UNIVERSITÉ CÔTE D'AZUR

ALGORITHM CREDULITY: HUMAN AND ALGORITHMIC ADVICE IN PREDICTION EXPERIMENTS

Documents de travail GREDEG
GREDEG Working Papers Series

Mathieu Chevrier
Brice Corgnet
Eric Guerci
Julie Rosaz

GREDEG WP No. 2024-03
https://ideas.repec.org/s/gre/wpaper.html

Algorithm Credulity
Human and Algorithmic Advice in Prediction Experiments

Mathieu Chevrier‡, Brice Corgnet†, Eric Guerci⁑ and Julie Rosaz*

GREDEG Working Paper No. 2024-03
This version: December 2024

ABSTRACT

This study examines algorithm credulity by which people rely on faulty algorithmic advice without critical evaluation. Using a prediction task comparing human and algorithm advisors, we find that participants are more likely to follow the same deficient advice when issued by an algorithm than by a human. We show that algorithm credulity reduces expected earnings by 13%. To explain this finding, we posit that people are more likely to perceive as credible an unpredictable and deficient piece of advice when produced by an algorithm than by a human. Overall, our results imply that humans might be particularly susceptible to the influence of deficient algorithmic advice.

JEL CODES: C92; D91.

Keywords: Algorithm credulity, algorithmic advice, artificial intelligence, laboratory experiments.

This work has been supported by the French government, through the UCAJEDI and EUR DS4H Investments in the Future projects managed by the National Research Agency (ANR) with the reference number ANR-15-IDEX-0001 and ANR-17-EURE-0004 and by Région Bourgogne-Franche-Comté, through IACITE project n°2022-Y-13758. Brice Corgnet acknowledges that this research was performed within the framework of the LABEX CORTEX (ANR-11-IDEX-007) operated by the French National Research Agency.
‡Université Côte d'Azur, CNRS, GREDEG, France. Email : mathieu.chevrier@univ-cotedazur.fr
†Emlyon business school, GATE UMR 5824, F-69130 Ecully, France. Email : corgnet@em-lyon.com
⁑ Université Côte d'Azur, CNRS, GREDEG, France. Email : eric.guerci@univ-cotedazur.fr
*CEREN EA 7477, Burgundy School of Business, Université Bourgogne Franche-Comté, Dijon, France. Email : julie.rosaz@bsb-education.com

1. Introduction

Humans are not credulous by nature. Throughout evolution, we have developed mechanisms of epistemic vigilance, enabling us to preempt deceptive attempts by, for example, deciphering others' intentions and probing their logical reasoning (see Sperber et al., 2010; Mercier, 2020).

Yet, the advent of artificial intelligence (Trajtenberg, 2018) poses new threats to our ability to detect faulty advice. Because algorithms do not have intentions (Langley et al., 2022) and because their codes are often difficult to decipher (Haibe-Kains et al., 2020), the mechanisms of epistemic vigilance are likely to be limited when humans interact with machines.[1] These limitations will give rise to the phenomenon of *algorithm credulity*, which we define as the tendency to rely on faulty algorithmic advice with less critical evaluation than for the same advice issued by a human. Anecdotal evidence of algorithm credulity is legion. One news story recounts how a person intending to pick up a friend in Brussels, located less than 150 km from her residence, found herself more than 1,500 km away in Zagreb (Croatia) due to a GPS error before correcting her course (Grenoble, 2013). Similar stories have been reported, such as a man submerging his car in an Alaskan harbor after blindly following GPS instructions (Murphy, 2012).

Nonetheless, the scientific study of algorithm credulity is scant (see Aroyo et al., 2021; Krügel et al., 2022, 2023), despite being the focus of new regulations (Zuiderveen Borgesius, 2018). We are not aware of any study demonstrating that people are more likely to follow hurtful algorithmic advice than the same advice given by a human. Our first research question is, therefore, to ask whether the phenomenon of algorithm credulity is real. Can it be observed in a controlled experiment? If so, under what conditions do individuals blindly follow algorithmic advice?

To investigate algorithm credulity, we posit that individuals are more likely to perceive random advice as more credible when dispensed by an algorithm than by a human. This is the case because advice that is highly random leads to a high level of uncertainty, which is a key characteristic of complexity (Shannon, 1948; Caplin et al., 2022; Mononen, 2023). Because complexity is perceived as inherent to algorithms (Lehmann et al., 2022), random advice is likely to be perceived as more credible when issued by an algorithm. The same undecipherable advice issue by a human is more likely to be perceived as deception attempt (Mercier, 2020).

To test for algorithm credulity model, we deploy an incentivized prediction task in a between-subject laboratory design, which resembles commonly used tasks in the literature (see Dietvorst et al., 2014; Logg et al., 2019; Lehmann et al., 2022). Our task consists in predicting the color of a ball, black or white, drawn at random with replacement from an urn. We use an abstract task purposefully to isolate the underlying mechanisms driving algorithm credulity.

[1] When asked "What are your motivations?", ChatGPT answered "As an artificial intelligence I don't possess personal motivations, emotions, or consciousness." on February 2, 2024.

Not only forecasting tasks capture critical features of actual jobs (Lehmann et al., 2022) but they also capture essential features of human decision-making.

In the Human (Algo) treatment, another human participant (an algo) makes a color prediction that can be either followed or ignored. Participants are not told how the advice series are produced but the 60 repetitions of the prediction task give extensive opportunities for learning by experience. In our design, we could compare how people responded to the same piece of advice when issued by a human or an algo, thus allowing us to control for people's natural tendency to discount advice relative to their own judgement (see Woolley and Risen, 2018). We also vary the type of advice — Optimal, Biased, or Random. In addition to the No-Advice baseline, we thus administered the following six treatments: Human-Optimal, Algo-Optimal, Human-Bias, Algo-Bias, Human-Random, and Algo-Random. To make sure Human and Algo treatments were directly comparable, we ensured participants received the exact same series of advice for a given pair of treatments.

Consistent with algorithm credulity, people were more likely to perceive a deficient algorithmic advisor to be credible, thus being more likely to follow its advice than they were to follow a deficient human advisor, and this effect occurred despite extensive opportunities for learning. Random pieces of advice were perceived as more credible when issued by an algo than by a human. To our knowledge, our study is the first to offer causal evidence for algorithm credulity and to identify the underlying mechanisms behind algorithm credulity phenomenon. Our findings extend the work Lehmann et al. (2022) who showed that algorithms that are perceived as excessively complex continue to be used. In our work, we show that the former phenomenon can trigger algorithm credulity, where deficient advice is more likely to be adopted when issued by an algorithm than by a human. Our findings carry significant implications. Instead of systematically safeguarding human decision makers from deficient advice, algorithms can amplify this risk by promoting the uncritical acceptance of unnecessarily complex recommendations.

2. Literature review

2.1. Prediction tasks and human biases

We study algorithmic advice in the context of a prediction task because forecasting skills are critical to human decision making in a broad range of activities (Fildes et al., 2009; Önkal et al., 2009; Zellner et al., 2021). In prediction tasks, advice, whether it is algorithmic or human, is potentially beneficial because human predictions are often biased. This is especially the case when predicting variables that lack regularity because humans have a natural inclination to

search for patterns even when none exists (Huettel et al., 2002; Pinker, 2022). Searching for patterns can be seen as a direct implication of the law of small numbers (Tversky and Kahneman, 1971; Rabin, 2002) according to which people mistakenly believe that a small sample should accurately reflect the underlying distribution. This leads lay people and even experts to overestimate the predictability of random series (Tetlock, 2006; Liang 2019; Pinker, 2022). Relying on the law of small numbers leads people to fall prey to the "gambler's fallacy" in which they mistakenly believe that if a particular outcome has occurred more frequently at the beginning of a series, it is less likely to occur next (Clotfelter and Cook, 1993; Rabin, 2002; Ayton and Fischer, 2005; Croson and Sundali, 2005; Dohmen et al., 2009; Huber et al., 2010). Another prominent bias in forecasting multiple draws of a random variable relates to "probability matching" (Fisher, 1930) according to which people tend to produce a forecasting series that resembles the underlying distribution of the random variable. For instance, consider a person asked to make a series of bets. The person can bet on whether the roll of a fair die will land on 6 (with a chance of occurrence $\frac{1}{6}$) or on any of the other five numbers (with a chance of occurrence $\frac{5}{6}$). In this scenario, one should always bet on the latter event. However, a person following probability matching will predict that the die will land on 6 in one out of every six bets.

Using our prediction task, we will evaluate whether algorithmic advice exacerbates, by inducing credulity, rather than tames well-known human biases.

2.2. Algorithmic advice in prediction tasks

Because algorithms are based on preset rules that can prevent biases and lead to optimal choices based on Bayesian inference (Meehl, 1954; Dawes, 1979), they offer an opportunity to debias human forecasting, which has traditionally been found to be difficult (Fischhoff, 1982; Larrick, 2004). As advanced by Edwards and Fasolo (2001), computer-based "decision tools will be as important in the 21st century as spreadsheets were in the 20th" (p. 581). However, despite their superior performance, evidence has shown that people tend to prefer human to algorithmic advice (Dzindolet et al., 2002; Önkal et al, 2009; Dietvorst et al., 2014; Castelo et al., 2019; Jung and Seiter, 2021; see reviews in Burton et al., 2020; Jussupow and Benbasat, 2021; Mahmud et al, 2022). In Dietvorst et al. (2014), algorithm aversion may have emerged because people perceive algorithmic errors as systematic, whereas human errors are seen as opportunities for learning and developing a better intuition about the task. This mechanism leads people to be more inclined to follow a human than an algorithmic advice after observing

a mistaken advice (Madhavan and Wiegmann, 2007; Highhouse, 2008; Prahl and Van Swol, 2017; Burton et al., 2020). However, in the absence of feedback about potential mistakes, Logg et al. (2019) have shown that people prefer to follow algorithmic advice over human advice, thus coining the term of algorithm appreciation. Algorithm appreciation applies to a wide range of situations such as the perception of visual stimuli and forecasts regarding song popularity and romantic attraction. Algorithm appreciation only vanishes when people can use their own estimate instead of that of advisors and when people are experts (Logg et al., 2019).

Our study of algorithmic advice naturally connects to the growing experimental literature on the use of automated systems in work-related decisions (e.g., Alekseev, 2020; Dargnies et al., 2022; Fumagalli et al., 2022; Corgnet, 2023; Corgnet et al., 2023; see also Chugunova and Sele, 2022; March 2021 for reviews). For example, Fumagalli et al. (2022) use experiments to evaluate workers' preferences for human and algorithmic recruiters. They show that people rate human and algorithmic recruiters differently even when they use the exact same information for their hiring decisions. Low performers prefer human recruiters whereas high performers prefer algorithmic recruiters because they are perceived as giving more weight on task performance. Using a real-effort online experiment, Dargnies et al. (2022) also report that workers generally prefer to be evaluated by a human manager than an algorithm despite its superior ability in selecting the best employee. Managers also prefer to make the hiring decision themselves rather than delegating it to an algorithm. Furthermore, this aversion to the algorithmic recruiter continues to be observed even when the details of the automated decision are disclosed. That said, the authors show that the use of algorithmic recruiting could be enhanced when extensive feedback is provided on the relatively poor performance of human managers in hiring the best worker. Interestingly, in the context of dismissal decisions, Corgnet (2023) find that individuals are more inclined to favor algorithmic decisions due to their perceived fairness.

The existing literature thus suggests that people generally prefer human to algorithmic advice unless there exists extensive feedback on the superior performance of algorithms over humans. Beyond feedback, another important moderating dimension of algorithm aversion is the nature of the task. Algorithms will be trusted more for machine-like tasks that do not involve subjective evaluations and only require computational skills and statistical knowledge such as when a GPS device calculates the optimal route between two locations (Castelo et al., 2019).

In this study, we posit that people can follow algorithmic advice more than identical human advice even when the advice is deficient, and feedback is extensive. We argue that it is

not feedback per se but perceived credibility of the advisor that will trigger the widespread adoption of algorithmic advice (French and Raven, 1959; see also Birnbaum and Stegner, 1979, Muir, 1994; Muir and Moray, 1996; Liu, 2010; Bonaccio and Dalal, 2006, Bailey et al., 2022).

3. Design

3.1. Prediction Task

We implement a prediction task which consists in predicting the color of the ball that is randomly drawn by a computer from an urn with black and white balls (see Instructions in Appendix A.1). In our task, the composition of the urn is known and displayed on participants' screens. In the first (last) 30 rounds, 55% of the balls (11 out of 20) in the urn are white (black). Before participants start the 60 rounds of the experiments, they complete 5 unpaid practice rounds. At the end of each round, a ball is randomly drawn from the urn with replacement. For the sake of comparability across treatments, we randomly drew 4 series before conducting the experiments and used these preset series for the 4 independent sessions for each treatment (see Table A1 in Appendix A and Figure 1). In the first (last) 30 rounds, the urn is composed of 11 (9) white balls and 9 (11) black balls. After making an initial prediction, participants either received a recommendation from a human or algorithmic advisor (advice treatments) or made their final choice without any recommendation (baseline) (see Figure 2). At the end of each round, participants received feedback about the actual draw as well as their initial choice, the advisor's recommendation (in advice treatments) and their final choice. The feedback screen explicitly reported whether the advice was correct or not.

We purposefully chose a quantitative task that does not involve subjective evaluations and only require computational skills and statistical knowledge for two reasons. This type of tasks is more likely to alleviate algorithm aversion and is therefore an adequate testbed for identifying algorithm credulity (Castelo et al., 2019; Hertz and Wiese, 2019). Second, this simple task has an optimal rule—always select the more prevalent ball in the urn—that is easy to identify and not cognitively demanding to implement. Nonetheless, as explained in Section 2.1, participants may exhibit cognitive biases that prevent them from easily exploiting the optimal strategy. Although simple, the task remains rich in terms of possible strategies employed by the participants.

3.2. Treatments

3.2.1. Three types of advice (optimal, biased and random)

We used a between-subject experimental design with 7 treatments that differ in the presence and type of advice. In the No-Advice baseline, participants received no recommendation. In the advice treatments, participants received an advice ('White' or 'Black' ball) after making their initial choice but before making their final decision. Participants did not know the advice rules and when they applied.

In Optimal treatments, the recommendation of the advisor was always to pick the ball color that appeared most frequently in the urn, thus leading to a correct answer in 57.1% of the rounds on average. We are thus in a case in which the optimal advice is often wrong.

In Biased treatments, the advice series reflected common biases observed in human choices in predictions tasks: the gambler's fallacy (Clotfelter and Cook, 1993; Rabin, 2002; Ayton and Fischer, 2005; Croson and Sundali, 2005; Dohmen et al., 2009; Huber et al., 2010) and probability matching (Fisher, 1930; Nowak and Sigmund, 1993; Gaissmaier and Schooler, 2008).

The advising rule based on probability matching recommended to bet on the ball color that was drawn in the previous round. In a long series, the draws would reflect the underlying distribution of ball colors so that the advice series would also reflect the underlying distribution in line with probability matching.

The advising rule based on the gambler's fallacy recommended the ball color that was *not* drawn in the previous round. This simple advising rule captures the essence of the gambler's fallacy according to which streaks are seen as less likely than alternating outcomes. This rule applies in Rounds 16 to 45 in Series 1 and 2, as well as in Rounds 1 to 15 and 46 to 60 in Series 3 and 4. When the gambler's fallacy rule did not apply, the probability matching rule applied so that both rules were used as frequently (see Figure 1).

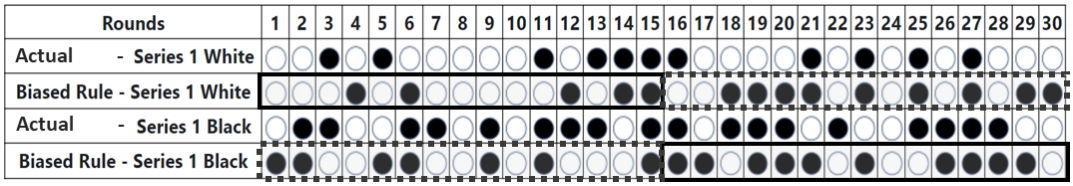


Figure 1. *Biased advice for Series 1*. The first (last) 30 rounds used the urn containing more white (black) balls. The probability matching rule applied during the first 15 rounds of 'Series 1 White', and the last 15 rounds of 'Series 1 Black' (see ☐ frame). The gambler's fallacy applies during the last 15 rounds of 'Series 1 White', and the first 15 rounds of 'Series 1 Black' (see ⬚ frame). Actual draws for Series 1 are also shown in rows 2 and 4 of the figure.

The biased advice was correct in 55.4% of the rounds on average. Biased treatments allowed us to study the extent to which participants followed a deficient advice, which was when the 'Black' ('White') ball was recommended in any of the first (last) 30 rounds. One critical

difference between Biased and Optimal treatments was that the biased advice series was characterized by a high variety in recommendations whereas the optimal advice always recommended betting on the 'White' ball in the first 30 rounds and on the 'Black' ball in the last 30 rounds. This means, that the optimal advice series was perfectly predictable, and in information-theoretic terms, this implies the Shannon entropy of the advice series was equal to zero (Shannon, 1948). In the case of the biased advice, the series exhibited more variety in recommendations and was thus less predictable. However, once the underlying rule was learned, the biased series was perfectly predictable based on the draw of the previous ball.

In Random treatments, we produced an advice series that was fully random and thus not predictable based on previous draws. To construct the sequence of random advice, we maximized the entropy of the series, that is its unpredictability (Gauvrit et al., 2017).[2] Although the actual level of Shannon entropy was similar in Random and in Biased treatments, the random series of advice was unpredictable based on observed draws and could not be learned over the course of the experiment. The random advice provided the correct prediction in 54.3% of the cases, similarly to the biased advice.

3.2.2. Two types of advisors (human and algo)

For each of the three types of advice (Optimal, Biased, and Random), we considered two types of advisors (Human and Algo), thus leaving us with a 3×2 between-subject factorial design with the following advice treatments in addition to the No-Advice baseline: Algo-Optimal, Human-Optimal, Algo-Bias, Human-Bias, Algo-Random, and Human-Random (see Figure 2).

To ensure that the Human and Algo treatments were directly comparable, we recruited 12 human advisors, with 4 advisors (one per series) for each of the 3 types of advice. Advisors were recruited in a separate session that took place before the main experiment. When making predictions in the task, they were incentivized to produce the exact same series as the one generated by the corresponding algorithmic advice. They received a 5-euro participation bonus for picking the suggested color ball advice in each round, and all did so. In our design, human and algorithmic advice are thus identical.[3]

---

[3] Unlike the standard "Wizard of Oz" technique in human-computer interaction research (Cross and Ramsey, 2021), we relied on actual humans and algorithms. However, our approach is not exempt of deception by omission as we do not tell participants that the human advice was produced using a specific incentivized rule. Deception by omission is often deemed acceptable in experimental economics (Hey, 1998; Wilson, 2016), especially when it is unlikely to impact the reputation of the lab and participants' behavior in future participations, which is the case in our current study (Davis and Holt, 1993). Yet, opinions diverge, and this remains a gray area (Charness et al., 2022).

In the instructions, algorithmic advisors were described as follows (see Appendix A.1): "*The algorithmic advisor is a computer program designed to assist you in making a prediction*." We did not provide a detailed description of the algorithm but provided ample opportunities to learn the effective rule used by the advisor in the 60 rounds of the experiment. Human advisors were described as follows: "*The advisor is another participant who completed a previous session of a similar prediction task. The human advisor aim is to assist you in making a prediction."*

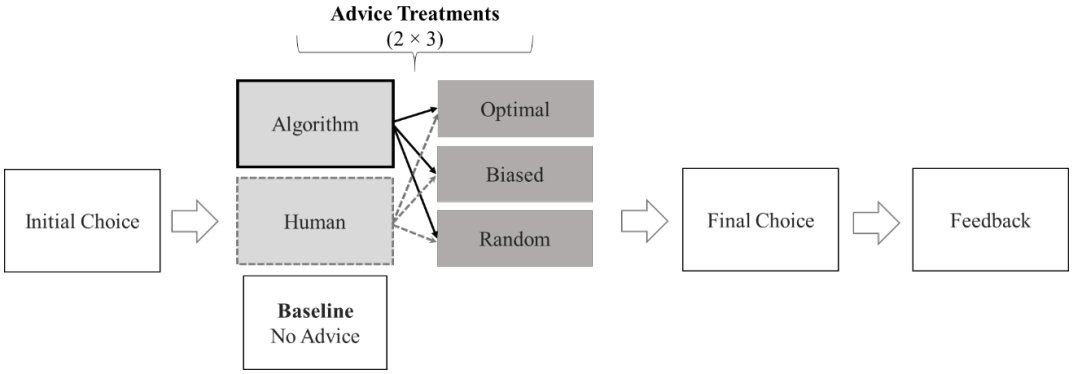


Figure 2. *Sequence of events during the prediction task across advice treatments and baseline for a single round*. Each participant makes an initial choice before receiving advice (in the advice treatments) or not (in the baseline). Each participant is assigned to one of the six possible treatments and a baseline for all 60 rounds of the experiment. After observing the advice (advice treatments) or not (baseline), participants make their final choice in which they can modify their initial choice. A ball is then drawn at random from the urn (with replacement) and feedback is received by participants.

3.3. Lab questionnaire

At the end of all treatments, we elicited cognitive reflection skills adapting the Cognitive Reflection Test (CRT, henceforth, see Frederick, 2005; Toplak et al., 2014; Primi et al., 2016), and risk attitudes (Holt and Laury, 2002). These test scores will serve as controls in our regression analyses (see Appendix A.1, Part 2). At the end of the Random treatments and the Human-Bias treatment, we elicited additional Likert scales to assess participants' perceptions about the advice (see Appendix A.1, Part 3).

3.4. Online questionnaire and protocol

When registering online for the lab experiment, participants had to complete demographic questions regarding age, gender, and education along with additional questionnaires (see Appendix A.2). They completed the NARS (Negative Attitudes Toward Robots scales, Nomura et al., 2006) for which the three subscales (Interactions, Social Influence and Emotions) showed acceptable reliability (Cronbach's $\alpha > 0.70$, see Appendix A.2). They also completed the TAM 2 (Technology Acceptance Model 2, Venkatesh and Davis, 2000). However, this scale exhibited

limited reliability (see Appendix A.2) so that we used NARS scores rather than TAM 2 as an individual control in our regression analyses.

In addition to NARS, CRT scores, and attitudes toward risk (number of safe choices in the Holt and Laury (2002) task), we use education and gender as additional controls. We do not include age as a control because our student sample showed limited variance in this measure, and the correlation between age and education level was substantial ($\rho = 0.67$). The online questionnaire also elicited the Belief in Good Luck Scale (Darke and Freedman, 1997) (see Appendix A.2) and an Implicit Association Test on Humans and Algorithms (Chien et al., 2019). These scores are not included in the analysis of the present study.

The experiment was conducted at the LEEN (Experimental Economics Laboratory in Nice).[4] We used oTree (Chen et al., 2016) to code the experiment and the survey, and the Heroku cloud application platform to run the experiment through their dedicated server. The experiment was conducted between May 2022 and April 2023. In total, we held 34 sessions with an average of 10 participants per session, resulting in 356 participants overall, averaging 50 participants per treatment.[5] Overall, 34.3% of the participants were male, and the age of the participants ranged from 18 to 25 years old. Our pairwise Rank Sum Tests indicate no treatment differences in terms of gender and age (all p-values > 0.1), suggesting that our randomization was effective. Incentives for the prediction task were structured such that two rounds out of 60 were randomly selected for payment, one from the first 30 rounds and one from the last 30 rounds. Participants were given a bonus of 7.5€ per correct prediction. The payment scheme was known to participants. The risk elicitation task (Holt and Laury, 2002) was incentivized and participants earned between 0.1€ and 3.8€. Moreover, participants were paid a 5€ show-up fee. Therefore, participants could earn a maximum of 23.8€ and a minimum of 5.1€. Including the show-up fee, the average payout was 14.8€ for a one-hour experiment.

4. Hypotheses

People follow a piece of advice when they trust it comes from a reliable source (Bonaccio and Dalal, 2006; Bailey et al., 2022). To assess whether a source of advice is credible, decision

[4] Participants were recruited from the LEEN Orsee Database (Greiner, 2015) which consists of a pool of 5,180 people from Université Côte d'Azur.

[5] We decided to include 48 participants per treatment after conducting a statistical power analysis. This involved running 1,000 Monte Carlo simulations, with 48 participants performing the same task 60 times across seven treatments and an assumed effect size of 0.20. The resulting statistical power was estimated to be 0.826 using a Rank Sum Test. We thus aimed at recruiting 12 participants for each of the four series for each treatment. The distribution of participants across treatments was as follows: 50 for the No-Advice baseline, 49 for Algo-Optimal, 51 for Human-Optimal, 52 for Algo-Bias, 50 for Human-Bias, 53 for Algo-Random, and 51 for Human-Random.

makers need to understand the basic rules and principles used by the advisor (Masuda et al., 2022). People will more likely trust advisors they understand. For example, the emergence of anti-science movements suggests people might prefer to believe non-academic experts with simple views of the world than academic pundits (Olshansky et al., 2020). This behavior is reminiscent of a fundamental human need for understanding (Fiske, 2018), leading people to prefer intuitive principles to counterintuitive scientific theories (Rutjens et al., 2018; Mede and Schäfer, 2020). For the same reason as people may reject scientific knowledge, they can also oppose advice based on artificial intelligence that is perceived to be too complex. This argument has pushed prominent authors to promote the field of AI explainability (Huang et al., 2018; Gunning and Aha, 2019; Russell, 2019; Vaughan and Wallach, 2020), which posits that by facilitating the understanding of AI predictions, algorithms will be more positively perceived and more likely to be trusted. However, this literature, which is inspired by attribution theories in Cognitive Psychology (Malle, 2004; Miller, 2019), assumes that explainability is as critical for endorsing algorithmic advice as it is for endorsing human advice. In the absence of controlled experiments, it is difficult to compare algorithmic and human advice as they differ in many dimensions, other than explainability, such as accuracy and cost.

Following Lehmann (2022), we posit that people will not downplay the credibility of algorithms as much as that of humans when both issue the same random advice. To explain this prediction, we contend that people apply different standards to human and algorithmic advice to evaluate their respective credibility. As part of the evolution of epistemic vigilance tools in humans (Sperber et al., 2010; Mercier, 2020), people will tend to attribute a high degree of credibility to a highly coherent human-based message. This is the case because, as emphasized in the AI explainability literature, algorithms are typically presented as black boxes and left unexplained (Huang et al., 2018; Gunning and Aha, 2019; Vaughan and Wallach, 2020). The lack of transparency of algorithms is widespread and people have thus presumably developed lower standards for algos than for humans in terms of explainability. This could be the reason why people continue to adopt algorithmic advice even when it is perceived to be excessively complex (Lehmann et al., 2022). In contrast, complex human advice can be viewed as both a sign of expertise and a deceptive attempt, thus being evaluated as less credible than the same algorithmic advice (Mercier, 2020). Explanations are also typically embedded in a social context (Sperber and Mercier, 2017) that is often absent of the interaction between humans and machines. This further limits people's expectations regarding the explainability of algorithms. Overall, we posit that people will perceive any advice that is apparently random as more credible if issued by an algo than if issued by a human.

In Random treatments, we thus expect the higher adoption of algorithmic advice compared to human advice to be more pronounced in which advice series are unpredictable. When decision makers cannot identify a predictable pattern in an advice series, they will be more likely to associate it with a lack of expertise when the advisor is a human than when the advisor is an algo. That is, people will be less likely to tolerate a an apparently random advice coming from a human than from an algorithm. People think of algorithms as rule-based predictable machines that produce consistent predictions (see Logg et al., 2019) whereas human experts are more likely to be impacted by idiosyncratic factors such as emotions and motivational factors (Kuhl et al., 1985; Raghunathan and Pham, 1999; Finucane et al., 2000; Loewenstein and Lerner, 2003; Bechara, 2005; Lerner et al., 2015). It follows that a random series of advice will more likely be seen as following a predictable pattern when issued by the algo than when issued by a human.

Hypothesis 1 (Random treatments)
In Random treatments, decision makers will be more likely to follow algorithmic advice than human advice, and this effect will be sustained over time.

In Optimal treatments, the advisor always recommends picking the most-frequent color ball so that the series of advice is perfectly predictable and has zero entropy. In that case, we do not expect differences in perceived credibility across advice types so that decision makers will be as likely to follow the algorithmic as the human advice (see Hypothesis 2).

Hypothesis 2 (Optimal treatments)
In Optimal treatments, decision makers will be as likely to follow algorithmic advice as human advice.

In Biased treatments, the series exhibits more variety and is less predictable. Yet, unlike the random advice series, the biased series is perfectly predictable once the underlying rule is learned, and the draw of the previous ball is observed. As a result, we expect that the positive impact of the algo on advice adoption will become insignificant over time in Biased treatments (see Hypothesis 3), thus resembling the prediction for the Optimal treatments (see Hypothesis 3). By contrast, the Random treatments maintain the unpredictability of the advice series over time so that the positive impact of the algorithm on advice adoption will persist (see second part of Hypothesis 1).

Hypothesis 3 (Biased treatments)

In Biased treatments, decision makers will be more likely to follow algorithmic advice than human advice, but this effect will decrease over time.

5. Results

5.1. Dependent variables

As primary metrics to evaluate our hypotheses, we use the Follow Ratio, the Switch Ratio and the Optimal Ratio.

The Follow Ratio is the proportion of times a participant's final choice is identical to the advice. The Follow Ratio thus measures the level of agreement between the decision maker and the advisor. To assess algorithm credulity more directly, we also measure people's inclination to follow bad advice by computing the 'Follow Bad Advice Ratio' as the proportion of times a decision maker follows the advice that consists in picking the least frequent color ball in the urn. An alternative measure to the Follow Ratio is the weight of advice used in the Judge Advisor System literature (Sniezek and Buckey, 1995) that captures the extent to which the advice impacts a participant's original choice.

The Switch Ratio is the proportion of times participants followed the advice although their initial decision differed from the advice. However, since our task is based on a repeated binary choice, we focus on the Follow Ratio as it is unaffected by endogeneity and treatment comparison issues.[6] We will only refer to the Switch ratio for robustness checks.

The Optimal Ratio is the proportion of times a participant picks the most frequent ball color from the urn. By definition, the higher the Optimal Ratio, the higher a participant's expected earnings. The Optimal Ratio for the advisors was 100%, 51.7%, and 50% in Optimal, Biased and Random treatments, respectively.[7]

5.2. Mistakes in the prediction task

For our prediction task to exhibit any of the hypothesized treatment differences, it is necessary that participants make mistakes and do not mechanically pick the optimal rule. As expected, given previous literature (Fisher, 1930; Clotfelter and Cook, 1993; Nowak and Sigmund, 1993; Rabin, 2002; Ayton and Fischer, 2005; Croson and Sundali, 2005; Gaissmaier and Schooler, 2008; Dohmen et al., 2009; Huber et al., 2010), participants did not consistently apply the optimal rule. In our baseline treatment, only 6% of participants decided to implement the optimal rule in every round. Additionally, 8.33% of participants applied the optimal rule in 90% of rounds (54 out of 60 rounds). Prior to receiving advice, participants applied the optimal rule

[6] Predictable advice, such as in the Optimal treatments, can influence a participant's choice in future rounds. Participants could learn the advisor's rule and apply it. By contrast, in Random treatments, participants cannot learn the rule used for the advice.

[7] These ratios are calculated as an average over the four series.

90% of the time in 12% (10.7%) {10.5%} of cases when the advisor was optimal (biased) {random}. Proportion tests revealed no significant differences, indicating that regardless of the advisor type, on average, 10.95% of participants used the optimal rule at least 90% of the time.

5.3. Hypotheses tests

In line with Hypotheses 1 and 3, decision makers were more likely to follow algorithmic rather than human advice in Random (61.7% vs 57.6%, Rank Sum Test, p-value = 0.043, see also Figure 3 (left panel) and Table 1 regressions (3) and (6)), and Biased treatments (61.1% vs 56.5%, Rank Sum Test, p-value = 0.010, see also regressions (2) and (5)). A similar pattern emerged when considering the Switch Ratio (Figure 3, right panel, and Table B1 in Appendix B) instead of the Follow. Our results suggest that decision makers were more likely to follow bad advice in Algo than in Human treatments. Indeed, as shown in Table B2 in Appendix B (see 'Algo Dummy'), decision makers were significantly more likely to follow a bad advice when it was issued by an algo than when issued by a human in both Biased (53.3% vs 44.5%, Rank Sum Test, p-value = 0.009) and Random treatments (54.7% vs 42.0%, Rank Sum Test, p-value = 0.012).[8] It is also interesting to note that under human advice, decision makers followed the bad advice as often as predicted by probability matching, which is the probability of picking the least-frequent color ball (45%). Under algorithmic advice, people followed the bad advice even more often. It follows that participants were impacted negatively by the presence of algorithmic advice.[9]

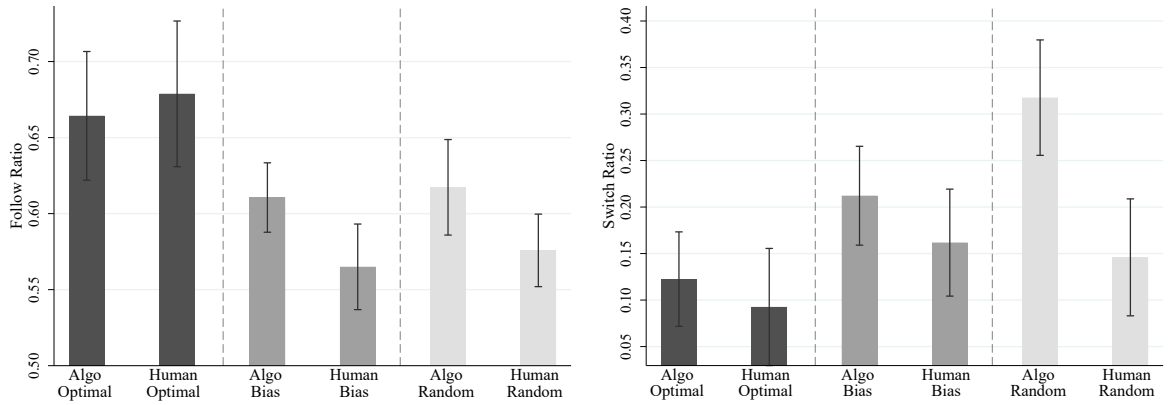


Figure 3. *Follow Ratio (left panel) and Switch Ratio (right panel) across advice treatments*.

[8] This is supported by the Switch ratio when the advisor gives bad advice: for the Biased treatment (24.45% vs. 18.23%, p-value = 0.02) and the Random treatment (32.18% vs. 16.62%, p-value < 0.001).

[9] We are not aware of a prediction experiment in which people chose the least-frequent color ball significantly more often than its frequency of occurrence (Sign Rank Tests, p-values = 0.030 and 0.011 for Biased and Random treatments).

In line with Hypothesis 2, Figure 2 also shows that decision makers are as likely to follow optimal algo as optimal human advice. This result is confirmed in Table 1 (see 'Algo Dummy' in regressions (1) and (4)).

Table 1. OLS regressions with robust standard errors at the individual level and series fixed effects for 'Follow Ratio' across treatments.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	Follow Ratio					
Treatment	Optimal	Bias	Random	Optimal	Bias	Random
Algo Dummy	-0.0152	0.0447***	0.0409**	-0.0559◊	0.0376**	0.0440**
	(0.0327)	(0.0169)	(0.0196)	(0.0310)	(0.0176)	(0.0198)
CRT				0.0350***	-0.0084**	-0.0116**
				(0.0106)	(0.0038)	(0.0049)
Years of Study				0.0231◊	0.0050	0.0196**
				(0.0121)	(0.0073)	(0.0085)
Risk Aversion				0.0018	-0.0004	0.0037
				(0.0063)	(0.0035)	(0.0041)
NARS				0.0003	0.0016	-0.0001
Interactions				(0.0040)	(0.0023)	(0.0020)
NARS				0.0038	-0.0017	0.0054***
Social Influence				(0.0048)	(0.0023)	(0.0019)
NARS				-0.0007	-0.0030	-0.0045**
Emotions				(0.0054)	(0.0023)	(0.0022)
Constant	0.7201****	0.5549****	0.5857****	0.5840****	0.6397****	0.5534****
	(0.0314)	(0.0164)	(0.0261)	(0.1017)	(0.0482)	(0.0676)
Observations	100	102	104	100	102	104
R^2	0.0247	0.2265	0.1239	0.1854	0.2824	0.2681
F-statistic	0.828	8.016	4.335	2.534	4.726	4.882

Robust standard errors in parentheses. 'Algo Dummy' is a dummy that takes value one if a person has been assigned to a treatment with an algorithmic advisor and value zero otherwise. Details of NARS scales are provided in Appendix A.2. **** p<0.001, *** p<0.01, ** p<0.05, ◊ p<0.1

To study the potential learning effects captured in Hypotheses 1 and 3, we compare participants' follow decisions in the first and second iterations, defined as the first and last 30 rounds of the experiment, using panel regressions (see Table 2). In line with Hypothesis 1, there were no learning effects in Random treatments as the variable 'Iteration × Algo Dummy' was not significant (see regressions (3) and (6)). In Random treatments, the coefficient for 'Algo Dummy + Iteration × Algo' is positive (Coefficient Tests, p-values = 0.054 and 0.049 for regressions (3) and (6)) showing that algorithmic advice increased the Follow Ratio in the second iteration. In line with Hypothesis 3, algorithmic advice led to a higher 'Follow Ratio' in the first but not in the second iteration in Biased treatments. Indeed, the coefficient for 'Algo Dummy' is positive and significant whereas 'Iteration × Algo' is negative and significant in regressions (2) and (5). Furthermore, 'Algo Dummy + Iteration × Algo' is not significantly different from zero (Coefficient Tests, p-values = 0.553 and 0.484 for regressions (2) and (5)).

In Table B3 in Appendix B, we show similar results when using the 'Follow Bad Advice Ratio' as dependent variable.

Table 2. Linear panel regressions with random effects and robust standard errors at the individual level and series fixed effects for 'Follow Ratio' in a given iteration across treatments.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	Follow Ratio in a given iteration					
Treatment	Optimal	Bias	Random	Optimal	Bias	Random
Algo Dummy	0.0500	0.2161****	0.0235	0.0093	0.2091****	0.0266
	(0.0577)	(0.0449)	(0.0342)	(0.0628)	(0.0442)	(0.0351)
Iteration Number	0.0346	0.0380**	0.0092	0.0346	0.0380**	0.0092
	(0.0245)	(0.0169)	(0.0151)	(0.0249)	(0.0172)	(0.0153)
Iteration × Algo Dummy	-0.0435	-0.1143****	0.0116	-0.0435	-0.1143****	0.0116
	(0.0388)	(0.0271)	(0.0217)	(0.0394)	(0.0275)	(0.0220)
CRT				0.0350****	-0.0084**	-0.0116**
				(0.0104)	(0.0037)	(0.0048)
Years of Study				0.0231◊	0.0050	0.0196**
				(0.0119)	(0.0071)	(0.0084)
Risk Aversion				0.0018	-0.0004	0.0037
				(0.0061)	(0.0034)	(0.0040)
NARS				0.0003	0.0016	-0.0001
Interactions				(0.0039)	(0.0022)	(0.0020)
NARS				0.0038	-0.0017	0.0054***
Social Influence				(0.0047)	(0.0022)	(0.0019)
NARS				-0.0007	-0.0030	-0.0045**
Emotions				(0.0053)	(0.0022)	(0.0022)
Constant	0.7028****	0.5359****	0.5811****	0.5666****	0.6207****	0.5488****
	(0.0310)	(0.0184)	(0.0264)	(0.0991)	(0.0476)	(0.0661)
Observations	200	204	208	200	204	208
R^2	0.0230	0.2045	0.1016	0.1419	0.2392	0.2144
χ^2-statistic	5.580	46.690	—	31.89	61.84	52.47

Robust standard errors in parentheses. "—" when the test statistic is not available. 'Algo Dummy' is a dummy that takes value one if a person has been assigned to a treatment with an algorithmic advisor and value zero otherwise. Iteration Number equals 1 (2) in the first (last) 30 rounds of the experiment. Details of NARS scales are provided in Appendix A.2. **** p<0.001, *** p<0.01, ** p<0.05, ◊ p<0.1

The absence of learning effects beyond the ones identified in the Biased treatments, which are attributed to the increase in the predictability of the advice, is evidenced by the positive and non-significant coefficient for 'Iteration Number' in regressions (1), (3), (4) and (6). Although experience with the task did not significantly decrease the 'Follow Ratio' and the 'Follow Bad Advice Ratio', individual factors, and in particular cognitive skills might reduce algorithm credulity. We investigate this possibility in the next section.

5.4. Individual factors

Previous analyses (see 'CRT' coefficient in Table 1) suggest that cognitive reflection skills could help people disregard the advice in Biased and Random treatments. This finding is in line with previous research showing that cognitive reflection is one key dimension of critical thinking and rational behavior (Cokely et al., 2009, Oechssler et al., 2009; Campitelli and Labollita, 2010; Toplak et al., 2011; Liberali et al., 2012, Corgnet et al., 2018). These studies have shown that people with high CRT scores are more likely to avoid common heuristics such as those used in designing the advice series in Biased treatments. It is thus reasonable to expect people with high cognitive reflection skills to better disentangle noise from information, thus disregarding the advice in Biased and Random treatments. However, the results in Table B4 in Appendix B suggest the ability of people with high cognitive reflection skills to disregard biased and random advice might not apply to algorithmic advice. Indeed, the coefficient associated with 'CRT', which is significant in the case of non-optimal human advice (see regression (3)), turns out to be non-significant in the case of non-optimal algorithmic advice (see regression (6)) and decreases in magnitude by about 60%. Furthermore, the treatment effects observed in Figure 3 continue to hold for participants with high cognitive skills, as defined by scoring in the top quartile on the CRT (see Figure B1 in Appendix B).

Future research should investigate whether this exploratory result regarding cognitive skills is robust and whether we can interpret it as evidence that humans have not yet fully developed the cognitive apparatus to evaluate the reliability of algorithmic advice.

5.5. Welfare consequences

In Figure 4, we show the Optimal Ratio across treatments. In line with our discussion of heuristics and biases in the design section, participants generally did not follow the optimal strategy. Across all treatments, only 3.6% of participants achieved an Optimal Ratio of 100%. We observe that under 'No Advice', participants picked the optimal choice in 63.4% of the cases (SD = 16.4%), which is lower than in Optimal treatments (Algo-Optimal and Human-Optimal) (Mean = 67.2%, SD = 16.1%, Rank Sum Test, p-value = 0.061). Optimal treatments led to a higher Optimal Ratio than Biased (Mean = 59.9%, SD = 14.4%, Rank Sum Test, p-value < 0.001) and Random treatments (Mean = 61.3%, SD = 15.1%, Rank Sum Test, p-value = 0.002). Overall, algorithmic advice led to a lower Optimal Ratio (Mean = 60.3%, SD = 13.2%) than human advice (Mean = 65.2%, SD = 17.1%, Rank Sum Test, p-value = 0.005).

In line with our hypotheses, the largest difference in Optimal Ratios between algorithmic and human advice was observed for Random treatments (57.1% vs 65.6%, Rank Sum Test, p-value = 0.019) (see Figure 4). This implies expected earnings were 13.0% lower

in Algo-Random than in Human-Random, which corresponds to an expected loss of 1.40€ for a one-hour experiment. As expected, the differences in Optimal Ratios between algorithmic and human advice were less pronounced in Biased (57.9% vs 62.0%, Rank Sum Test, p-value = 0.117) and in Optimal treatments (66.4% vs 67.9%, Rank Sum Test, p-value = 0.478).

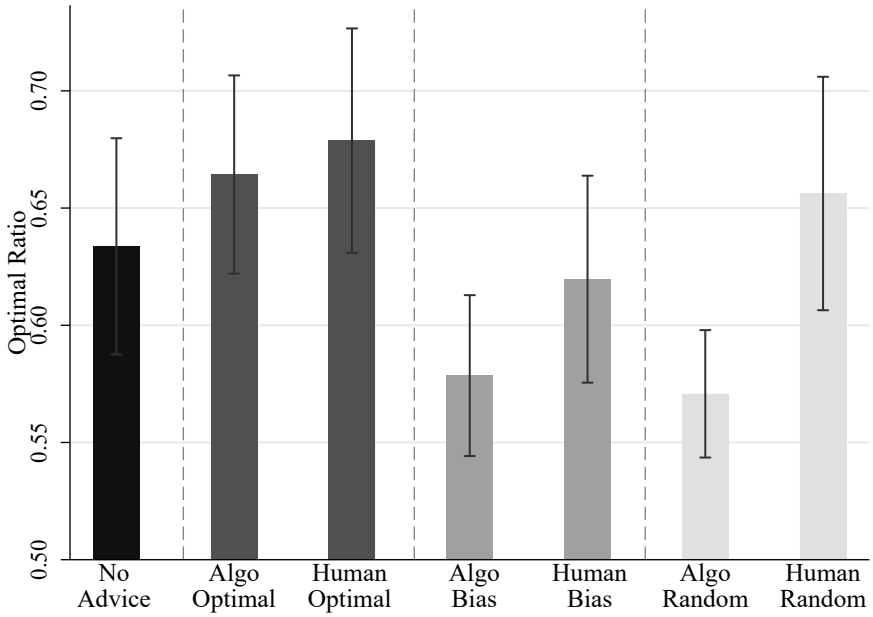


Figure 4. *Optimal Ratio across treatments*.

In Table B5 (Appendix B), a regression analysis confirms that optimal advice was beneficial to participants overall (see 'Optimal' in regression (1)) and that algorithmic advice led to significantly fewer optimal choices (see 'Algo Dummy' in Table B5).

These findings are in line with the results on algorithmic appreciation in Logg (2019). Interestingly, our findings emerge even though algorithmic and human advisors were highly inaccurate in the task. Indeed, optimal advisors provided the accurate answer in only 57.1% of the cases, and biased {random} advisors do even worse (55.4% {54.3%}). Our findings thus seem to contradict the seminal result of Dietvorst et al. (2014) that decision makers do not follow inaccurate algorithms. However, unlike the tasks considered in Dietvorst et al. (2014) where participants predict student performance or the number of airline passengers in U.S. states, our prediction task is abstract and only requires statistical reasoning for which algorithms are more likely to be perceived as experts than humans (Hertz and Wiese, 2019).

5.6. Underlying mechanisms for algorithm credulity

In this section, we investigate the potential mechanisms underlying algorithm credulity. To that end, we use additional scales elicited in the Random treatments (see Appendix A.1, Part 3). We

focused on random treatments because this is the treatment in which we should expect to observe the most significant and persistent difference between algorithmic and human advice (see Hypothesis). Underlying algorithm credulity is the idea that people will be more likely to discredit an apparently random advice coming from a human expert than when coming from an algo. Our main assumption is that participants decrease their epistemic vigilance when they receive advice from an algorithm compared to a human. As a result, we expect random pieces of advice to be perceived as more intelligible when issued by an algo than by a human. To measure the perceived intelligibility of the advice, we used four items to construct a Perceived Intelligibility Index. Each of the items were rated on a 1 to 7 Likert scale. The items are "I understood the choices made by the advisor.", "I understood the strategy used by the advisor.", "I found the piece of advice to be complex." (reversed item), "I found the piece of advice to be simple." In an open-ended question, participants were also asked to briefly describe the strategy adopted by the advisor.

In line with our conjectured mechanisms, we found that perceived intelligibility was indeed high when the advice was algorithmic (Mean = 9.87, SD = 3.72) than when it was human (Mean = 8.08, SD = 3.37) (Rank Sum Test, p-value = 0.020). This result is also corroborated by our analysis of the open-ended question regarding the description of the strategy used by the advisors in Random treatments. Participants described the strategy of the algorithmic advisor in more detail than that of the human advisor. Their responses used about twice more words (19.2 vs 9.6) and characters (112.5 vs 51.7) for the random algorithmic advisor than for the random human advisor (Rank Sum Tests, p-values = 0.004 and < 0.001). Furthermore, people were more likely to describe the algorithmic advisor as using probability calculations (23.5% vs 7.8%, Proportion Test, p-value = 0.029).[10] Moreover, they rated the random algorithmic advice as twice less likely to be either random or unintelligible (21.6%) than the human random advice (41.2%) (Proportion Test, p-value = 0.033).

Finally, in regression (1) in Table 3, we show that participants perceive a random algorithmic advisor as more intelligible than a random human advisor. This increased level of perceived intelligibility also leads to an increase in the level of perceived competence (see regression (2)). This relationship holds because people, in line with Bayesian reasoning, tend to downplay information coming from uncertain sources and producing unclear messages (Knill

[10] To analyze the text, two of the authors rated each of the 102 available answers without knowing which treatment they corresponded to. Both raters produced similar ratings. One rater categorized 3.9% (19.6%) of the answers as invoking probability calculations for the Human (Algo) Random treatment whereas the other rater reported the following percentages: 7.8% (23.5%). For the sake of conservatism, we report in the main text the tests and p-values associated with the ratings that are least favorable to our hypothesis.

and Pouget, 2004; Friston, 2012; Mercier, 2020). In turn, numerous studies have emphasized that perceived competence is critical for developing trust in other humans (see Mayer et al., 1995; Fiske et al., 2007 for reviews). This conjecture is confirmed in regression (3), and regression (4) shows that the higher the level of trust in the advice also leads to more frequent adoption of the advice. Interestingly, in regression (6) of Table 1, the 'Algo Dummy' variable plays a key role in explaining the follow ratio in the Random treatments. However, in regression (4) of Table 3, 'Trust' plays a significant role in explaining why people tend to follow advice, leading to a 73.7% decrease in the 'Algo Dummy' coefficient. This suggests that, as observed in Table 1, the 'Algo Dummy' primarily captures the effect of Trust in the advisor. As a result, participants follow the algorithmic advisor more often because they trust it more than the human advisor (6.40 vs 4.88, Rank Sum Test, p-value < 0.004). This greater trust in the algorithm stems from the perception of the algorithm as more competent, which, in turn, is based on the perception of the algorithm as being more intelligible.

6. Discussion

Casual observations of human behavior suggest that people may blindly follow algorithmic advice, a phenomenon we refer to as algorithm credulity. In this study, we provide the first causal evidence of algorithm credulity in a series of controlled experiments. To test our three hypotheses, we developed a simple experimental paradigm. Our main result shows that people followed deficient algorithmic advice more often than deficient human advice, leading to a loss of 13% in expected profit.

We interpret this phenomenon as suggesting that the mechanisms of epistemic vigilance that humans have developed for evaluating the credibility of human communications may be less effective when applied to algorithms. Humans may still need to learn how to distrust seemingly sophisticated, yet flawed, algorithmic advice. This learning must happen quickly, as the risk of algorithmic credulity—which, as we have shown, is amplified by unpredictable algorithmic advice—is likely to increase with the recent emergence of AI systems (see Zuiderveen Borgesius, 2018), such as Large Language Models, that do not rely on logical rules and generate unpredictable outcomes. Finally, we hope that this paper, by demonstrating the existence of algorithmic credulity, will encourage future research to further study the underlying mechanisms of the phenomenon and help develop policies that protect humans against it.

Table 3. OLS regressions with robust standard errors at the individual level and series fixed effects for 'Perceived Intelligibility', 'Perceived Competence', 'Trust', and 'Follow Ratio' as dependent variable in treatment Random.

Dependent Variable	(1) Perceived Intelligibility	(2) Perceived Competence	(3) Trust	(4) Follow Ratio
Algo Dummy	1.7671**			0.0105
	(0.7753)			(0.0161)
Perceived Intelligibility		0.2593****		
		(0.0616)		
Perceived Competence			0.6696****	
			(0.0995)	
Trust				0.0220****
				(0.0030)
CRT	-0.2347	-0.2794**	-0.1570	-0.0032
	(0.2159)	(0.1086)	(0.1357)	(0.0043)
Years of Study	-0.0005	0.1110	0.0766	0.0167***
	(0.1713)	(0.0998)	(0.1082)	(0.0062)
Risk Aversion	-0.3020	0.2940	-0.1095	-0.0003
	(0.3626)	(0.1878)	(0.1980)	(0.0032)
NARS Interactions	-0.0939	0.0345	0.0239	0.0000
	(0.0920)	(0.0444)	(0.0613)	(0.0017)
NARS Social Influence	0.1916**	0.0586	0.0616	0.0026
	(0.0776)	(0.0559)	(0.0455)	(0.0017)
NARS Emotions	0.0234	-0.0240	-0.0716	-0.0032*
	(0.0830)	(0.0580)	(0.0576)	(0.0017)
Constant	7.1245***	3.5346**	1.0225	0.4709****
	(2.2758)	(1.5762)	(1.3796)	(0.0532)
Observations	104	104	104	104
R^2	0.1777	0.3353	0.4354	0.5589
F-statistic	3.387	5.436	13.16	18.15

Robust standard errors in parentheses. 'Algo Dummy' is a dummy that takes value one if a person has been assigned to a treatment with an algorithmic advisor and value zero otherwise. 'Perceived Intelligibility', 'Perceived Competence', 'Trust' are variables computed using the scales in Appendix A1, Part 3.[11] **** p<0.001, *** p<0.01, ** p<0.05, ◊ p<0.1

7. References

Alekseev, A. (2020). The economics of babysitting a robot. Available at SSRN 3656684.

Aroyo, A. M., De Bruyne, J., Dheu, O., Fosch-Villaronga, E., Gudkov, A., Hoch, H., ... & Tamò-Larrieux, A. (2021). Overtrusting robots: Setting a research agenda to mitigate overtrust in automation. Paladyn, Journal of Behavioral Robotics, 12(1), 423-436.

[11] To measure perceived competence, we constructed the Perceived Competence Index using two related items, rated on a 1 to 7 Likert scale, "The advisor used a successful prediction strategy." and "The advisor used an effective prediction strategy". The level of trust in the advisor was measured constructing a Trust Index with the following two items rated on a 1 to 7 Likert scale: "Overall, I trusted the advisor." and "At the end of the experiment, I trusted the advisor."

Ayton, P., & Fischer, I. (2005). The Hot Hand Fallacy and the Gambler's Fallacy: Two faces of Subjective Randomness? Memory & Cognition, 32, 1369-1378.

Bailey P., Leon T., Ebner N., Moustafa A., Weidemann, G. (2022). A meta-analysis of the weight of advice in decision-making. Current Psychology.

Bechara, A. (2005). Decision making, impulse control and loss of willpower to resist drugs: A neurocognitive perspective. Nature Neuroscience, 8(11), 1458-1463.

Birnbaum, M. H., & Stegner, S. E. (1979). Source credibility in social judgment: Bias, expertise, and the judge's point of view. Journal of Personality and Social Psychology, 37(1), 48-74.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. Organizational Behavior and Human Decision Processes, 101(2), 127-151.

Burton, J. W., Stein, M. K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. Journal of Behavioral Decision Making, 33(2), 220-239.

Campitelli, G., & Labollita, M. (2010). Correlations of cognitive reflection with judgments and choices. Judgment and Decision Making, 5, 182-191.

Caplin, A., Dean, M., & Leahy, J. (2022). Rationally inattentive behavior: Characterizing and generalizing Shannon entropy. Journal of Political Economy, 130(6), 1676-1715.

Castelo, N., Bos, M., & Lehmann, D. (2019). Task-Dependent Algorithm Aversion. Journal of Marketing Research, 56, 002224371985178.

Charness, G., Samek, A. & van de Ven, J. (2022). What is considered deception in experimental economics? Experimental Economics, 25, 385-412.

Chen, D. L., 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.

Chien, S.-E., Chu, L., Lee, H.-H., Yang, C.-C., Lin, F.-H., Yang, P.-L., Wang, T.-M., & Yeh, S.-L. (2019). Age Difference in Perceived Ease of Use, Curiosity, and Implicit Negative Attitude toward Robots. ACM Transactions on Human-Robot Interaction, 8(2), 9:1-9:19.

Chugunova, M., & Sele, D. (2022). We and It: An interdisciplinary review of the experimental evidence on how humans interact with machines. Journal of Behavioral and Experimental Economics, 99, 101897.

Clotfelter, C. T., & Cook, P. J. (1993). The « Gambler's Fallacy » in Lottery Play. Management Science, 39(12), 1521-1525.

Cokely, E., & Kelley, C. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. Judgment and Decision Making, 4(1), 20-33.

Corgnet, B., Desantis, M., & Porter, D. (2018). What Makes a Good Trader? On the Role of Intuition and Reflection on Trader Performance. The Journal of Finance, 73: 1113-1137.

Corgnet, B. (2023). An experimental test of algorithmic dismissals. ESI Working Paper 23-02.

Corgnet, B., Hernán-González, R., & Mateo, R. (2023). Peer effects in an automated world. Labour Economics, 85, 102455.

Croson, R., & Sundali, J. (2005). The Gambler's Fallacy and the Hot Hand : Empirical Data from Casinos. Journal of Risk and Uncertainty, 30, 195 209.

Cross, E., & Ramsey, R. (2021). Mind Meets Machine: Towards a Cognitive Science of Human–Machine Interactions. Trends in Cognitive Sciences, 25, Issue 3, 200-212.

Darke, P. R., & Freedman, J. L. (1997). The belief in good luck scale. Journal of Research in Personality, 31(4), 486-511.

Dargnies, M. P., Hakimov, R., & Kübler, D. (2022). Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. CESifo Working Paper 9968.

Davis, D. D., & Holt, C. A. (1993). Experimental economics. Princeton University Press.
Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. American Psychologist, 34(7).
Dietvorst, B., Simmons, J., & Massey, C. (2014). Algorithm Aversion : People Erroneously Avoid Algorithms After Seeing Them Err. Journal of experimental psychology. General, 144.
Dohmen, T., Falk, A., Huffman, D., Marklein, F., & Sunde, U. (2009). Biased probability judgment : Evidence of incidence and relationship to economic outcomes from a representative sample. Journal of Economic Behavior & Organization, 72(3), 903 915.
Dzindolet, M., Pierce, L., Beck, H., & Dawe, L. (2002). The Perceived Utility of Human and Automated Aids in a Visual Detection Task. Human factors, 44, 79 94.
Edwards, W., & Fasolo, B. (2001). Decision technology. Annual review of psychology, 52(1), 581-606.
Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. International Journal of Forecasting, 25(1), 3-23.
Finucane, M.L., Alhakami, A., Slovic, P. and Johnson, S.M. (2000), The affect heuristic in judgments of risks and benefits. J. Behav. Decis. Making, 13: 1-17.
Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic & A. Tversky (Eds.), Judgment Under Uncertainty: Heuristics and Biases (pp. 422–444). Cambridge University Press.
Fisher, R. A. (1930). The genetical theory of natural selection. Clarendon Press.
Fiske, S.T., Cuddy, A.J.C., Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. Trends in Cognitive Sciences, 11, 2, 77-83.
Fiske, S. T. (2018). Social beings: Core motives in social psychology. John Wiley & Sons.
Frederick, S. (2005). Cognitive Reflection and Decision Making. Journal of Economic Perspectives, 19(4), 25-42.
French, J. R. P., Jr., & Raven, B. (1959). The bases of social power. In D. Cartwright (Ed.), Studies in social power, 150-167. University of Michigan.
Friston, K. (2012). The history of the future of the Bayesian brain. NeuroImage, 62(2), 1230-1233.
Fumagalli, E., Rezaei, S., & Salomons, A. (2022). OK computer: Worker perceptions of algorithmic recruitment. Research Policy, 51(2), 104420.
Gaissmaier, W., & Schooler, L. J. (2008). The smart potential behind probability matching. Cognition, 109(3), 416-422.
Gauvrit, N., Zenil, H., Soler-Toscano, F., Delahaye, J. P., & Brugger, P. (2017). Human behavioral complexity peaks at age 25. PLoS Computational Biology, 13(4), e1005408.
Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. Journal of the Economic Science Association, 1(1), 114-125.
Grenoble, R. (2013). Sabine Moreau, Belgian Woman, Drives 900 Miles Off 90-Mile Route because Of GPS Error. https://www.huffpost.com/entry/sabine-moreau-gps-belgium-croatia-900-miles_n_2475220
Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. AI Magazine, 40(2), 44-58.
Haibe-Kains, B., Adam, G. A., Hosny, A., Khodakarami, F., Waldron, L., Wang, B., ... & Aerts, H. J. (2020). Transparency and reproducibility in artificial intelligence. Nature, 586(7829), E14-E16.
Hertz, N., & Wiese, E. (2019). Good advice is beyond all price, but what if it comes from a machine?. Journal of Experimental Psychology: Applied, 25(3), 386.
Hey, J. D. (1998). Experimental economics and deception: A comment. Journal of Economic Psychology, 19(3), 397-401.

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. Industrial and Organizational Psychology,1(3), 333-342.

Holt, C. A., & Laury, S. K. (2002). Risk Aversion and Incentive Effects. American Economic Review, 92(5), 1644 1655.

Huang, L., Fong, N. M., & Tiedens, L. Z. (2018). The Influence of Power on Perspective-Taking in Political Negotiation: How Power Shaped the Construal of Progress in Peace Talks. Organizational Behavior and Human Decision Processes, 145, 1-16.

Huber, J., Kirchler, M., & Stöckl, T. (2010). The hot hand belief and the gambler's fallacy in investment decisions under risk. Theory and Decision, 68(4), 445 462.

Huettel, S. A., Mack, P. B., & McCarthy, G. (2002). Perceiving patterns in random series: dynamic processing of sequence in prefrontal cortex. Nature neuroscience, 5(5), 485-490

Jung, Markus & Seiter, Mischa. (2021). Towards a better understanding on mitigating algorithm aversion in forecasting: an experimental study. Journal of Management Control. 32. 1-22.

Jussupow, E., Benbasat, I., & Heinzl A. (2020). Why are we averse towards Algorithms? A comprehensive literature Review on Algorithm aversion. European Conference on Information Systems.

Knill, D.C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. Trends in Neurosciences, 27, 12, 712-19.

Krügel, S., Ostermaier, A., & Uhl, M. (2022). Zombies in the Loop? Humans Trust Untrustworthy AI-Advisors for Ethical Decisions. Philosophy & Technology, 35(1), 1-37.

Krügel, S., Ostermaier, A., & Uhl, M. (2023). Algorithms as partners in crime: A lesson in ethics by design. Computers in Human Behavior, 138, 107483.

Kuhl, J., Beckmann, J. (1985). Historical Perspectives in the Study of Action Control. In: Kuhl, J., Beckmann, J. (eds) Action Control. SSSP Springer Series in Social Psychology. Springer, Berlin, Heidelberg.

Langley, C., Cîrstea, B. I., Cuzzolin, F., & Sahakian, B. J. (2022). Theory of Mind in Humans and in Machines. Frontiers in Artificial Intelligence, 5, 3, 379-423.

Larrick, R. P. (2004). Debiasing. Blackwell handbook of judgment and decision making, 316-338.

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. Annual review of psychology, 66(1), 799-823.

Lehmann, C. A., Haubitz, C. B., Fügener, A., & Thonemann, U. W. (2022). The risk of algorithm transparency: How algorithm complexity drives the effects on the use of advice. Production and Operations Management, 31(9), 3419-3434.

Liang A., (2019) "Inference of preference heterogeneity from choice data", Journal of Economic Theory, 179, 275-311.

Liberali, J.M., Reyna, V.F., Furlan, S., Stein, L.M. & Pardo, S.T. (2012), Individual Differences in Numeracy and Cognitive Reflection, with Implications for Biases and Fallacies in Probability Judgment. Journal of Behavioral Decision Making, 25: 361-381.

Liu, C. Y., & Huang, K. P. (2005). An empirical study of travelers' external information search activity in Internet-based hotel reservation systems. Tourism Management, 26(5), 709-721.

Loewenstein, G. F., & Lerner, J. S. (2003). The role of affect in decision making. In R. J. Davidson, H. H. Goldsmith, & K. R. Scherer (Eds.), Handbook of Affective Sciences (pp. 619-642). Oxford University Press.

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

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human-human and human-automation trust: An integrative review. Theoretical Issues in Ergonomics Science, 8(4), 277-301.

Mahmud, H., Islam, A.K.M. N., Ahmed, S.I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. Technological Forecasting and Social Change, Elsevier, vol. 175(C).

Malle, B. (2004). How the Mind Explains Behavior: Folk Explanation, Meaning and Social Interaction. The MIT Press.

March, C. (2021). Strategic interactions between humans and artificial intelligence: Lessons from experiments with computer players. Journal of Economic Psychology, 87, 102426.

Masuda, T., Mikami, R., Sakai, T., Serizawa, S., & Wakayama, T. (2022). The net effect of advice on strategy-proof mechanisms: an experiment for the Vickrey auction. Experimental Economics, 25, 902–941.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. The Academy of Management Review, 20(3), 709-734.

Mede, N. G., & Schäfer, M. S. (2020). Science-related populism: Conceptualizing populist demands toward science. Public Understanding of Science, 29(5), 473-491.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence, 149. University of Minnesota Press.

Mercier, H. (2020). Not born yesterday: The science of who we trust and what we believe. Princeton University Press.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence, 267, 1-38.

Mononen, L. (2023). On Preference for simplicity and probability weighting. Working Paper available at: https://drive.google.com/file/d/1VD9lbZdvisUJBsvc1Vfd5PYo6ZAFOnDV/edit.

Murphy, K. (2012). GPS fails; man follows directions straight into Alaska harbor. https://www.latimes.com/nation/la-xpm-2012-aug-23-la-na-nn-whittier-harbor-alaska-20120823-story.html

Muir, B. M. (1994). Trust in automation: I. Theoretical issues in the study of trust and human intervention in automated systems. Ergonomics, 37(11), 1905-1922.

Muir, B. M., & Moray, N. (1996). Trust in automation: II. Experimental studies of trust and human intervention in a process control simulation. Ergonomics, 39(3), 429-460.

Nomura, T., Kanda, T., & Suzuki, T. (2006). Experimental investigation into influence of negative attitudes toward robots on human-robot interaction. AI Soc., 20, 138-150.

Nowak, M., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. Nature, 364, 56-58.

Oechssler, J., Roider, A., & Schmitz, P. W. (2009). Cognitive abilities and behavioral biases. Journal of Economic Behavior & Organization, 72(1), 147-152.

Olshansky, A., Peaslee, R. M., & Landrum, A. R. (2020). Flat-smacked! Converting to flat eartherism. Journal of Media and Religion, 19(2), 46-59.

Önkal, D., Goodwin, P., Thomson, M., Gönül, 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.

Pinker, S. (2022). Rationality: What it is, why it seems scarce, why it matters. Penguin.

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

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A.,& Hamilton, J. (2016) The Development and Testing of a New Version of the Cognitive Reflection Test Applying Item Response Theory (IRT). Journal of Behavioral Decision Making, 29: 453-469.

Rabin, M. (2002). Inference by Believers in the Law of Small Numbers. The Quarterly Journal of Economics, 117(3), 775-816.

Raghunathan, R.,& Pham, M. T. (1999). All negative moods are not equal: Motivational influences of anxiety and sadness on decision making. Organizational Behavior and Human Decision Processes, 79(1), 56-77.

Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Penguin.

Rutjens, B. T., Sutton, R. M., & van der Lee, R. (2018). Not all skepticism is equal: Exploring the ideological antecedents of science acceptance and rejection. Personality and Social Psychology Bulletin, 44(3), 384-405.

Shannon, C.E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal, 27, 379-423 & 623-656.

Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in Judge-Advisor decision making. Organizational Behavior and Human Decision Processes, 62(2), 159-174.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. Mind & Language, 25(4), 359-393.

Tetlock, P. (2009). Expert Political Judgment: How Good is It? How can We Know? Princeton University Press.

Toplak, M., West, R., & Stanovich, K. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics and biases tasks. Memory & Cognition, 39, 1275-1289.

Toplak, M., West, R., & Stanovich, K. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. Thinking & Reasoning, 20(2), 147-168.

Trajtenberg, M. (2018). Artificial Intelligence as the Next GPT : A Political-Economy Perspective. In The Economics of Artificial Intelligence : An Agenda (p. 175 186). University of Chicago Press.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. Psychological Bulletin, 76(2), 105-110.

Vaughan, J. W., & Wallach, H. (2020). A human-centered agenda for intelligible machine learning. Machines We Trust: Getting Along with Artificial Intelligence.

Venkatesh, V., & Davis, F. (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. Management Science, 46, 186-204.

Wilson, B. J. (2016). The meaning of deceive in experimental economic science. In The Oxford handbook of professional economic ethics. Oxford University Press.

Woolley, K., & Risen, J. L. (2018). Closing your eyes to follow your heart: Avoiding information to protect a strong intuitive preference. Journal of Personality and Social Psychology, 114(2), 230.

Zellner, M., Abbas, A. E., Budescu, D. V., & Galstyan, A. (2021). A survey of human judgement and quantitative forecasting methods. Royal Society open science, 8(2), 201187.

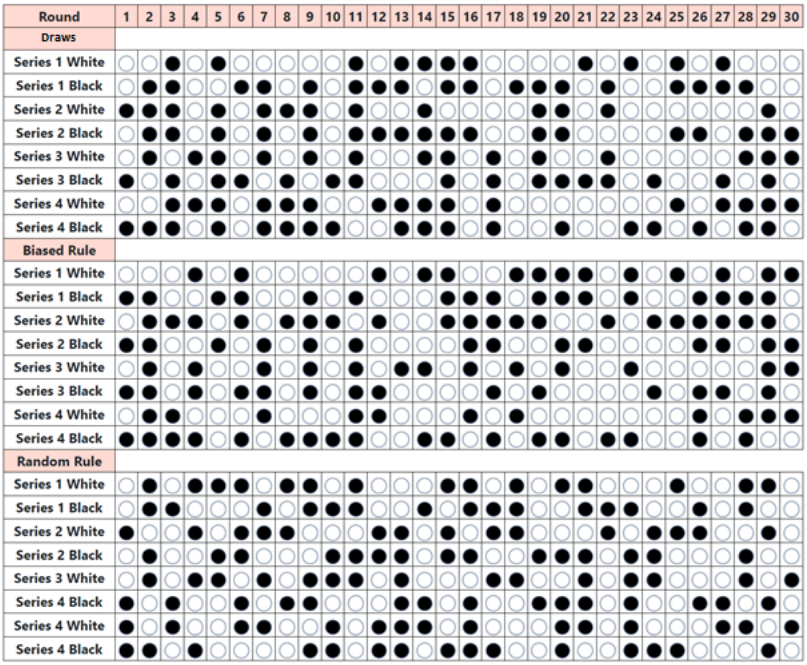
Zuiderveen Borgesius, F. (2018). Discrimination, artificial intelligence, and algorithmic decision-making. Council of Europe, Directorate General of Democracy, 42.

Appendices

Appendix A (Instructions[12])

Table A1. Series of draws and advice rules.

Series 'White' means that there are more white balls than black balls in the urn (first 30 rounds). Series 'Black' means that there are more black balls than white balls in the urn (last 30 rounds). We omit the optimal advice in this table because it always consists in recommending white (black) when the most frequent ball is white (black).



A.1. Lab instructions

Below, are the are translated instructions from French for the advice treatments. We use [] to refer to the Algo treatments instructions when needed. The instructions for the baseline simply omit all the references to the advice stage.

Welcome instructions (for every experiment conducted at LEEN):

Welcome to the Experimental Economics Laboratory of Nice (the LEEN). By agreeing to participate in this experiment, you express your full agreement with the Laboratory's regulations, available on the website or upon request. You will participate in an experiment where your decisions will be anonymous and will partly determine your final payment. Therefore, we invite you to read the following instructions carefully. In addition to earnings collected in the experiment and regardless of your decisions, a fixed amount of 5 euros will be given to you to cover

[12] The original instructions and their English translation are provided on OSF: https://osf.io/6kxmp/ .

your travel expenses. A variable amount will be added based on your decisions during the experiment. Your earnings will be paid individually and confidentially at the end of the experiment.

In order not to alter the results of the experiment, we ask that you do not communicate with other participants. We also ask that you kindly turn off your mobile phones and refrain from using them throughout the duration of the experiment. If these rules are breached, the experiment will be interrupted, and earnings will be forfeited. If you encounter a technical problem, please simply raise your hand, and wait for the experimenter to come to you. Everybody in this room has access to the same instructions and will participate in the same experiment.

Part 1

Description:

1. This part begins with a classic prediction task that will span 5 rounds, aiming to familiarize you with the task. This is a practice and unpaid task.
2. Baseline (no-advice) treatment:

   The first paid task will also be a prediction task. The prediction involves determining the color of the ball drawn at random by a computer. This task will be paid and will last for 30 rounds.

   Human treatment

   The first paid task will also be a prediction task that will be done with the help of a human advisor. The human advisor is another participant who completed a previous session of a similar prediction task. The human advisor aim is to assist you in making a prediction. Here, the prediction involves determining the color of the ball drawn randomly by a computer. This task is paid and lasts for 30 rounds.

   Algo treatment

   The first paid task will also be a prediction task that will be done with the help of an algorithmic advisor. The algorithmic advisor is a computer program designed to assist you in making a prediction. Here, the prediction involves determining the color of the ball drawn randomly by a computer. This task is paid and lasts for 30 rounds.
3. The second paid task will be similar to the first paid task. It will also last for 30 rounds.

In total, you can earn a maximum of 15 euros for the prediction task, and a minimum of 0. At the end of the experiment, you will also complete a risk preference task for which you can earn 0.10, 1.60, 2.00, or 3.85 euros. These payments depend on your decisions and on chance.

Training Task Procedure:

The purpose of this task is to allow you to practice with the task for which you will later be paid. There are 5 rounds.

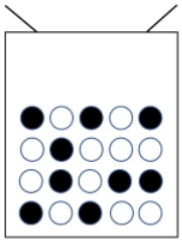
Steps:

1. A ball will be drawn from this urn at random. Your goal is to guess the color of the drawn ball (**Figure A1**).
2. You will predict the color of the ball drawn at random by the computer.
3. You will observe the color of the drawn ball (**Figure A2**).
4. The drawn ball will be placed back into the urn, and you will repeat this task 5 times.

Choice 5/5

The ball drawn previously has been placed back into the urn

A ball has been randomly drawn by the computer.
You must guess which ball has been drawn. At the end of the round, the ball is placed back into the urn, and then the urn is shuffled.



Previously :

Round	1	2	3	4	5
Balls drawn by the computer	○	●	○	●	
Your choices	○	○	●	●	
Your results	✓	✗	✗	✓	

In your opinion, what is the color of the ball drawn in the 5th round ?

○ Black
○ White

Next

Figure A1. Decision screen for the training phase (Round 5).

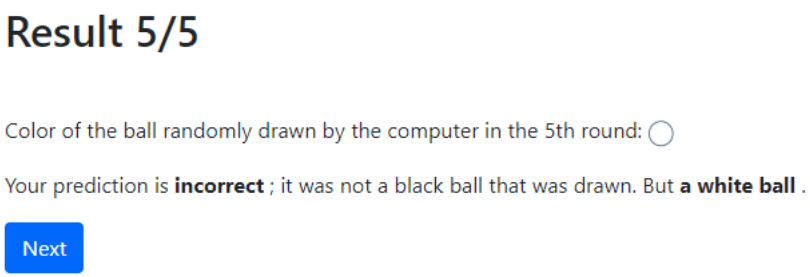


Figure A2. Feedback screen for the training phase (Round 5).

Procedure for the 1st Task:

A ball is drawn from this urn. Your goal is to guess the color of the drawn ball.

The urn contains 11 white balls and 9 black balls.

1. You will predict the color of the ball drawn by the computer.
2. Another participant [an algorithm] will provide their prediction. You will then be asked to make a final prediction of the color of the selected ball. You are free to change your decision, only the color of the ball you select for the final prediction is considered for your payment.
3. You observe the color of the drawn ball. You will find out if you and the advisor made a correct prediction.
4. The drawn ball will be placed back into the urn, and you will repeat this 30 times.

Once you have finished this 1st task, you will complete a 2nd task for 30 rounds, which is similar to the first one. In total, you will play 60 rounds.

Payments:

Two of the 60 rounds (One round in the 1st task and another round in the 2nd task) will be randomly selected for payment, so it is in your best interest to always make the best possible prediction. If your prediction is correct for the selected round, you will earn 7.5 euros. If your prediction is incorrect, you earn 0 euro. If both of your predictions are correct, you earn 15 euros.

Note: We only keep the prediction made after receiving the advice for payment. Therefore, only the final prediction made during a round counts for payments.

Example of payments: Round 8 of the first task and round 23 of the second task are randomly selected for payment. You predicted that a black ball would be drawn in round 8, but a white

ball was drawn. In round 23, after receiving advice from the other participant [algorithm], you predicted that a white ball would be drawn, and indeed a white ball was drawn. Therefore, you earn 7.5 euros because the prediction made for round 23 (task 2) was correct. Had the prediction for round 8 been correct, you would have earned 15 euros.

Procedure for the 2nd Task:

The same as for 1st task, except that:

The urn contains 9 white balls and 11 black balls.

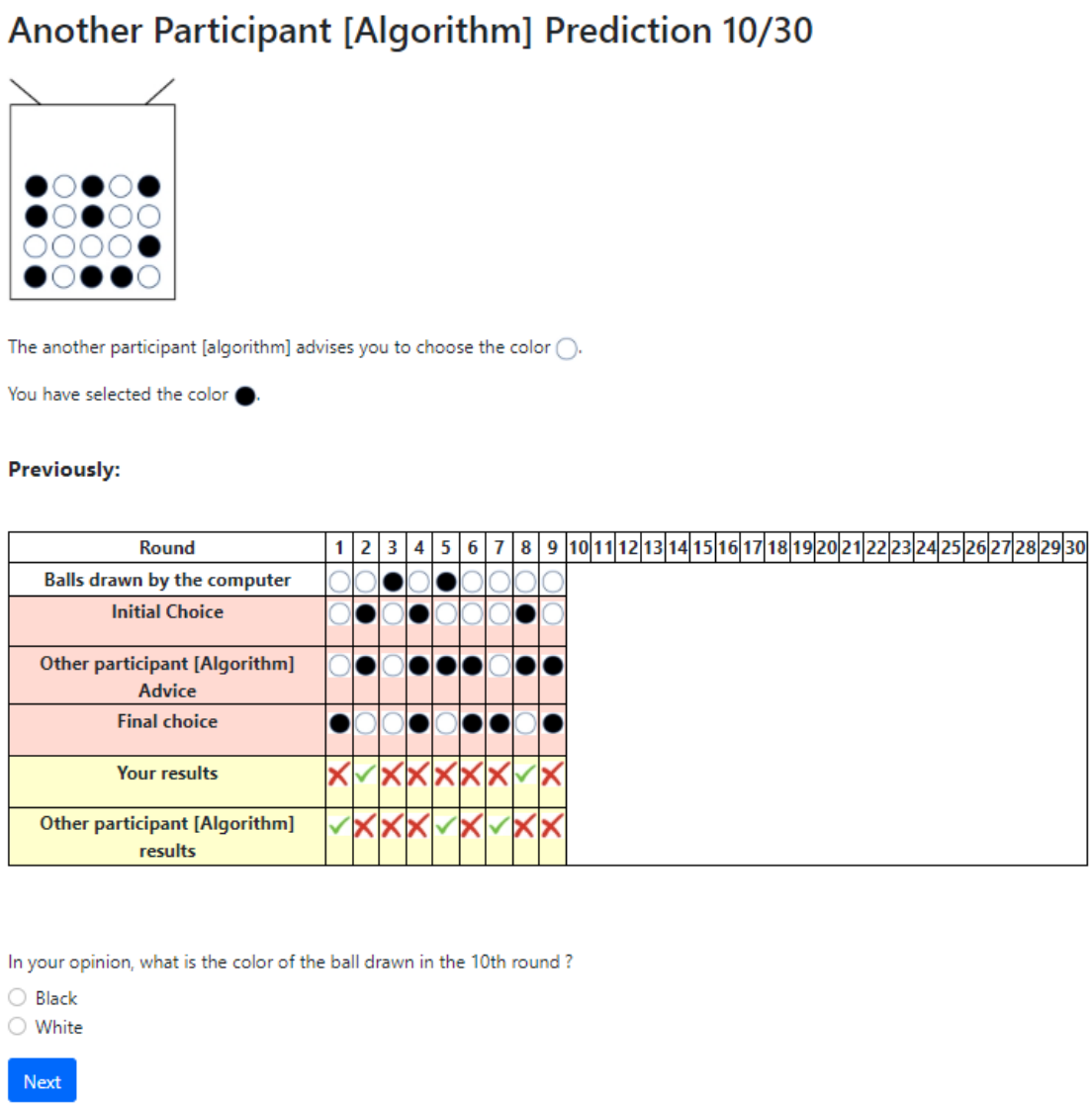


Figure A3. Decision screen for the main prediction task (Round 10).

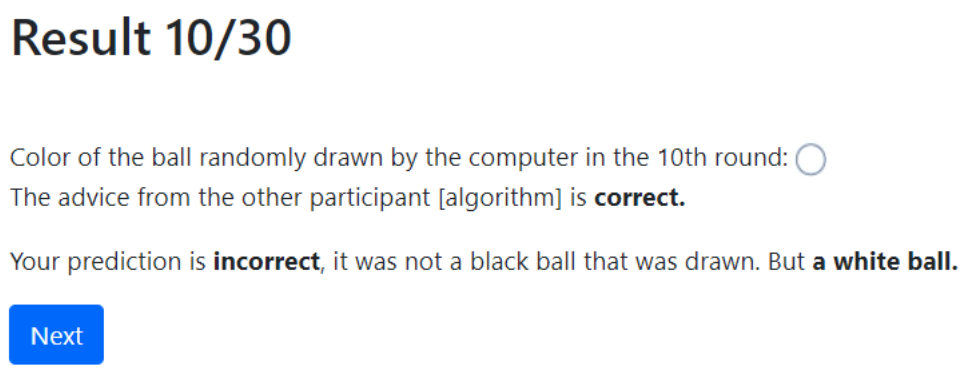


Figure A4. Decision screen for the main prediction task (Round 10).

Part 2

1) Questionnaire (This is the CRT 7 adapted from Toplak et al., 2014). Participants have 7 minutes to answer all the questions.

    1. If 2 nurses take 2 minutes to measure the blood pressure of 2 patients, how long would it take 200 nurses to measure the blood pressure of 200 patients? [Correct answer: 2 minutes; intuitive answer: 200 minutes]

    2. Soup and salad cost a total of 5.50 euros. The soup costs 5 euros more than the salad. How much does the salad cost? [Correct answer: 0.25 euro; intuitive answer: 0.5 euro]

    3. Sally is making sun tea. Every hour, the concentration of the tea doubles. If it takes 6 hours for the tea to be ready, how long would it take for the tea to reach half of its final concentration? [Correct answer: 5 hours; intuitive answer: 3 hours]

    4. If John can drink one barrel of water in 6 days, and Mary can drink one barrel of water in 12 days, how long would it take them to drink one barrel of water together? [Correct answer: 4 days; intuitive answer: 9 days]

    5. Jerry received both the 15th highest and the 15th lowest marks in the class. How many students are in the class? [Correct answer: 29 students; intuitive answer: 30 students]

    6. A man buys a pig for 60 euros, sells it for 70 euros, buys it back for 80 euros, and then sells it again for 90 euros. How much has he made? [Correct answer: 20 euros; intuitive answer: 10 euros]

7. Simon decided to invest 8,000 euros in the stock market one day early in 2008. Six months after he invested, on July 17, the stocks he had purchased had decreased by 50%. Fortunately for Simon, from July 17 to October 17, the stocks he had purchased increased by 75%. At this point, Simon: (a) broke even in the stock market, (b) is ahead of where he started, (c) has lost money. [Correct answer: c, because the value at this point is 7,000 euros; intuitive answer: b].

2) Risk aversion was elicited using the task in Holt and Laury (2002), but with euros. We decided to use the single switching point version of the task so that once participants switched from option A to option B, or vice versa, they could not switch back.

Part 3 (Only for Random treatments and for Human-Bias treatment)

For all these scales, the answers were on a 7-item Likert Scale from completely agree to completely disagree. All reversed items are marked with *.

Perceived intelligibility

1. I found the piece of advice to be simple.*
2. I found the piece of advice to be complex.
3. I understood the strategy used by the advisor.
4. I understood the choices made by the advisor.

Perceived competence

1. The advisor used a successful prediction strategy.
2. The advisor used an effective prediction strategy.

Trust in the advisor

1. At the beginning of the experiment, I trusted the advisor.
2. In the middle of the experiment, I trusted the advisor.
3. At the end of the experiment, I trusted the advisor.
4. Overall, I trusted the advisor.

Attention

1. I did not put in much effort while completing this task.
2. I was attentive while performing this task.
3. This task required a lot of energy from me.*
4. I was conscientious during this task.

5. I made a significant effort to complete this task.*

6. I did not put in much effort while completing this task.

Emotion

1. This task frustrated me.

2. This task emotionally impacted me.

Participant strategy

1. I used a successful strategy to maximize the number of correct choices.

2. I could have used a more effective strategy.

Other

1. If the advisor had been more successful, I would have trusted them more.

2. I was more effective than the advisor in making correct choices.

3. I was lucky during the prediction tasks.

4. The advisor was lucky during the prediction tasks.

Open questions

1. Briefly describe the best strategy, the one to maximizes the number of correct choices.

2. Briefly describe the strategy used by the advisor.

A.2. Online Questions

All reversed items are marked with *.

1) NARS (Nomura et al., 2006)

We used the French translation validated by Dinet and Vivian (2015). Participants responded using a 7-item Likert scale. Here are the questions in English.

Subscale 1 (NARS Interactions). Negative Attitudes toward Situations and Interactions with Robots. [Cronbach's $\alpha = 0.71$]

1. I would feel uneasy if I was given a job where I had to use robots.
2. The word "robot" means nothing to me.
3. I would feel nervous operating a robot in front of other people.
4. I would hate the idea that robots or artificial intelligences were making judgments about things.
5. I would feel very nervous just standing in front of a robot.
6. I would feel paranoid talking with a robot.

Subscale 2 (NARS Social Influence). Negative Attitudes toward Social Influence of Robots. [Cronbach's $\alpha = 0.76$]

1. I would feel uneasy if robots really had emotions.
2. Something bad might happen if robots developed into living beings.
3. I feel that if I depend on robots too much, something bad might happen.
4. I am concerned that robot would be a bad influence on children.
5. I feel that in the future society will be dominated by robots.
6. I feel that in the future, robots will be commonplace in society.

Subscale 3 (NARS Emotions). Negative Attitudes toward Emotions in Interaction with Robots. [Cronbach's $\alpha = 0.84$]

1. I would feel relaxed talking with robots.
2. If robots had emotions, I would be able to make friends with them.
3. I feel that I could make friends with robots.
4. I feel comforted being with robots that have emotions.
5. I feel comfortable being with robots.

2) The Belief in Good Luck Scale

We utilized a French translation provided by the Côte d'Azur University's translation department for the items in the scale, which can be found in Darke and Freedman (1997). Participants responded using a 5-item Likert scale. [Cronbach's $\alpha = 0.695$]

1. Luck plays an important part in everyone's life.
2. Some people are consistently lucky, and others are unlucky.
3. I consider myself to be a lucky person.
4. I believe in luck.
5. I often feel like it's my lucky day.
6. Nobody can win at games of chance in the long-run.
7. I consistently have good luck.
8. I tend to win games of chance.
9. It's a mistake to base any decisions on how lucky you feel.*
10. Luck works in my favor.
11. I don't mind leaving things to chance because I'm a lucky person.
12. Even the things in life I can't control tend to go my way because I'm lucky.

13. I consider myself to be an unlucky person.
14. There is such a thing as luck that favors some people, but not others.
15. Luck is nothing more than random chance.*

3) Implicit Association Test (IAT)

Here are the words we have translated from French to construct our IAT based on Chien et al. (2019).

Table A2. Concepts and words used in the IAT.

Concept	**Algorithm**	**Human**	**Negative** (*Chien et al., 2019*)	**Positive** (*Chien et al., 2019*)
Words to associate with each concept:	Computer Informatics System Algorithm Probability	Living Emotion Human Man Woman	Abuse Resentment Abandoned Hate Fear	Ecstasy Surprise Satisfaction Happiness Trust

4) Technology Acceptance Model 2

We utilized a French translation provided by the Côte d'Azur University's translation department for the TAM 2 questions, which can be found in Venkatesh and Davis (2000). Compared to the original version, we replaced the word 'system' with 'algorithm' to align with the focus of our study. Participants responded using a 7-item Likert scale. Cronbach's α was not acceptable in three out of the nine dimensions of the TAM 2 scale. This might be due to the placement of this questionnaire at the very end of the online survey, thus being impacted by participants fatigue and lack of attention.

1. Using an algorithm in my job increases my productivity.
2. Using an algorithm enhances my effectiveness in my job.
3. I find algorithm to be useful in my job.
4. My interaction with an algorithm is clear and understandable.
5. Interacting with an algorithm does not require a lot of my mental effort.
6. I find the algorithm to be easy to use.
7. I find it easy to get an algorithm to do what I want it to do.
8. People who influence my behavior think that I should use an algorithm.

9. People who are important to me think that I should use an algorithm.
10. My use of an algorithm is voluntary.
11. My supervisor does not require me to use an algorithm.
12. Although it might be helpful, using an algorithm is certainly not compulsory in my job.
13. People in my organization who use an algorithm have more prestige than those who do not.
14. People in my organization who use an algorithm have a high profile.
15. Having algorithm is a status symbol in my organization.
16. In my job, usage of an algorithm is important.
17. In my job, usage of an algorithm is relevant.
18. The quality of the output I get from an algorithm is high.
19. I have no problem with the quality of the algorithm's output.
20. I have no difficulty telling others about the results of using an algorithm.
21. I believe I could communicate to others the consequences of using an algorithm.
22. The results of using an algorithm are apparent to me.
23. I would have difficulty explaining why using an algorithm may or may not be beneficial.

The Cronbach's α correspond to the nine dimensions of the TAM 2 model are shown in brackets: Intention to use [0.78], Perceived Usefulness [0.85], Perceived Ease of Use [0.80], Subjective Norm [0.70], Voluntariness [0.46], Image [0.75], Job relevance [0.75], Output Quality [0.64], Result Demonstrability [0.69].

Appendix B. Additional analyses

Table B1. OLS regressions with robust standard errors at the individual level and series fixed effects for 'Switch Ratio' across treatments.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	Switch Ratio					
Treatment	Optimal	Bias	Random	Optimal	Bias	Random
Algo Dummy	0.0234	0.0546	0.1637****	0.0402	0.0804**	0.1587***
	(0.0446)	(0.0363)	(0.0448)	(0.0525)	(0.0377)	(0.0538)
CRT				-0.0168	-0.0059	-0.0149
				(0.0164)	(0.0087)	(0.0110)
Years of Study				0.0207	-0.0302◊	-0.0095
				(0.0229)	(0.0169)	(0.0252)
Risk Aversion				0.0038	-0.0031	0.0090
				(0.0075)	(0.0078)	(0.0094)
NARS				0.0043	0.0039	-0.0003
Interactions				(0.0084)	(0.0042)	(0.0061)
NARS				0.0090	0.0038	0.0075◊
Social Influence				(0.0076)	(0.0051)	(0.0044)
NARS				-0.0004	0.0024	0.0016
Emotions				(0.0069)	(0.0046)	(0.0073)
Constant	0.1254**	0.0956***	0.1817**	-0.0648	0.0270	0.0271
	(0.0460)	(0.0319)	(0.0733)	(0.1168)	(0.1288)	(0.1648)
Observations	36	52	60	36	52	60
R^2	0.0378	0.1771	0.2387	0.1562	0.2865	0.3057
F-statistic	0.307	3.651	4.573	0.988	2.787	5.267

Robust standard errors in parentheses. 'Algo Dummy' is a dummy that takes value one if a person has been assigned to a treatment with an algorithmic advisor and value zero otherwise. Details of NARS scales are provided in Appendix A.2.
**** p<0.001, *** p<0.01, ** p<0.05, ◊ p<0.1

Table B2. OLS regressions with robust standard errors at the individual level and series fixed effects for 'Follow Bad Advice Ratio' as dependent variable.

	(1)	(2)	(3)	(4)
Dependent Variable	Follow Bad Avice Ratio			
Treatment	Bias	Random	Bias	Random
Algo Dummy	0.0887**	0.1253***	0.0858**	0.1361****
	(0.0349)	(0.0394)	(0.0401)	(0.0386)
CRT			-0.0085	-0.0437****
			(0.0086)	(0.0111)
Years of Study			0.0097	0.0220
			(0.0176)	(0.0178)
Risk Aversion			-0.0042	0.0200**
			(0.0058)	(0.0093)
NARS			0.0064	-0.0020
Interactions			(0.0048)	(0.0041)
NARS			-0.0036	0.0067
Social Influence			(0.0050)	(0.0042)

NARS			-0.0053	-0.0066
Emotions			(0.0043)	(0.0046)
Constant	0.4352****	0.4279****	0.5663****	0.3744**
	(0.0330)	(0.0561)	(0.1046)	(0.1439)
Observations	102	104	102	104
R^2	0.0662	0.1339	0.1070	0.3606
F-statistic	1.995	3.685	1.711	6.981

Robust standard errors in parentheses. 'Algo Dummy' is a dummy that takes value one if a person has been assigned to a treatment with an algorithmic advisor and value zero otherwise. Details of NARS scales are provided in Appendix A.2. **** p<0.001, *** p<0.01, ** p<0.05, ◊ p<0.1

Table B3. Panel regressions with random effects and robust standard errors at the individual level and series fixed effects for 'Follow Bad Advice Ratio' in a given iteration as dependent variable.

	(1)	(2)	(3)	(4)
Dependent Variable	Follow Bad Advice Ratio			
Treatment	Bias	Random	Bias	Random
Algo Dummy	0.3180****	0.1677*	0.3130****	0.1917***
	(0.0892)	(0.0970)	(0.0718)	(0.0683)
Iteration Number	0.0157	0.0157	0.0157	0.0157
	(0.0450)	(0.0504)	(0.0279)	(0.0302)
Iteration × Algo Dummy	-0.1504***	-0.0371	-0.1504****	-0.0371
	(0.0572)	(0.0625)	(0.0422)	(0.0416)
CRT			-0.0082	-0.0437****
			(0.0084)	(0.0109)
Years of Study			0.0104	0.0220
			(0.0172)	(0.0174)
Risk Aversion			-0.0043	0.0200**
			(0.0057)	(0.0091)
NARS	0.0067	0.0041	0.0061	-0.0020
Interactions	(0.0042)	(0.0038)	(0.0047)	(0.0040)
NARS	-0.0037	0.0024	-0.0032	0.0067*
Social Influence	(0.0041)	(0.0039)	(0.0049)	(0.0041)
NARS	-0.0053	-0.0033	-0.0056	-0.0066
Emotions	(0.0035)	(0.0040)	(0.0042)	(0.0045)
Constant	0.4965****	0.3915***	0.5405****	0.3508**
	(0.1071)	(0.1239)	(0.1188)	(0.1451)
Observations	204	208	204	208
R^2	0.1185	0.1186	0.1296	0.2919
χ^2-statistic	31.35	16.22	35.48	73.75

Robust standard errors in parentheses. 'Algo Dummy' is a dummy that takes value one if a person has been assigned to a treatment with an algorithmic advisor and value zero otherwise. Iteration Number equals 1 (2) in the first (last) 30 rounds of the experiment. Details of NARS scales are provided in Appendix A.2. **** p<0.001, *** p<0.01, ** p<0.05, ◊ p<0.1

Table B4. OLS regressions with robust standard errors at the individual level and series fixed effects for 'Follow Ratio' as dependent variable for Human and Algo treatments, separately.

Dependent Variable	(1)	(2)	(3)	(4)	(5)	(6)
	Follow Ratio					
Treatment	Human Bias	Human Random	Human Bias & Random	Algo Bias	Algo Random	Algo Bias & Random
Random Dummy			0.0078			0.0070
			(0.0184)			(0.0211)
CRT	-0.0101	-0.0148**	-0.0126***	-0.0081	-0.0043	-0.0053
	(0.0063)	(0.0067)	(0.0044)	(0.0048)	(0.0096)	(0.0046)
Years of Study	0.0008	0.0128	0.0117	0.0088	0.0345**	0.0171◊
	(0.0125)	(0.0096)	(0.0076)	(0.0096)	(0.0150)	(0.0088)
Risk Aversion	-0.0011	0.0047	0.0023	0.0003	0.0095	0.0032
	(0.0074)	(0.0047)	(0.0036)	(0.0040)	(0.0081)	(0.0033)
NARS	0.0025	0.0020	0.0033	-0.0004	0.0003	-0.0025
Interactions	(0.0040)	(0.0030)	(0.0021)	(0.0026)	(0.0038)	(0.0023)
NARS	-0.0082*	0.0014	-0.0018	0.0007	0.0189****	0.0045**
Social Influence	(0.0049)	(0.0022)	(0.0019)	(0.0026)	(0.0036)	(0.0022)
NARS	0.0019	-0.0039	-0.0028	-0.0041	-0.0128**	-0.0029
Emotions	(0.0056)	(0.0025)	(0.0024)	(0.0027)	(0.0051)	(0.0023)
Constant	0.6493****	0.6094****	0.6197****	0.6734****	0.4053***	0.5721****
	(0.0774)	(0.0816)	(0.0554)	(0.0596)	(0.1200)	(0.0575)
Observations	50	51	101	52	53	105
R^2	0.2939	0.3415	0.2124	0.3104	0.4104	0.2329
F-statistic	1.956	4.008	3.131	2.208	5.174	3.099

Robust standard errors in parentheses. 'Random Dummy' is a dummy that takes value one if a person has been assigned to a Random treatment and value zero otherwise. Details of NARS scales are provided in Appendix A.2.

**** p<0.001, *** p<0.01, ** p<0.05, ◊ p<0.1

Table B5. OLS regressions with robust standard errors at the individual level and series fixed effects for 'Optimal Ratio' as dependent variable across all treatments, including individual controls.

Dependent Variable	(1)	(2)
	Optimal Ratio	
Algo Dummy	-0.0474***	-0.0495***
	(0.0174)	(0.0172)
Bias Dummy	-0.0095	-0.0121
	(0.0286)	(0.0286)
Random Dummy	0.0039	0.0112
	(0.0296)	(0.0279)
Optimal Dummy	0.0620**	0.0681**
	(0.0295)	(0.0281)
CRT		0.0206****
		(0.0048)
Years of Study		0.0085
		(0.0067)
Risk Aversion		-0.0026

		(0.0033)
NARS		-0.0004
Interactions		(0.0019)
NARS		0.0005
Social Influence		(0.0020)
NARS		0.0008
Emotions		(0.0020)
Constant	0.6618****	0.6020****
	(0.0283)	(0.0559)
Observations	356	356
R^2	0.0657	0.1637
F-statistic	4.126	4.763

Robust standard errors in parentheses. 'Algo Dummy' is a dummy that takes value one if a person has been assigned to a treatment with an algorithmic advisor and value zero otherwise. 'Bias Dummy' ('Random Dummy') ['Optimal Dummy'] is a dummy that takes value one if a person has been assigned to a Bias (Random) [Optimal] treatment and value zero otherwise. Details of NARS scales are provided in Appendix A.2.
**** p<0.001, *** p<0.01, ** p<0.05, ◊ p<0.1

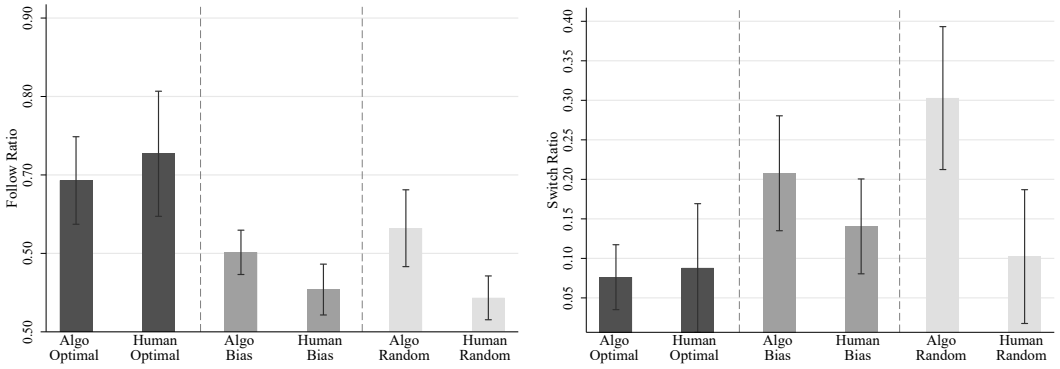


Figure B1. *Follow Ratio (left panel) and Switch Ratio (right panel) across advice treatments for participants scoring in the top quartile on the CRT.*

DOCUMENTS DE TRAVAIL GREDEG PARUS EN 2024
GREDEG Working Papers Released in 2024

2024-01 Davide Antonioli, Alberto Marzucchi, Francesco Rentocchini & Simone Vannuccini
Robot Adoption and Product Innovation

2024-02 Frédéric Marty
Valorisation des droits audiovisuels du football et équilibre économique des clubs professionnels : impacts d'une concurrence croissante inter-sports et intra-sport pour la Ligue 1 de football

2024-03 Mathieu Chevrier, Brice Corgnet, Eric Guerci & Julie Rosaz
Algorithm Credulity: Human and Algorithmic Advice in Prediction Experiments