ex ante*; robust evidence exists showing that even inanimate products (e.g., airbags and vaccines) suffer reduced uptake due to betrayal aversion. We answer this question through an experimental financial market where there is an identical risk of betrayal from either a human or algorithmic financial advisor. We find that the willingness to delegate to human experts is significantly reduced by betrayal aversion, while no betrayal aversion is exhibited towards algorithmic experts. The impact of betrayal aversion towards financial advisors is considerable: the resulting unwillingness to take the advice of the human expert leads to a 20% decrease in subsequent earnings, while no loss in earnings is observed in the algorithmic expert condition. This study has significant implications for firms, policymakers, and consumers, specifically in the financial services industry.

Key words: Algorithm Adoption, Artificial Intelligence, Behavioral Decision Making, Experimental Economics

1. Introduction

People often seek expert advice on investing for retirement, choosing a healthcare plan, and treating illness. However, individual uptake of expert advice has consistently been found to be sub-optimal (Woodhouse and Nieuwsma 1997). In a number of contexts, including healthcare and financial planning, failing to adhere to expert advice can result in significant negative consequences (Seiders et al. 2015). Further increasing complexity, expert advice is rapidly being augmented by algorithms. Consumers are faced with the decision of whether to let an algorithm manage their pension investments (Lourenço et al. 2020), suggest potential dating partners (Prahl and Van Swol 2017),

or control what digital content they read and watch (Prahl and Van Swol 2017, Van Swol 2011). This trend of algorithmic advice continues to increase, and likely will for years to come (Tetlock and Gardner 2016). Importantly, in many of these instances, evidence-based forecasts made by algorithms outperform those made by their human counterparts, making this new form of expert advice even more valuable (Beck et al. 2011, Dawes 1979, Highhouse 2008).

One potential mechanism that restricts the acceptance of expert advice is the fear of betrayal. Economics literature has shown that decision makers are often strongly averse to possibility of a betrayal, even when controlling for monetary risk (Bohnet and Zeckhauser 2004). Take, for example, an expert financial advisor whose incentive structure may encourage them to occasionally make a self-interested financial decision: a betrayal. If an investor chooses to hire the advisor, that advisor could either 1) perform as expected or 2) betray the investor. The total utility for the investor would then consist of the expected *monetary* return and the *emotional* cost of being betrayed (betrayal aversion). An investor who, following a betrayal, experiences a positive emotional cost is considered betrayal averse and may choose not to hire the advisor to avoid experiencing the emotional disutility even if the expected monetary return is larger than any alternatives.

Our work aims to address what changes, if any, occur to levels of exhibited betrayal aversion when an expert is replaced by an algorithm. Critically, many accounts of betrayal aversion bypass the need for human intentions completely and introduce a significant potential for betrayal aversion to persist for non-human advisors. For example, Koehler and Gershoff (2003) show that safety products such as airbags or vaccines can elicit feelings of betrayal in their users. Other work shows that humans mindlessly attribute social rules and expectations to computers, even when they know the computers lack intentionality (Nass and Moon 2000, Nass et al. 1994). Currently, no work has examined if betrayal aversion persists when a human expert is augmented or replaced by an algorithm. This disconnect is essential to understanding the role of algorithms in motivating the acceptance of expert advice. Therefore, we ask: *are humans averse to algorithmic betrayal?*

Given the rapid growth of algorithm use in financial market trading¹ and its potential to increase efficiency, transparency, and capacity in the trading process (Bell and Gana 2012), we contextualize our work within the financial services industry specifically. To answer our central research question, we conduct an economic lab experiment in which participants play a 40-round financial trading game using a simulated market structure and a real financial trading algorithm designed to give advice that would be comparable to what consumers would receive from both human and algorithmic experts in a real-life setting. Participants are given the choice in each round to either make their own decision on how much of their endowment to use to purchase a risky asset or

¹ <https://therobusttrader.com/what-percentage-of-trading-is-algorithmic/>

use the advice of an expert. No deception was used in our experiment and all participants were compensated based on their actual investment decisions.

The experiment employs a two factor (2x3) between-subjects design. The first manipulated factor is the *presentation* of the expert as either human or algorithm; importantly, the underlying algorithm that generates the advice is identical for both the human and the algorithm expert. The second manipulated factor is the presence of a salient betrayal risk where treated participants are informed that there is a slight chance that the human or algorithmic expert may make a self-interested decision on their behalf. To isolate the emotional response to the betrayal (betrayal aversion), we include an error condition where participants learn that there is a slight chance of an accidental error occurring. Ex ante, one would expect the error condition to elicit similar impacts on the perception of the monetary performance of the expert, while the lack of intentionality would limit the presence of betrayal aversion. We recruited 275 participants from Amazon Mechanical Turk (mTurk) to take part in the experiment over a video sharing platform in, what we believe to be, one of the first digitally-face-to-face economic experiments conducted on mTurk.

Our analysis captures changes in advisor usage trends in the betrayal risk (both human and algorithm), error risk, and control groups. When in a human advisor condition, we find a roughly 16% decrease ($\beta = -0.159$, $p = 0.024$) in advisor usage when informed of a betrayal risk but no decrease when only informed of a risk of accidental error ($\beta = 0.007$, $p = 0.924$). This result implies that the effect we measure is due to betrayal aversion and not a perceived decrease in expected monetary return. Interestingly, we do not find this effect in the algorithm conditions ($\beta = 0.055$, $p = 0.431$). These findings imply that substituting an algorithmic advisor for a human advisor can significantly attenuate betrayal aversion. We also observe that betrayal aversion results in significant economic consequences, resulting in a loss of \$2.15 ($p = 0.072$), an almost 20% decrease in overall returns for affected participants. Analysis of exit questions substantiates that betrayal aversion may be a key aspect of this difference: participants in the betrayal condition with a human advisor reported feeling more concerned about being misled ($p = 0.006$), having their trust violated ($p = 0.010$), and most importantly, feeling betrayed ($p = 0.001$). We do not find these effects when the advisor was an algorithm, nor in the error-risk condition.

We find similar effects in a second experiment that utilizes the same experimental design while 1) drawing from a different population, undergraduate students recruited from a university economics lab and 2) utilizing a different human advisor. We find that risk of betrayal decreased uptake of the human advisor by 12% ($p = 0.0901$) and earnings by \$2.13 ($p = 0.003$). The effects were again attenuated when assigned to an algorithm advisor ($p = 0.026$). This second experiment demonstrates the robustness of the findings to research platform, participant sample, and advisor selection. Note that the lab utilized explicitly prohibits any form of deception in experimental

procedures, so this experiment reduces concerns about the credibility of the information provided to participants when compared to the mTurk sample.

Our work contributes to the nascent literature exploring the acceptance, or lack thereof, of expert algorithms. The majority of the literature (see Logg et al. (2019) for a notable exception) highlights decision-maker's preference towards human experts (Dietvorst et al. 2015, Longoni et al. 2019) and identifies traits that algorithms *lack* that may lead an individual to exhibit "algorithm aversion" (Castelo et al. 2019). Proposed solutions to the nonacceptance of algorithmic experts have focused on either enhancing the human involvement in the algorithmic decision making by allowing users to add feedback (Dietvorst et al. 2018), making an algorithm more human-like by highlighting the affective abilities of the tool (Castelo et al. 2019) or anthropomorphizing the algorithms (Schanke et al. 2021). Our results highlight an important nuance to the conclusions of prior work: reducing the human element associated with algorithmic experts can remove some of the traditional barriers to the acceptance of expert advice.

We also contribute to this literature by identifying a potential strength of algorithmic experts and introducing a novel phenomenon to the algorithmic adoption literature – betrayal aversion. This phenomenon, largely unstudied by the information systems discipline, has been shown by behavioral and experimental economists to significantly influence an individual's decision-making. Given that the economics literature has shown the phenomenon to exist even when the betrayal comes from inanimate products or objects, its role in the human-algorithm relationship is essential to explore. However, no work, in economics or information systems, has explored the extent to which betrayal aversion may persist (or not) when a human is replaced by an algorithm. By bridging these two literatures, we show that despite prior evidence showing the potential of inanimate objects to elicit betrayal aversion, algorithms may attenuate the phenomenon instead.

We also provide valuable insights to the Financial Technology (FinTech) literature. Algorithmic trading services have grown substantially and trading services augmented by artificial intelligence are likely in the near future (Gomber et al. 2018). Recent estimates show that between 60 and 75% of the total trading volume in the U.S. involves some form of algorithmic trading, but more research is needed to understand the implications of this trend (Gomber et al. 2018, Hendershott et al. 2021, Alt et al. 2018, Kou 2019, Cao et al. 2020). We specifically respond to Hendershott et al. (2021) by examining the role of algorithmic financial advice and providing a potential solution to a significant barrier to the acceptance of expert advice within financial services.

2. Conceptual Background: Algorithms and Betrayal

Our work examines two primary areas of the literature. First, we look to previous work on algorithm adoption and aversion that focuses on specific drivers of use and disuse to better understand

predictors of algorithmic acceptance. Then, we explore the betrayal aversion literature in economics and introduce it as a previously unexamined factor in algorithm uptake. In the section that follows, we contextualize our work to financial technology and examine the theoretical expectations of the effect of betrayal aversion on the use of trading algorithms.

2.1. Algorithm Adoption and Aversion

Prior work considering the use and adoption of information technology has strived to consider user's social and emotional beliefs (Benbasat and Wang 2005, Qiu and Benbasat 2005, Venkatesh and Davis 2000). This has led to the perception of IT artifacts as "social actors" (Al-Natour and Benbasat 2009, Reeves and Nass 1996). In the Computers are Social Actors (CASA) paradigm, humans have the same social rules and expectations of technology that they have of other humans (Nass et al. 1994). This paradigm now has renewed importance as digital ecosystems built on powerful customization algorithms are influencing the daily lives of their users (Parker et al. 2017). Artificial intelligence and algorithmic decision makers are augmenting or replacing their human counterparts, even in domains like medicine (Jussupow et al. 2021), and online services are increasingly using algorithms to determine the content users see (Orlikowski and Scott 2015). Given the increased prevalence of algorithm-run services and platforms (Ransbotham et al. 2018), research that considers the complexities of algorithmic use and adoption is of growing importance.

While algorithm adoption has grown, some research has shown that overall, people still prefer human decision makers over algorithmic decision makers (Diab et al. 2011, Eastwood et al. 2012). Dietvorst et al. (2015) terms this algorithm aversion, when individuals avoid using algorithms, after they see them make an error (See Burton et al. (2020) for a review).

A number of potential drivers to this phenomenon have been introduced. First, false expectations lead to an individual's unwillingness to use and accept an algorithm. This may occur because individuals believe that human error is random whereas algorithmic error demonstrates that the entire system is flawed and did not simply produce a one-time error (Dietvorst et al. 2015, 2018, Highhouse 2008). Alternatively, individuals may value forming a professional relationship with those from whom they seek advice (Alexander et al. 2018, Önköl et al. 2009, Prahll and Van Swol 2017). Further, Castelo et al. (2019) finds that algorithm aversion is dependent on task type, for example, when a task is perceived as subjective, individuals are less inclined to trust an algorithm. However, when individuals perceive that an algorithm is able to learn from experience, the effect of algorithm aversion is lessened (Berger et al. 2021). Additionally, when an individual gains experience with an algorithm, and receives feedback outlining the superior performance of the algorithm, aversion decreases over time (Filiz et al. 2021).

Research also shows that human decision makers fear relinquishing control, thus preferring a certain level of power over an algorithm's advice (Colarelli and Thompson 2008, Scherer et al. 2015).

Dietvorst et al. (2018) further explores the integration of human-in-the-loop decision making where the human decision maker oversees the algorithm's processes, creating a perception of control.

There is also a small but growing literature considering the opposite effect: that people prefer algorithms to humans. Some work has shown that when individuals are asked to remember important information, they opt to outsource the task to algorithms, showing a preference towards algorithms over their own abilities (Sparrow et al. 2011). Further, Logg et al. (2019) shows that individuals rely more heavily on algorithmic advice than advice from others or their own judgement, a phenomenon they term "algorithm appreciation." While most of the past work has proposed solutions to this debate by focusing on making algorithms more human (Schanke et al. 2021, Wilson et al. 2017), no work, to our knowledge, has considered the potential for algorithms to bypass traditional barriers to the willingness to seek advice often faced by human experts. We contribute to this literature by using insights from behavioral economics to explore the role of betrayal aversion in algorithm adoption.

2.2. Betrayal Aversion

Betrayal aversion is defined as the strong dislike for violations of trust norms implicit in a relationship between two parties (Aimone et al. 2015). Importantly, past work has shown that the degree of betrayal aversion depends on the extent to which the betrayers have a duty to protect. In Koehler and Gershoff (2003), they show that participants view a betrayal as more severe when a member of the military commits treason than when its committed by an individual in an unrelated career. The same phenomenon can be seen with criminal punishment where, for example, people believe that day care workers who abuse a child should be punished more harshly than a janitor who abuses a child in the same way. This has led to demands to revise sentencing guidelines to make them sensitive to the extent to which the defendant's role included ensuring an individual's wellbeing (Shnoor 2009).

While many considerations of betrayal aversion have focused on human-to-human interactions, other work has acknowledged that non-human agents elicit fears of betrayal in the same way that humans do. Koehler and Gershoff (2003) show this phenomenon through the elicitation of opinions on different vehicle air bags. They give participants the option to choose between Air Bag A that carries a 2% chance of driver death when involved in a serious accident or Air Bag B that carries a 1% chance of death in a serious car accident plus a 0.01% chance of the air bag causing the death when the driver would have otherwise survived. From a purely risk-minimizing perspective, Air Bag B is preferred. However, only 32.6% of participants in the experiment chose it. Similar findings from the same paper are produced when considering smoke alarms and vaccines. Other work has shown that consumers who feel betrayed by products punish the firm by way of complaints or negative reviews (Grégoire and Fisher 2008).

Importantly, past work has focused on removing elements of risk and trust to isolate betrayal aversion as the primary explanatory variable. The general method of studying betrayal aversion through this lens involves having one group of participants play a trust game with random chance determining the player's financial return, instead of another human player. This work shows that participants require a higher minimum acceptable probability (MAP) of earning a return in a trust game played with a human compared to a similar game involving random chance, providing evidence of non-monetary betrayal aversion (Bohnet et al. 2008, 2010, Bohnet and Zeckhauser 2004, Hong and Bohnet 2007).

This phenomenon has expanded beyond economic decision making to other fields as well. Some work has shown that managers are influenced by betrayal aversion when developing relationships with their employees (Birnberg and Zhang 2010). The results showed that some managers spend more money to prevent betrayal than they could have conceivably lost from the betrayal itself, implying that betrayal aversion, not just the betrayal, may lead to a decrease in expected monetary return. As firms increasingly are incorporating algorithmic tools and services into their underlying business models, the presence of betrayal aversion in both consumers and employees is likely to influence subsequent adoption and use.

3. Theoretical Development

Betrayal aversion with respect to algorithms has not yet been explored. To lend concreteness to our theoretical expectations, we focus on algorithm use in the financial services context. As mentioned previously, algorithmic trading (AT) has grown substantially in financial markets around the world. The use of AT has the ability to increase efficiency, transparency, and capacity in the trading process (Bell and Gana 2012). These increases are realized through the rapid collection of public information (Chakrabarty et al. 2015) and the ease in converting that information into prices (Brogaard et al. 2014). Additionally, AT can enhance liquidity and increase the informativeness of price quotes (Hendershott et al. 2011).

Although investors and traders have increasingly developed trading algorithms designed to mimic the trends and actions of human traders (Hendershott et al. 2021), consumer unwillingness to adopt often superior algorithms can result in lower returns (Ge et al. 2021). In addition, if betrayal aversion persists in this context, the phenomenon should be considered in the design and marketing of algorithmic tools. While we defer to future work for specific design recommendations, we highlight for the first time the potential of algorithmic betrayal aversion and open the door for further explorations. In the rest of this section, we use the context of trading algorithms to explore the role of betrayal aversion in algorithmic use and leverage common assumptions of utility maximizing agents (e.g. Bernard et al. (2015)). What follows is not a formal analytical model but rather

a conceptual exercise to help organize competing dynamics and articulate ex ante expectations around algorithmic betrayal aversion.

Assume that a decision maker has a basic utility function such that:

$$U(E, \sigma) = E - \frac{1}{2}R\sigma^2 \quad (1)$$

where E is the expected return or mean outcome of a financial market trade that follows a well-defined probability distribution function with a variance of σ^2 . R is a constant risk aversion parameter that reflects the decision maker's risk preference (reflective of the Arrow-Pratt risk aversion measure first introduced by Pratt (1964) and Arrow (1971)). Assume that the decision maker has a choice between two financial strategies for the trading decision. They can choose to make their own decision (S), or they can choose to use the recommendation of a financial advisor (A) such that:

$$U(E_i, \sigma_i) = E_i - \frac{1}{2}R\sigma_i^2, \text{ where } i \in [A, S] \quad (2)$$

Before making the choice, they are told the following:

$$E_A > E_S \quad (3)$$

Therefore, an individual would choose to use the advisor's recommendation over their own decision if the increase in expected return is higher than the difference in the variance scaled by the risk aversion parameter such that:

$$E_A - E_S > \frac{1}{2}R(\sigma_A^2 - \sigma_S^2) \quad (4)$$

However, they are also told that if A is chosen, there is a small probability of event B, or a betrayal, also occurring. This probability can be expressed as E_B (the mean probability of a betrayal occurring). The betrayal occurs when the advisor makes a self-interested choice that may or may not align with the ideal trading strategy of the decision maker, with equal probability. In other words, the betrayal may increase or decrease the investor's return. However, it is classified as a betrayal because the advisor makes the choice that best suits their own interests with no regards for the investor. If the goals of the advisor and the investor happen to align when a betrayal occurs, the investor may still financially benefit².

² While a real-world betrayal may come with financial consequences, for the sake of simplicity, we assume here that the expected return is constant. This allows us to isolate the betrayal aversion parameter more easily, increasing the clarity of our expectations.

Therefore, the betrayal does not decrease their expected monetary return such that $E_{(A|B=1)} = E_{(A|B=0)}$. However, given the extensive empirical findings described in the previous section, we posit that the utility function also includes a non-monetary element β , or betrayal aversion, in which an individual exhibits an emotional cost from a betrayal that is distinct from any consideration of expected monetary returns. The difference in utility of choosing A over S would then be:

$$\Delta U = \left[E_A - \frac{1}{2} R \sigma_A^2 - \beta E_B \right] - \left[E_S - \frac{1}{2} R \sigma_S^2 \right] \quad (5)$$

where β represents the level of betrayal aversion the individual exhibits and E_B is the expected betrayal outcome [0,1] as described above.

Let us assume that there is an individual with a unique risk aversion parameter, R^* , who is indifferent between the two trading strategies assuming $E_A > E_S$ and $\sigma_A^2 > \sigma_S^2$:

$$E_A - \frac{1}{2} R^* \sigma_A^2 = E_S - \frac{1}{2} R^* \sigma_S^2 \quad (6)$$

$$R^* = \frac{2[E_A - E_S]}{\sigma_A^2 - \sigma_S^2} \quad (7)$$

If this individual also exhibits some level of betrayal aversion and the probability of a betrayal is non-zero, such that $\beta > 0; E_B > 0$, they will choose to make their own trading decision over taking the advisor's, as shown below:

$$E_A - \frac{1}{2} R^* \sigma_A^2 - \beta E_B > E_S - \frac{1}{2} R^* \sigma_S^2 \quad (8)$$

$$-\beta E_B < (E_S - E_A) + \frac{1}{2} \frac{2[E_A - E_S]}{\sigma_A^2 - \sigma_S^2} (\sigma_A^2 - \sigma_S^2) \quad (9)$$

$$\beta E_B > 0 \quad (10)$$

Therefore, in line with the central proposition of Bohnet and Zeckhauser (2004), we hypothesize:

HYPOTHESIS 1. *Holding the objective monetary return constant, the willingness to outsource a decision to a human expert will decrease when there is a chance of betrayal.*

This hypothesis is consistent with previous research on betrayal aversion towards a human counterpart. Since previous work has not examined betrayal aversion within financial investing, we first aim to explore whether betrayal aversion exists in this context. We then examine whether levels of betrayal aversion differ when a human advisor is replaced with an algorithmic advisor. Prior work has shown that humans mindlessly apply social rules and expectations to computers, even

when they know that the computers lack feelings and intentionality (Nass and Moon 2000, Nass et al. 1994). Therefore, we anticipate betrayal aversion persisting for algorithmic experts. In other words, our expectations imply that:

$$\beta_{Algorithm} > 0 \quad (11)$$

leading us to hypothesize:

HYPOTHESIS 2. Holding the objective monetary return constant, the willingness to outsource a decision to an algorithmic expert will decrease when there is a chance of betrayal.

Prior work has shown that the intensity of an individual's emotional response to betrayal depends on both the significance and depth of the relationship between parties and the magnitude of the harm caused by the betrayal (Rachman 2010). Revisiting our theoretical model, we can then further differentiate betrayal aversion, β into the magnitude of harm, H , and the depth of the relationship D , such that:

$$\beta = H + D \quad (12)$$

If one controls for the magnitude of harm (the decrease in expected return, in our case, $H=0$), the variance, and the risk preference between human and algorithm contexts, then it would follow that betrayal aversion would rely on the perceived significance of the decision-maker's relationship with the algorithm. Algorithms, generally, lack the ability to display social intelligence, and therefore the relationship between humans and algorithms is limited (Frey and Osborne 2017, Rafaeli et al. 2016). If this detracts from the perceived significance of the relationship a human feels they have with an algorithmic expert, we can assume that:

$$D_{Human} > D_{Algorithm} \quad (13)$$

Further showing that:

$$\beta_{Human} > \beta_{Algorithm} \quad (14)$$

Therefore, we hypothesize:

HYPOTHESIS 3. Holding the objective monetary return constant, there will be a smaller decrease in the willingness to outsource a decision to an algorithm compared to a human when there is a chance of betrayal.

4. Financial Investment Game

The following section presents the experimental design for the financial investment game our participants were tasked with completing. We first discuss the variation of the experimental factors before providing detail as to the simulated financial market and the investment algorithm that was designed to predict fluctuations in said market. Finally, we discuss our recruitment method and the procedure completed by participants.

4.1. Design

We utilize a two-factor (2x3) between-subjects design. Our two factors differentiate the presentation of the advisor (human or algorithm) and the additional risk of some disutility (betrayal risk, accidental error risk, or no additional risk). The experimental task is a 40-round financial investing game, consisting of 10 baseline rounds (where the additional risk factor was not introduced) and 30 principal rounds. Each participant was assigned to either the human advisor or algorithm advisor treatment for the duration of the experiment. Participants start the task with a trading endowment which they can invest in a risky financial asset in our experimental market. The experiment is incentivized so that the cumulative return from 10 baseline rounds and 30 principal rounds determine a participant's earnings.

In both the human and algorithm advisor treatments, participants begin each round by deciding whether to choose their own level of investment or utilize an expert advisor. In the initial 10 baseline rounds there was no possibility of betrayal. At the start of the 30 principal rounds which followed, participants in the betrayal condition read a disclaimer indicating that there was incentive misalignment with the human/algorithm advisor that could result in occasional negative returns. Participants also learned that, historically, the human/algorithm advisor had outperformed participants' investment decision, thus ensuring the expected return for the advisor/algorithm would be higher than the expected return of investing themselves. For control, both the human and algorithm advisors offered the same advice, and we verified that the decision rule used by both advisor types (discussed below) truly did outperform participants' investment decisions. We included a shortened version of the disclaimer during each round to ensure the saliency of the treatment. The full disclaimer can be found in Appendix A.

4.1.1. Market Structure & Algorithm Design The risky asset in our experimental market follows a two-state (good or bad) Markov-switching Gaussian random walk with a state switching probability of 35%³, adapted from Zhang (2020). In other words, if the previous round was a good

³ Simulations were run to identify an optimal switch rate that balanced the predictability of market fluctuations with the desire for the algorithm to consistently outperform the individual investors.

state, there is a 35% chance that the following round would be a bad state and a 65% that it would be a good state. The expected return of the good state is described by:

$$r_t = \mu_1 dt + \sigma_1 dZ_t \quad (15)$$

where $\mu_1 = 0.10$, $\sigma_1 = 0.10$, and Z_t is white noise. The expected return of the bad state is described by:

$$r_t = \mu_2 dt + \sigma_2 dZ_t \quad (16)$$

where $\mu_2 = -0.05$, $\sigma_2 = 0.05$. If the asset price increased, participants received the highest profit if they use their entire endowment to purchase the asset. If the asset price decreased, participants received the lowest loss by purchasing none of the asset. This allows us to account for the fictive error occurring in the market process. We can define the fictive error as:

$$f^+ = (100\% \times r_t^+) - (Allocation \times r_t^+), f^- = (0\% \times r_t^-) - (Allocation \times r_t^-) \quad (17)$$

where r^+ is a positive return on the asset and r^- is a negative return on the asset, *Allocation* is the percentage of the participant's endowment they used to purchase the risk asset, $Allocation \times r_t^+$ is the experienced gain for the positive return, and $Allocation \times r_t^-$ is the experienced loss for the negative return.

In both the human and algorithm conditions, the suggestion from the advisor was derived using the same Bayesian investment model. This allowed for a well-defined probability that the market was in a good state. Assume that r_t is a price change of the risky asset that the participant observed in period t . We then know that the probability that the market is currently in a good state is:

$$q_t = Pr(s_t = good | r_t, r_{t-1}, \dots, r_2, r_1) \quad (18)$$

Therefore, we know:

$$q_t(q_{t-1}, r_t) = \frac{Pr(r_t | s_t = good) Pr(s_t = good | q_{t-1})}{Pr(r_t | s_t = good) Pr(s_t = good | q_{t-1}) + Pr(r_t | s_t = bad) Pr(s_t = bad | q_{t-1})} \quad (19)$$

Further, we know the expected return in a good state from Equation 15. We can define the distribution function of the expected return $Pr(r_t | s_t = good)$ as $f_{good}(r_t) = \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{(r_t - \mu_1)^2}{2\sigma_1^2}}$. Similarly, in a bad state, the distribution function would be $f_{bad}(r_t) = \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-\frac{(r_t + \mu_2)^2}{2\sigma_2^2}}$. Given the updated belief that the previous trial was in a good state q_{t-1} , $Pr(s_t = good | 1_{t-1}) = (1-p)q_{t-1} + p(1-q_{t-1})$ and $Pr(s_t = bad | q_{t-1}) = pq_{t-1} + (1-p)(1-q_{t-1})$ where p is the state switch rate described above ($p = 0.35$). Therefore, we can rewrite Equation 19 as the following:

$$q_t(q_{t-1}, r_t) = \frac{f_{good}(r_t)((1-p)q_{t-1} + p(1-q_{t-1}))}{f_{good}(r_t)((1-p)q_{t-1} + p(1-q_t)) + f_{bad}(r_t)(pq_{t-1} + (1-p)(1-q_{t-1}))} \quad (20)$$

The algorithm then derives $q_t(q_{t-1}, r_t)$ based on the previous round's return to determine the probability of this round being in a good state. We then derive the optimal amount of the endowment to invest using the maximization of the utility function subject to the relationships between expected return and standard deviation of the risky asset as follows:

$$y^* = \frac{q_t\mu_1 + (1-q_{t-1})\mu_2}{0.02A\sigma^2} \quad (21)$$

where A is a measure of risk tolerance for an individual. To simplify the experiment, rather than measure individual risk preferences in advance we used the average risk aversion level ($A = 5.17$) from Holt and Laury (2002) shown in Equation 21. We piloted 10 rounds of the trading task with a group of volunteers and had them make their own investment decisions in each round and recorded their investment returns. Afterwards, we simulated returns data using the rule used by the human/algorithm advisors with the same stream of market returns. We found that the returns earned by the advisor were 65% higher on average than that earned by the participants, verifying that we were accurate in informing participants that utilizing the advisor would yield higher investment earnings was accurate.

4.1.2. Participants & Procedure We recruited 275 participants through Amazon Mechanical Turk. We restricted sign-ups to those located in the US along with minimum requirements for the number of HITs completed and the acceptance rates of those HITs. We advertised our study as a financial game with the opportunity to make on average \$20-\$30/hour for their participation⁴.

Participants signed up for a designated time slot using a Qualtrics survey. They received email reminders 2 days prior to their time slot and 1 hour before the experiment was scheduled to begin. Emails were sent via automated scripts utilizing the mTurk API. Experiment sessions took place over a 2-week period. Even though this is an individual task, for efficiency, multiple participants completed the experiment at once. Each session began with participants joining a Zoom meeting room where the experimenter gave instructions. In the human condition, a third-year MBA student was introduced using the following statement: "This is Brandon. He is trained in the financial task you are being asked to complete today and will be aiding in the execution of the experiment." The student came to each session dressed professionally and the same human advisor was used for the

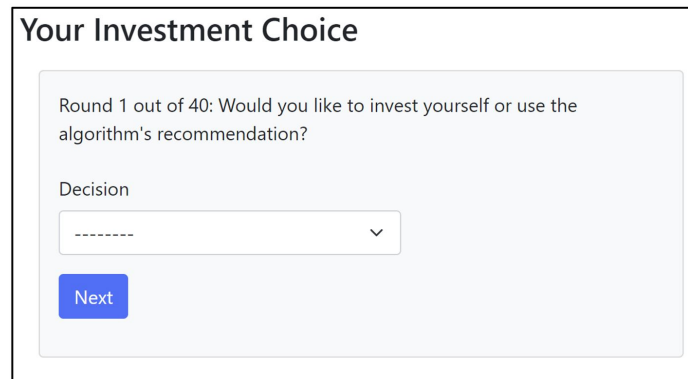
⁴ The average mTurk requester offers \$11/hour for work. We wanted our earnings potential to be particularly attractive to workers given the fact that most workers may not have experience with face-to-face HITs over Zoom and we wanted to complete data collection in a short period of time.

duration of data collection to ensure no effects of advisor appearance or other characteristics. The procedure was the same in the algorithm advisor treatment, except that no advisor was introduced.

Participants were next sent a link that directed them to the online experiment. The experiment interface, including both the financial market and the investment algorithm, was developed using oTree Version 5.9 (Chen et al. 2016) and was deployed via Heroku, a commercial cloud hosting provider. Upon clicking the link, they were directed to the experiment and began by going through a brief tutorial that explained the market process and the task that they were being asked to complete. Then they were given the opportunity to invest in 10 unpaid practice rounds where they had to make their own investment decision in each round. After completing the practice rounds, participants were directed to a short comprehension quiz (see Appendix B). They were informed that they were being granted an endowment of 1000 research points with which to play the game, which translates to \$5USD. They were also compensated an additional \$5 which was added to their earnings at the end of the experiment.

At the onset of the 40 payoff determining rounds, participants were informed that the following rounds would be the same as the practice rounds with two key differences. First, these rounds would determine their true payoff at the end of the experiment. Second, they learned that they had the option in each round to make their own investment decision or to utilize a financial advisor/algorithm to make the decision for them. Before making their decision, they were directed to a waiting page where they saw a short video of the available advisor, either the human advisor or a graphic that visualized an algorithm working and were informed that the advisor was generating their investment recommendation based on market data (see Appendix C). This was to ensure that timing between rounds remained consistent and participants behavior was not impacted by perceptions of the length of time required to decide oneself relative to the advisor. It also improved the realism of the task since the advisor took a non-trivial amount of time to generate the recommendation. Importantly, prior to the first principal round, participants had no knowledge of the betrayal or error treatment and were only told that they had the choice to use an advisor's (either an algorithm or a human) recommendation and that those who chose to use the advisor's recommendation earned more on average.

To avoid using deception, our human financial advisor took part in the actual experiment beyond being present in the Zoom session. His interface (shown in Appendix D) would report the suggested investment, using the same rule as the algorithm, for each participant in each round and would ask him to submit the recommendation. We kept the design as simple as possible to avoid any possibility of human error while avoiding deception. While participants may have believed that the advisor was manually developing investment recommendations, they were only ever told that the advisor would be submitting recommendations.



Your Investment Choice

Round 1 out of 40: Would you like to invest yourself or use the algorithm's recommendation?

Decision

----- ▾

Next

Figure 1 Decision Page

After the 10 baseline rounds participants in an assigned betrayal condition received a disclaimer that there was a small chance that the advisor or algorithm would intentionally over-invest even if he/it was not confident that it was a good market. Likewise, if the participant was in an error-risk condition, they were told that there was a small chance of an accidental error occurring. Again, the full disclaimers can be found in Appendix A.


Figure 1 shows the decision page for the control group. In the betrayal treatment and error treatment, this decision was accompanied by a small additional disclaimer stating, “The algorithm will occasionally over-invest even when it is not clear that it is a good market” and “The algorithm will occasionally make an accidental error”, respectively, to ensure the saliency of our treatment. The decision of whether to use the advisor was made at the beginning of each round.

If the participant chose to invest themselves, they were directed to the investment page (Figure 2). Participants chose what percent of their current endowment they wanted to use to purchase the risky asset. They were next directed to the results page (Figure 3) which displayed how much they invested, the percentage change in asset price for that round, their current round and cumulative returns (in research points), their total research points, as well as two graphs: one that tracks the price movement of the asset and the other that tracks their research points. If a participant chose to use the advisor’s suggestion, they bypassed the investment page and went directly to the results.

After completing each of the investment decisions, participants answered a short demographic survey that asked for their gender, race, ethnicity, education, employment, experience with investing, and familiarity with algorithm design. They also answered exit questions to measure their perceptions of trust, betrayal, and regret (see Appendix E). Participants earned \$5 compensation in addition to the USD equivalent of their ending research points where 100 points equaled one dollar.

Your Investment Choice

You currently have 1000.0 research points. What percent of these points would you like to invest in the asset for Round 1?



This will invest 500 of your points in the asset.

Next

Figure 2 Investment Page



Figure 3 Results Page

4.2. Estimation Approach

Our main dependent variable for our analyses is the participant's choice to use the advisor's recommendation or make their own decision in each round. Given the multi-round repeated structure of our experiment, we use a panel random effects model for our estimation.

$$Use_{it} = \beta_1 \times Advisor_i + \beta_2 \times Risk_i + \beta_3 (Advisor_i \times Risk_i) + \alpha \times Demographics_i + \delta_t + \theta_i + u_i \quad (22)$$

The dependent variable, Use_{it} , is a indicator that equals 1 if participant i chose to use the advisor in round t and 0 otherwise. $Advisor_i$ is a vector of binary indicators for the advisor treatment condition assigned to each participant [algorithm vs. human]. $Risk_i$ is a vector of binary indicators for the risk treatment condition assigned to each participant [control vs. error risk vs. betrayal risk]. We also include interactions between advisor presentation and risk treatments. $Demographics_i$ is a vector of controls capturing heterogeneity in individual demographics (e.g., gender, race, ethnicity, education, work status, experience with algorithms, experience with investing). δ_t includes round fixed effects and θ_i is the participant-specific random effect. The error term, u_i , is clustered on participant. Estimates on the randomly assigned treatments are assumed unbiased due to the lack of correlation with unobserved individual differences and the error term. This assumption is tested in Section 5.1. This estimation approach allows for a robust account of time trends in our data and corrects for the nonindependence of multiple observations of the decision to use the advisor from a single participant.

Additionally, we estimate treatment effects for total returns. Since the algorithm, on average, realizes higher returns than participant decisions, any decrease in advisor usage should be accompanied by a decrease in returns as well.

5. Results

Two hundred and seventy-five individuals took part in our financial trading game. Table 1 shows the distribution of participants across treatments in our experience. We find a roughly equal split for both the additional risk dimension (control vs. betrayal risk vs. error risk) and the advisor presentation dimension (algorithm vs. human). The average advisor usage was 56.44% for our baseline rounds and 42.59% in the 30 principal rounds.

	Control	Betrayal Risk	Error Risk	Algorithm	Human	Total
Sample Size	3,560	3,720	3,720	5,760	5,240	11,000
Individuals	89	93	93	144	131	275
Percent	32.4%	33.8%	33.8%	52.4%	47.6%	100%

Table 1 Breakdown of Sample by Experimental Conditions

Table 2 provides a description of our variables and summary statistics for each. We have a diverse sample of participants with a relatively even split between males and females and racial and ethnic statistics that are comparable to the national averages⁵. Fifteen participants stated that they had extensive investment experience and only 7 stated that they had extensive experience with algorithms.

⁵ <https://www.census.gov/quickfacts/fact/table/US/PST045221>

Variable	Description	Mean (1)	S.D. (2)
Use	The choice to use the advisor in a specific round	0.461	0.498
Earnings	The total earnings from the game	9.141	5.292
Male	Whether the individual is a male	0.455	0.498
Caucasian	Whether the individual is Caucasian	0.749	0.434
African American	Whether the individual is African American	0.0945	0.293
Asian	Whether the individual is Asian	0.124	0.329
Hispanic	Whether the individual is Hispanic	0.0655	0.247
Some College	Completed some college	0.0982	0.298
Bachelor's Degree	Completed bachelor's degree	0.702	0.457
Advanced Degree	Completed advanced degree	0.200	0.400
Student	Whether the individual is a student	0.0473	0.212
Unemployed	Whether the individual is unemployed	0.135	0.341
Full-Time Employed	Whether the individual is employed full-time	0.0255	0.158
Retired	Whether the individual is retired	0.578	0.494
† Investment Experience	The participant's experience with investing	1.735	0.551
† Algorithm Experience	The participant's experience with algorithms	1.498	0.549

† Likert Scale (1 = No Experience, 3 = Extensive Experience)

Table 2 Summary Statistics

5.1. Balance Checks

Before estimating our main effects, we evaluate the efficacy of random assignment by examining whether there are differences in demographic characteristics across treatment groups. Table 3 provides pair-wise comparisons for all the variables listed in Table 2 across each set of manipulations (additional risk and advisor presentation). We conduct 52 pairwise comparisons (13 variables \times 4 comparisons) and find that 48 of these comparisons identify insignificant differences between conditions. With an alpha of 0.1, we would expect that 5 comparisons are significant by random chance. We identify only 4 significant differences between these groups. Even so, we control for these variables in our analysis and continue to identify consistent results.

Next, we leverage our first 10 baseline rounds which precede the introduction of any betrayal or error risk. We evaluate whether any differences exist across our risk conditions in use of the expert advisor and total earnings prior to the introduction of the additional risk. Table 4 estimates the baseline effect of the experimental risk conditions. Additionally, Table 4 reports the baseline effect of the advisor presentation (human vs. algorithm) for the first 10 rounds.

We do not find any significant differences between risk conditions during the baseline rounds. For example, our primary advisor usage variable sees no significant difference between the betrayal risk treatment and our control ($p = 0.878$), the betrayal risk treatment and the error risk treatment ($p = 0.866$), or the error risk and control ($p = 0.996$). Interestingly, we also see no significant effects

Variable	Control vs. Betrayal	Control vs. Error	Betrayal vs. Error	Algorithm vs. Human
Male	0.663	0.109	0.238	0.180
Caucasian	0.302	0.882	0.233	0.810
African American	0.152	0.901	0.118	0.327
Asian	0.566	0.600	0.268	0.943
Hispanic	0.087*	0.341	0.422	0.780
Some College	0.408	0.609	0.176	0.981
Bachelor's	0.027**	0.140	0.459	0.472
Advanced Degree	0.320	0.243	0.861	0.952
Student	0.132	0.701	0.250	0.913
Full-Time Employed	0.109	0.367	0.011**	0.362
Other Employment	0.322	0.271	0.034**	0.328
Investment Experience	0.989	0.152	0.143	0.298
Algorithm Experience	0.900	0.693	0.583	0.412

Table 3 Comparison of Demographics between Conditions

on use between the algorithm and human conditions ($p = 0.770$). Those in the algorithm treatment chose to use the advisor, on average, 56.9% of the time while those in the human treatment used the advisor 55.6%. Estimating treatment effects on total earnings for the first 10 rounds similarly provides no significant results. Overall, we find substantial balance across individual characteristics and outcomes prior to the introduction of betrayal.

5.2. Main Effects

Absent the presence of betrayal or error, we find no significant difference in uptake between the human and algorithmic advisor ($p = 0.566$). However, this trend changes when the potential for betrayal is introduced. Figure 4 graphs the average advisor usage across each of the treatment sets. First, there appears to be a significant drop in human advisor usage for the betrayal risk treatment (Control $\mu = 45.1\%$; Error $\mu = 45.2\%$; Betrayal $\mu = 29.1\%$). However, we see no such drop off in advisor usage for the algorithm treatments (Control $\mu = 40.6\%$; Error $\mu = 48.9\%$; Betrayal $\mu = 46.2\%$). The lack of a negative effect in the error conditions implies that the betrayal effect for the human advisor is due to betrayal aversion and not a perceived decrease in expected return or increase in risk generally. Further, this provides preliminary evidence that substituting a human expert with an algorithm may attenuate betrayal aversion.

Our regression model confirms our summary results. The estimates are reported in Table 5. First, we find no significant effect of the algorithm presentation, absent additional betrayal or error risk ($p = 0.563$) (column (1)). On average, our estimates show that in the human conditions, adding betrayal risk decreases use by roughly 16% ($p = 0.023$) (column (1)). Conversely, introducing risk of an accidental error has no effect on usage ($p = 0.924$) (column (1)). This implies that betrayal

Variables	(1) Use	(2) Earnings	(3) Use	(4) Earnings
Betrayal Risk	0.00736 (0.0482)	0.998 (1.663)		
Error Risk	-0.000203 (0.0447)	2.722 (1.900)		
Algorithm			0.0110 (0.0376)	-1.017 (1.512)
Constant	0.889*** (0.127)	18.75*** (5.146)	0.888*** (0.125)	20.51*** (5.191)
Fixed Effects	YES	NO	YES	NO
Demographics	YES	YES	YES	YES
Observations	2,750	275	2,750	275
Number of ID	275	275	275	275

Robust standard errors in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 4 Baseline Effects – Ten Rounds Prior to Risk Treatment

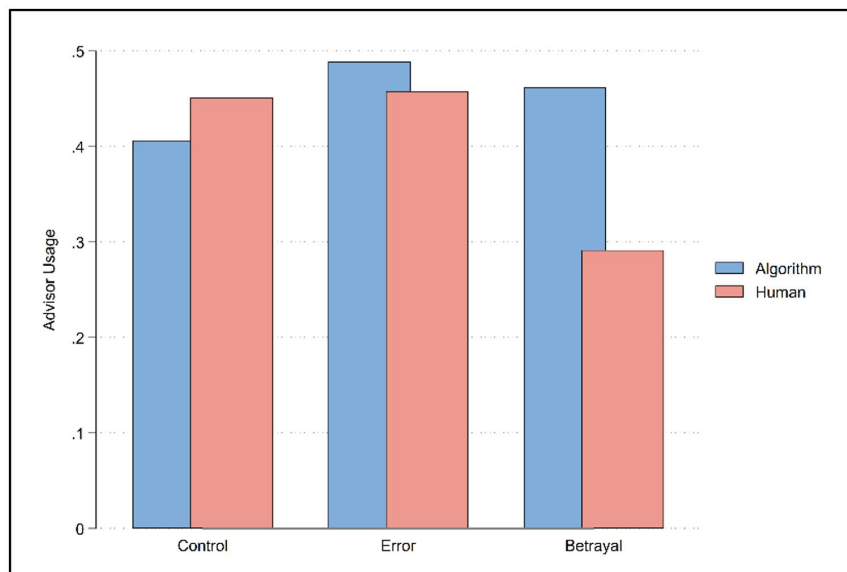


Figure 4 Advisor Usage Across Conditions

aversion is observed for the human conditions. Importantly, we find a positive and significant interaction effect of roughly +21% between betrayal risk and the algorithm treatment ($p = 0.031$) (column (1)). This implies that the betrayal aversion found in the human treatments is largely attenuated when the human advisor is switched out for an algorithm. These effects are consistent when controlling for time fixed effects and participant demographics (columns (2) and (3)).

Variables	Overall			1 – 10	11 – 20	21 – 30
	(1) Use	(2) Use	(3) Use	(4) Use	(5) Use	(6) Use
Betrayal Risk	-0.159** (0.0702)	-0.159** (0.0704)	-0.155** (0.0706)	-0.147** (0.0708)	-0.163** (0.0772)	-0.161* (0.0823)
Error Risk	0.00714 (0.0753)	0.00714 (0.0754)	0.00841 (0.0756)	0.0310 (0.0745)	0.0466 (0.0864)	-0.0492 (0.0866)
Algorithm	-0.0444 (0.0768)	-0.0444 (0.0770)	-0.0284 (0.0781)	-0.0106 (0.0770)	-0.0156 (0.0859)	-0.0651 (0.0885)
Algorithm × Betrayal	0.215** (0.0992)	.215** (0.0994)	0.178* (0.101)	0.141 (0.101)	0.206* (0.110)	0.214* (0.116)
Algorithm × Error	0.0754 (0.103)	0.0754 (0.103)	0.0680 (0.101)	0.00835 (0.104)	0.0544 (0.115)	0.140 (0.117)
Constant	0.451*** (0.0563)	0.421*** (0.0602)	0.600*** (0.146)	0.547*** (0.0598)	0.618*** (0.167)	0.709*** (0.166)
Fixed Effects	NO	YES	YES	YES	YES	YES
Demographics	NO	NO	YES	YES	YES	YES
Observations	8,250	8,250	8,250	2,750	2,750	2,750
Number of ID	275	275	275	275	275	275

Robust standard errors in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 5 Main Effect of Treatment on Use

5.2.1. Persistence of Treatment Effects We also explore heterogeneity across time for our treatment effects. Figure 5a and 5b plot average advisor usage across rounds for each condition. In Figure 5a, we can clearly see the effect of betrayal for the human conditions. It appears as if the effect remains relatively constant across rounds. In Figure 5b, algorithm usage rates for the betrayal treatment appear to drop initially before rising by the tenth principal round. This could imply that the algorithm experienced some initial betrayal effects but, unlike the human conditions, those effects dissipated quickly.

Table 5 above confirms these results. We estimate our model for the first 10 principal rounds (1-10) (column (4)), the next 10 principal rounds (11-20) (column (5)), and the last 10 principal rounds (21-30) (column (6)). While betrayal aversion towards the human advisor remains consistent across rounds, we find that in the first 10 treated rounds, substituting the human with an algorithm does not fully attenuate the betrayal aversion ($p = 0.266$) (column (3)). However, after at least 10 treated rounds, the betrayal aversion towards algorithms is completely attenuated ($p = 0.058$) (column (4)). As stated above, this provides evidence of some initial, but not persistent, betrayal aversion for algorithms.

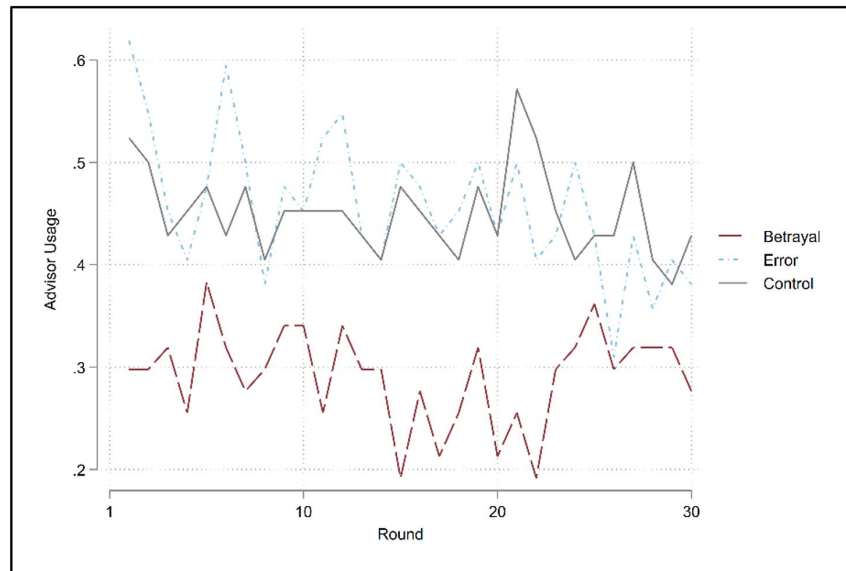


Figure 5a Human Usage Rates Across Treatments

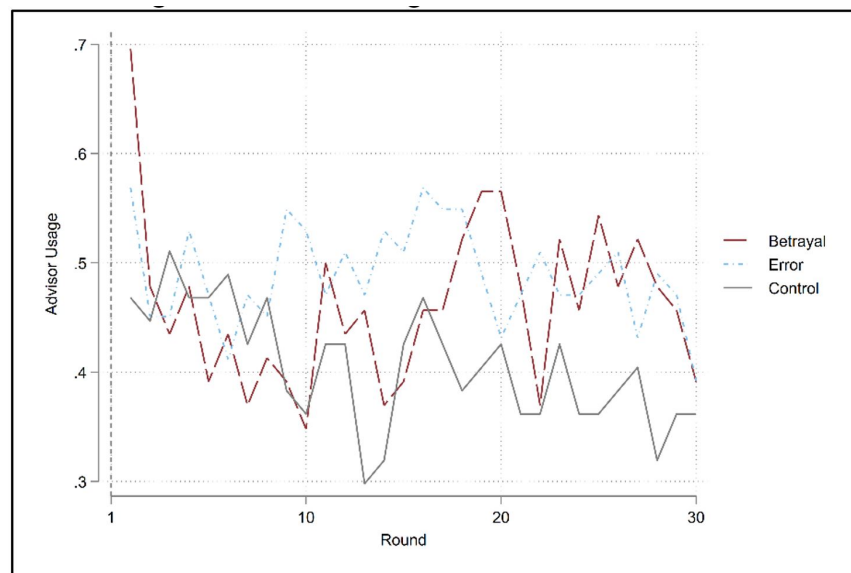


Figure 5b Algorithm Usage Rates Across Treatments

5.2.2. Investment Returns We estimate betrayal treatment effects for the total earnings received by each participant. The estimates in Table 6 present our findings. Our results indicate a \$2.15 decrease in total earnings for those in betrayal treatment with a human advisor ($p = 0.072$). We observe no significant effects of the error conditions for either advisor (Human $p = 0.195$; Algorithm $p = 0.281$) nor is there an effect of betrayal risk for the algorithm treatment ($p = 0.883$). Given that the average earnings for those in the control condition with a human advisor was \$10.29 (not including the \$5 show up fee), a decrease of \$2.15 would equate to a 21% drop in total earnings.

These findings highlight the real financial consequences of betrayal aversion, a clear antecedent to the acceptance of human expert advice.

Variables	Human (1) Earnings	Algorithm (2) Earnings
Betrayal Risk	-2.153* (1.186)	0.180 (1.223)
Error Risk	-1.502 (1.153)	-.118 (1.032)
Constant	10.29*** (0.933)	9.577*** (0.835)
Observations	131	144

Robust standard errors in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 6 Betrayal Effect on Total Earnings

5.2.3. Exit Questions In Table 7 we estimate the effect of our treatments on a number of exit questions related to feelings of being misled, having trust violated, and importantly, feeling betrayed. The questions were answered with a 5-point Likert scale where 1 = strongly disagree and 5 = strongly agree. Positive coefficients imply a stronger trend towards agreeing with the question. Column 1 reports estimates for responses to the question, “If you chose to invest with the advisor, you were concerned about being misled.” For the human condition, betrayal has a positive and statistically significant effect ($p = 0.006$). This effect is almost entirely cancelled out by the algorithm and betrayal interaction term ($p = 0.048$) implying that feelings of being misled from the betrayal risk were not a concern in the algorithm conditions. Column 2 reports estimates to the question asking if participants were concerned about their trust being violated. Again, we find similar results to the first question with the caveat that the algorithm interaction term is slightly insignificant ($p = 0.103$).

Our primary question of concern asked participants if they had concerns about feeling betrayed by the advisor. This was our main mechanism check given that the treatment does not explicitly use the word “betrayal” but implies a betrayal in the form of an incentive misalignment. The effect of the treatment in Column 3 is the largest coefficient ($p = 0.001$). This implies that our treatment worked as intended and elicited feelings of betrayal in participants. Additionally, the interaction term for the algorithm condition cancels out the treatment effect entirely ($p = 0.002$), again providing evidence that algorithms can attenuate betrayal aversion.

	(1)	(2)	(3)
Variables	Misled	Trust Violated	Feel Betrayed
Algorithm	0.245 (0.242)	0.0755 (0.236)	0.390 (0.250)
Betrayal Risk	0.671*** (0.240)	0.650** (0.252)	0.815*** (0.254)
Error Risk	0.0476 (0.256)	-0.0476 (0.249)	0.214 (0.264)
Algorithm \times Betrayal	-0.663** (0.333)	-0.568 (0.347)	-1.095*** (0.357)
Algorithm \times Error	0.0825 (0.339)	0.164 (0.338)	-0.185 (0.358)
Constant	3.095*** (0.183)	2.690*** (0.175)	2.738*** (0.177)
Observations	275	275	275

Robust standard errors in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 7 Exit Question Analysis

6. Robustness: Experiment 2

To validate the results of our first experiment, we conducted a condensed version of the investment game with a new sample: undergraduate students. Unlike the main study, we do not include an additional error treatment (only a betrayal risk treatment and the control). Finally, there is a within-subjects fee treatment that is introduced after the first 20 rounds which decreased advisor use to almost zero across conditions. Therefore, we chose to analyze only the first 20 rounds where there were no fees. All other procedures are identical to the main study.

There are several key benefits to this follow-up study. First, we utilize a university economics lab to recruit our sample. Researchers that recruit through this lab are prohibited from using any form of deception in their experiments. This is explicitly highlighted to participants, and they should therefore understand that any claims made throughout the experiment are true. This additional confidence in the veracity of information provided by the researchers help alleviate concerns around the believability of the information provided to participants in the first experiment. Additionally, we utilized a different human financial advisor to further provide robustness around advisor appearance and helps alleviate concerns of effects specific to a particular individual.

One hundred and twenty-three participants took part in our experiment. Table 8 shows the results of our panel random effects models that we used to estimate our treatment effects. Our effects from the main study are replicated here. Column 1 estimates a roughly 12% decrease in advisor

Variables	Human (1) Use	Algorithm (2) Use	All (3) Use	Human (4) Earnings	Algorithm (5) Earnings
Betrayal Risk	-0.123*	0.0774	-0.123*	-2.128***	-0.566
	(0.0712)	(0.0561)	(0.0708)	(0.709)	(0.651)
Algorithm			-0.118*		
			(0.0697)		
Algorithm × Betrayal			0.201**		
			(0.0902)		
Constant	0.423***	0.305***	0.423***	18.72***	17.62***
	(0.0560)	(0.0421)	(0.0557)	(0.593)	(0.555)
Number of ID	48	75	123	48	75

Robust standard errors in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 8 Betrayal Effect on Advisor Use and Earnings (Experiment 2)

usage when betrayal risk is introduced ($p = 0.091$). Conversely, we see no significant decrease in advisor usage when betrayal risk is introduced to the algorithm conditions ($p = 0.167$). Column 3 reports the interaction term between betrayal risk and the algorithm, confirming that switching the advisor from a human to an algorithm counteracts the measured betrayal aversion ($p = 0.026$).

Likewise, Column 4 shows that betrayal aversion decreased total earnings by roughly \$2.13 ($p = 0.003$), strikingly similar to the decrease found in the main study. Those in the algorithm conditions saw no such decrease in earnings ($p = 0.387$) (column (5)). The results from Experiment 2 confirm findings from Experiment 1 across the board. This level of robustness is difficult to attain even when replicating experimental procedures that draw from the same population. Provided that Experiment 2 drew from an entirely new population with large demographic differences (e.g., college-aged students versus a far older population), this is an impressive replication that provides significant richness to our findings.

7. Discussion

Our results show that while human experts elicit significant betrayal aversion, substituting those experts with algorithms may attenuate most, if not all, of the effect. Additionally, we highlight the real financial consequences of betrayal aversion and show a 20% decrease in earnings due to this phenomenon. Finally, we show that while algorithms may experience some initial betrayal aversion, it quickly subsides, whereas the aversion persists for human experts.

This work is not without limitations. First, we only focus on the context of financial investing. Future work could explore this avenue by studying contexts with higher stakes and exploring the

effect that stakes have on betrayal aversion and algorithm aversion. Moreover, determining if our findings are consistent across numerous settings would provide further insights. Additionally, the experiment is not able to measure individual level betrayal aversion due to the between-subjects design. Aimone et al. (2015) propose a betrayal aversion elicitation task that uses a within-subjects design to determine betrayal aversion at the individual level. Future research could adopt this framework to uncover heterogeneity around individual characteristics and their impact on betrayal aversion. Further, this paper only considers the role of betrayal aversion as it relates to algorithms. In reality, the choice to use an algorithm is likely influenced by a number of other factors including familiarity with the algorithm provider, transparency of the algorithm, autonomy over the algorithm, and others. Future research could incorporate these ideas and explore their initial effects and interactive effects with betrayal aversion on algorithm adoption and use.

Despite these limitations, our work has significant implications for research. Our findings highlight the value in looking to behavioral and experimental economics to uncover potential attributes of algorithmic tools that drive adherence to expert advice. Further, we show that studying the adoption of algorithms and the interactions that humans have with them should not be viewed as an isolated phenomenon. Instead, by comparing individual behavior with other humans versus an algorithm, we can better understand what mechanisms drive adoption. For example, one possible explanation for the attenuation of betrayal aversion by algorithms is that participants in the human condition were able to interact with their advisor. This element of social connection heightens trust, which subsequently increases concerns over betrayal. Researchers can use these findings to develop solutions to hesitation around accepting algorithmic tools that utilize the inherent strengths of the algorithm itself. Further, future research could extend this potential pathway by varying the levels of social connection that an individual has with an algorithm.

Finally, our findings highlight the need for behavioral economics to revisit previously well-accepted phenomena in light of advancing technologies. While the majority view of the literature is that inanimate objects can elicit betrayal aversion, we provide evidence that this may not always be the case. One potential reason for this deviation could be that past work considering inanimate objects and betrayal aversion focused exclusively on products meant to protect consumer's physical safety. It may be the case that while algorithms can attenuate betrayal aversion in a financial services setting, they may still elicit the fear when a consumer's physical outcomes are at risk. Future work can extend this thought by looking for instances of betrayal aversion with smart cities and Internet-of-Things devices that direct traffic or deploy emergency services, betrayal aversion elicited through online health portals, or the effect of algorithmic vaccine deployment on betrayal fears. Each of these pathways would prove fruitful in developing a deeper understanding of human decision-making and thought.

Industry could also benefit from the findings we present. Largely, firms across industries are adopting algorithmic solutions and determining which roles within their corporations could be augmented by algorithms. While efficiencies and improved operations play a large role in this consideration, acceptance from shareholders and customers is equally important. Our work provides valuable insight into the relationship humans have with algorithms and how that relationship leads to adoption and acceptance. For example, firms that provide face-to-face services that rely on consumer acceptance of their output may benefit from enhanced algorithmic participation, given the decrease in social expectations of algorithms from the consumer.

More specifically, financial services can utilize our results to improve adherence to financial planning. While firms are increasingly introducing algorithmic tools to industry customers, there is little work showing the value of these tools to individual consumers. By augmenting traditional financier involvement in strategic customer planning with advanced algorithmic tools, consumers may be more willing to accept the provided advice.

Overall, our results highlight the potential of algorithms to decrease the acceptance burden that many consumers face with “experts” that they encounter. Mitigating betrayal aversion can lead to increased demand for expert advice and better individual outcomes for consumers. The potential of algorithms from the perspective of efficiency and accuracy is only one piece of the larger puzzle. Algorithms may make individuals feel at ease and allow them to overlook some of the less impactful social rules that might have restricted their previous gains.

Appendix A: Risk Disclosures

Disclaimer for Betrayal Treatments (Verbiage in brackets refers to the human condition):

“In the past 10 rounds, you were given the option to use a *financial investment algorithm* [financial advisor, Brandon] to make your investment decisions. Historically, *the algorithm* [Brandon] has outperformed those who choose to invest themselves, where the majority of negative returns *it* [he] received were the result of random market volatility. However, *investment algorithms, like the one used in this experiment, are designed to earn revenue* [he is compensated] based on how frequently *it* [he] invests. Occasionally *the algorithm* [he] will intentionally over-invest even when *the algorithm* [he] is not confident that it is a good market. This incentive misalignment comes at the expense of increasing your risk and accounts for a small proportion of historical negative returns.”

Disclaimer for Error Treatments (Verbiage in brackets refers to the human condition):

“However, *investment algorithms, like the one used in this experiment,* [his recommendations] are not always perfect and *it* [he] has occasionally made accidental errors that result in a negative return. These errors come at the expense of increasing your risk and account for a small proportion of historical negative returns.”

Appendix B: Comprehension Quiz

Each question had the following possible answers: True, False

Question	Answer
You are being asked to invest your research points in a market with TWO possible assets.	False – there is only ONE asset.
You can invest 0% of your research points in the asset.	True.
If the CURRENT market is good, then there is a higher chance of the NEXT market being bad.	False – If the current market is good, there is a higher chance of the next market being good.
1000 research points translates to \$1 at the end of the experiment.	False – 1000 research points translates to \$5 at the end of the experiment.
If you have any questions throughout the experiment, you should ask them through Zoom only.	True.

Table 9

Appendix C: Waiting Page

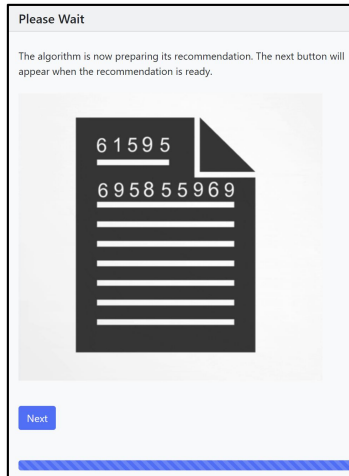


Figure 6a Algorithm

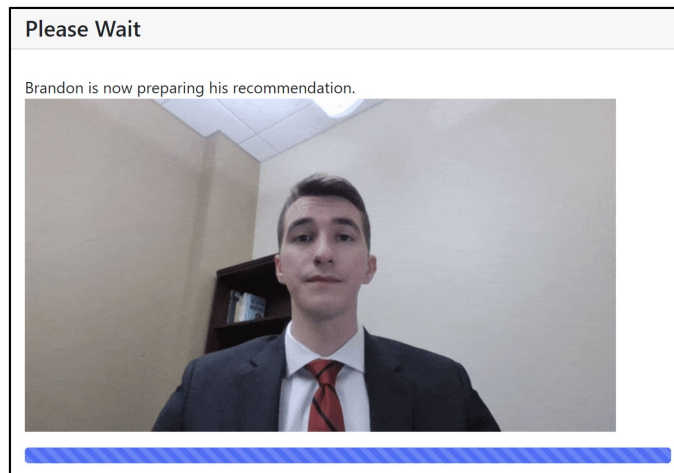


Figure 6b Human

Appendix D: Advisor Interface

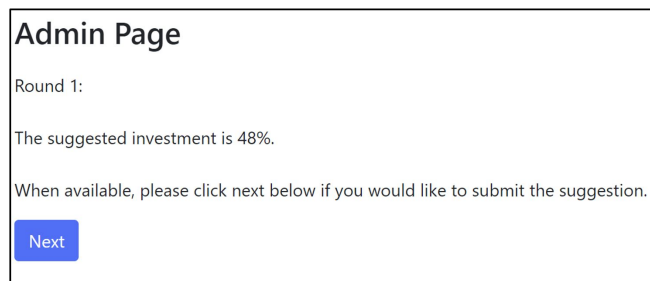


Figure 7

Appendix E: Exit Questions

Each question was answered using a Likert scale with the following responses: Strongly Agree, Agree, Neither Agree nor Disagree, Disagree.

Question

If I chose to invest with help, I was concerned about being misled.
If I chose to invest with help, I was concerned about my trust being violated.
If I chose to invest with help, I would feel betrayed if I received a negative return.
If I chose to invest with help, I trusted that a good investment decision would be made for me.
If I chose to invest with help, I felt like I made a mistake if I received a negative return.
If I chose to invest on my own, I felt like I made a mistake if I received a negative return.

Table 10

References

- Aimone J, Ball S, King-Casas B (2015) The betrayal aversion elicitation task: An individual level betrayal aversion measure. *PLoS ONE* 10(9):e0137491, URL <http://dx.doi.org/10.1371/journal.pone.0137491>.
- Al-Natour S, Benbasat I (2009) The adoption and use of it artifacts: A new interaction-centric model for the study of user-artifact relationships. *J. AIS* 10, URL <http://dx.doi.org/10.17705/1jais.00208>.
- Alexander V, Blinder C, Zak PJ (2018) Why trust an algorithm? performance, cognition, and neurophysiology. *Computers in Human Behavior* 89:279–288, ISSN 1873-7692(Electronic),0747-5632(Print), URL <http://dx.doi.org/10.1016/j.chb.2018.07.026>.
- Alt R, Beck R, Smits MT (2018) Fintech and the transformation of the financial industry. *Electronic Markets* 28(3):235–243, ISSN 1422-8890, URL <http://dx.doi.org/10.1007/s12525-018-0310-9>.
- Arrow KJ (1971) The theory of risk aversion. *Essays in the theory of risk-bearing* 90–120.
- Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, van de Vijver MJ, West RB, van de Rijn M, Koller D (2011) Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci Transl Med* 3(108):108ra113, ISSN 1946-6234, URL <http://dx.doi.org/10.1126/scitranslmed.3002564>.
- Bell D, Gana L (2012) Algorithmic trading systems: A multifaceted view of adoption. *2012 45th Hawaii International Conference on System Sciences*, 3090–3099 (IEEE), ISBN 1457719258.
- Benbasat I, Wang W (2005) Trust in and adoption of online recommendation agents. *J. AIS* 6, URL <http://dx.doi.org/10.17705/1jais.00065>.
- Berger B, Adam M, Rühr A, Benlian A (2021) Watch me improve—algorithm aversion and demonstrating the ability to learn. *Business & Information Systems Engineering* 63(1):55–68, ISSN 1867-0202, URL <http://dx.doi.org/10.1007/s12599-020-00678-5>.

- Bernard C, Chen JS, Vanduffel S (2015) Rationalizing investors' choices. *Journal of Mathematical Economics* 59:10–23, ISSN 0304-4068.
- Birnberg J, Zhang Y (2010) When betrayal aversion meets loss aversion: The effects of changes in economic conditions on internal control system choices. *Journal of Management Accounting Research* 23, URL <http://dx.doi.org/10.2308/jmar-10087>.
- Bohnet I, Greig F, Herrmann B, Zeckhauser R (2008) Betrayal aversion: Evidence from brazil, china, oman, switzerland, turkey, and the united states. *American Economic Review* 98(1):294–310, URL <http://dx.doi.org/10.1257/aer.98.1.294>.
- Bohnet I, Herrmann B, Zeckhauser R (2010) Trust and the reference points for trustworthiness in gulf and western countries. *The Quarterly Journal of Economics* 125(2):811–828, ISSN 1531-4650.
- Bohnet I, Zeckhauser R (2004) Trust, risk and betrayal. *Journal of Economic Behavior & Organization* 55(4):467–484, ISSN 1879-1751(Electronic),0167-2681(Print), URL <http://dx.doi.org/10.1016/j.jebo.2003.11.004>.
- Brogaard J, Hendershott T, Riordan R (2014) High-frequency trading and price discovery. *The Review of Financial Studies* 27(8):2267–2306, ISSN 1465-7368.
- Burton JW, Stein MK, Jensen TB (2020) A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* 33(2):220–239, ISSN 0894-3257, URL <http://dx.doi.org/https://doi.org/10.1002/bdm.2155>.
- Cao L, Yuan G, Leung T, Zhang W (2020) Special issue on ai and fintech: the challenge ahead. *IEEE Intelligent Systems* 35(2):3–6, ISSN 1541-1672.
- Castelo N, Bos MW, Lehmann DR (2019) Task-dependent algorithm aversion. *Journal of Marketing Research* 56(5):809–825, ISSN 0022-2437, URL <http://dx.doi.org/10.1177/0022243719851788>.
- Chakrabarty B, Moulton P, Wang X (2015) Attention effects in a high-frequency world. *Working Paper* .
- Chen DL, Schonger M, Wickens C (2016) otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance* 9:88–97, ISSN 2214-6350, URL <http://dx.doi.org/https://doi.org/10.1016/j.jbef.2015.12.001>.
- Colarelli SM, Thompson M (2008) Stubborn reliance on human nature in employee selection: Statistical decision aids are evolutionarily novel. *Industrial and Organizational Psychology* 1(3):347–351, ISSN 1754-9426, URL <http://dx.doi.org/https://doi.org/10.1111/j.1754-9434.2008.00060.x>.
- Dawes RM (1979) The robust beauty of improper linear models in decision making. *American Psychologist* 34(7):571–582, ISSN 1935-990X(Electronic),0003-066X(Print), URL <http://dx.doi.org/10.1037/0003-066X.34.7.571>.
- Diab D, Pui SY, Yankelevich M, Highhouse S (2011) Lay perceptions of selection decision aids in us and non-us samples. *International Journal of Selection and Assessment* 19, URL <http://dx.doi.org/10.1111/j.1468-2389.2011.00548.x>.

- Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144(1):114–126, ISSN 1939-2222(Electronic),0096-3445(Print), URL <http://dx.doi.org/10.1037/xge0000033>.
- Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64(3):1155–1170, URL <http://dx.doi.org/10.1287/mnsc.2016.2643>.
- Eastwood J, Snook B, Luther K (2012) What people want from their professionals: Attitudes toward decision-making strategies. *Journal of Behavioral Decision Making* 25, URL <http://dx.doi.org/10.1002/bdm.741>.
- Filiz I, Judek JR, Lorenz M, Spiwoks M (2021) Reducing algorithm aversion through experience. *Journal of Behavioral and Experimental Finance* 100524, ISSN 2214-6350, URL <http://dx.doi.org/https://doi.org/10.1016/j.jbef.2021.100524>.
- Frey CB, Osborne MA (2017) The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change* 114:254–280, ISSN 0040-1625, URL <http://dx.doi.org/https://doi.org/10.1016/j.techfore.2016.08.019>.
- Ge R, Zheng Z, Tian X, Liao L (2021) Human–robot interaction: When investors adjust the usage of robo-advisors in peer-to-peer lending. *Information Systems Research* ISSN 1047-7047.
- Gomber P, Kauffman R, Parker C, Weber B (2018) Special issue: Financial information systems and the fintech revolution. *Journal of Management Information Systems* 35(1):12–18.
- Grégoire Y, Fisher RJ (2008) Customer betrayal and retaliation: when your best customers become your worst enemies. *Journal of the Academy of Marketing Science* 36(2):247–261, ISSN 1552-7824, URL <http://dx.doi.org/10.1007/s11747-007-0054-0>.
- Hendershott T, Jones CM, Menkveld AJ (2011) Does algorithmic trading improve liquidity? *The Journal of Finance* 66(1):1–33, ISSN 0022-1082, URL <http://dx.doi.org/https://doi.org/10.1111/j.1540-6261.2010.01624.x>.
- Hendershott T, Zhang X, Zhao JL, Zheng Z (2021) Fintech as a game changer: Overview of research frontiers. *Information Systems Research* 32(1):1–17, ISSN 1047-7047, URL <http://dx.doi.org/10.1287/isre.2021.0997>.
- Highhouse S (2008) Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice* 1(3):333–342, ISSN 1754-9434(Electronic),1754-9426(Print), URL <http://dx.doi.org/10.1111/j.1754-9434.2008.00058.x>.
- Holt CA, Laury SK (2002) Risk aversion and incentive effects. *The American Economic Review* 92(5):1644–1655, ISSN 00028282, URL <http://www.jstor.org/stable/3083270>.
- Hong K, Bohnet I (2007) Status and distrust: The relevance of inequality and betrayal aversion. *Journal of Economic Psychology* 28(2):197–213, ISSN 0167-4870, URL <http://dx.doi.org/https://doi.org/10.1016/j.joep.2006.06.003>.

- Jussupow E, Spohrer K, Heinzl A, Gawlitza J (2021) Augmenting medical diagnosis decisions? an investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research* ISSN 1047-7047.
- Koehler JJ, Gershoff AD (2003) Betrayal aversion: When agents of protection become agents of harm. *Organizational Behavior and Human Decision Processes* 90(2):244–261, ISSN 0749-5978, URL [http://dx.doi.org/https://doi.org/10.1016/S0749-5978\(02\)00518-6](http://dx.doi.org/https://doi.org/10.1016/S0749-5978(02)00518-6).
- Kou G (2019) Introduction to the special issue on fintech. *Financial Innovation* 5(1):45, ISSN 2199-4730, URL <http://dx.doi.org/10.1186/s40854-019-0161-1>.
- Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151:90–103, ISSN 0749-5978, URL <http://dx.doi.org/https://doi.org/10.1016/j.obhdp.2018.12.005>.
- Longoni C, Bonezzi A, Morewedge CK (2019) Resistance to medical artificial intelligence. *Journal of Consumer Research* 46(4):629–650, ISSN 0093-5301.
- Lourenço CJS, Dellaert BGC, Donkers B (2020) Whose algorithm says so: The relationships between type of firm, perceptions of trust and expertise, and the acceptance of financial robo-advice. *Journal of Interactive Marketing* 49:107–124, ISSN 1094-9968, URL <http://dx.doi.org/https://doi.org/10.1016/j.intmar.2019.10.003>.
- Nass C, Moon Y (2000) Machines and mindlessness: Social responses to computers. *Journal of Social Issues* 56(1):81–103, ISSN 1540-4560(Electronic),0022-4537(Print), URL <http://dx.doi.org/10.1111/0022-4537.00153>.
- Nass C, Steuer J, Tauber ER (1994) Computers are social actors. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 72–78.
- Orlikowski WJ, Scott SV (2015) The algorithm and the crowd: Considering the materiality of service innovation. *MIS Q.* 39:201–216.
- Parker G, Van Alstyne M, Jiang X (2017) Platform ecosystems:: How deve/opers invert the firm. *MIS Quarterly* 41(1):255–266, ISSN 2162-9730.
- Prahl A, Van Swol L (2017) Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting* 36(6):691–702, ISSN 0277-6693, URL <http://dx.doi.org/https://doi.org/10.1002/for.2464>.
- Pratt JW (1964) Risk aversion in the small and in the large. *Econometrica* 32(1/2):122–136, ISSN 00129682, 14680262, URL <http://dx.doi.org/10.2307/1913738>.
- Qiu L, Benbasat I (2005) Online consumer trust and live help interfaces: The effects of text-to-speech voice and three-dimensional avatars. *International Journal of Human-Computer Interaction* 19(1):75–94, ISSN 1532-7590(Electronic),1044-7318(Print), URL http://dx.doi.org/10.1207/s15327590ijhc1901_6.

- Rachman S (2010) Betrayal: A psychological analysis. *Behaviour Research and Therapy* 48(4):304–311, ISSN 0005-7967, URL <http://dx.doi.org/https://doi.org/10.1016/j.brat.2009.12.002>.
- Rafaeli A, Altman D, Gremler DD, Huang MH, Grewal D, Iyer B, Parasuraman A, de Ruyter K (2016) The future of frontline research: Invited commentaries. *Journal of Service Research* 20(1):91–99, ISSN 1094-6705, URL <http://dx.doi.org/10.1177/1094670516679275>.
- Ransbotham S, Gerbert P, Reeves M, Kiron D, Spira M (2018) Artificial intelligence in business gets real. Report, MIT Sloan Management Review.
- Reeves B, Nass CI (1996) *The media equation: How people treat computers, television, and new media like real people and places*. The media equation: How people treat computers, television, and new media like real people and places. (New York, NY, US: Cambridge University Press), ISBN 1-57586-052-X (Hardcover).
- Schanke S, Burtch G, Ray G (2021) Estimating the impact of “humanizing” customer service chatbots. *Information Systems Research* 32(3):736–751, ISSN 1047-7047, URL <http://dx.doi.org/10.1287/isre.2021.1015>.
- Scherer LD, de Vries M, Zikmund-Fisher BJ, Witteman HO, Fagerlin A (2015) Trust in deliberation: The consequences of deliberative decision strategies for medical decisions. *Health Psychol* 34(11):1090–9, ISSN 0278-6133, URL <http://dx.doi.org/10.1037/hea0000203>.
- Seiders K, Flynn AG, Berry LL, Haws KL (2015) Motivating customers to adhere to expert advice in professional services: a medical service context. *Journal of Service Research* 18(1):39–58, ISSN 1094-6705.
- Shnoor B (2009) Loss of chance: A behavioral analysis of the difference between medical negligence and toxic torts. *Am. J. Trial Advoc.* 33:71.
- Sparrow B, Liu J, Wegner Daniel M (2011) Google effects on memory: Cognitive consequences of having information at our fingertips. *Science* 333(6043):776–778, URL <http://dx.doi.org/10.1126/science.1207745>.
- Tetlock PE, Gardner D (2016) *Superforecasting: The art and science of prediction* (Random House), ISBN 1847947158.
- Van Swol LM (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, ISSN 0169-2070.
- Venkatesh V, Davis F (2000) A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science* 46:186–204, URL <http://dx.doi.org/10.1287/mnsc.46.2.186.11926>.
- Wilson H, Daugherty P, Bianzino N (2017) When ai becomes the new face of your brand. *Harvard Business Review* 27.

Woodhouse EJ, Nieuwsma D (1997) When expert advice works, and when it does not. *IEEE Technology and Society Magazine* 16(1):23–29, ISSN 0278-0097.

Zhang X (2020) *Experiential and Neurobiological Influences on Economic Preferences and Risky Decision Making*. Thesis, Virginia Tech.

Önköl D, Goodwin P, Thomson M, Gönöl S, 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, ISSN 1099-0771(Electronic),0894-3257(Print), URL <http://dx.doi.org/10.1002/bdm.637>.