Journal of Economic Psychology 102 (2024) 102727

Contents lists available at ScienceDirect

Journal of Economic Psychology

journal homepage: www.elsevier.com/locate/joep

Beware the performance of an algorithm before relying on it: Evidence from a stock price forecasting experiment[☆]

Tiffany Tsz Kwan Tse [a,*], Nobuyuki Hanaki [a,b], Bolin Mao [c]

[a] Institute of Social and Economic Research, Osaka University, Japan
[b] University of Limassol, Cyprus
[c] Kyoto Institute of Economic Research, Kyoto University, Japan

ARTICLE INFO

JEL classification:
C90
G1
G4
G17

Keywords:
Algorithms
Financial market
Forecasting
Modification
Technology adoption

ABSTRACT

We experimentally investigated the relationship between participants' reliance on algorithms, their familiarity with the task, and the performance level of the algorithm. We found that when participants were given the freedom to submit any number as their final forecast after observing the one produced by the algorithm (a condition found to mitigate algorithm aversion), the average degree of reliance on high and low performing algorithms did not significantly differ when there was no practice stage. Participants relied less on the algorithm when there was practice stage, regardless of its performance level. The reliance on the low performing algorithm was positive even when participants could infer that they outperformed the algorithm. Indeed, participants would have done better without relying on the low performing algorithm at all. Our results suggest that, at least in some domains, excessive reliance on algorithms, rather than algorithm aversion, should be a concern.

1. Introduction

The use of artificial intelligence (AI) pervades various spheres of society, including financial markets, as noted, for example, by the OECD (2019). In both academia and industry, there is a growing trend of investigating and applying AI to predict stock prices (Bank of England, 2019; Gu, Kelly, & Xiu, 2020; Henrique, Sobreiro, & Kimura, 2019; Kolanovic & Krishnamachari, 2017) and to trade (Lewis, 2014; Liu et al., 2020; Meng & Khushi, 2019). Such a rise in the use of AI allows investors to utilize advice generated by AI in addition to their own judgment in making various decisions. Despite the widespread use of algorithms in financial transactions, as demonstrated by the prevalence of algorithmic trading, it is not yet well understood how individual investors trust and utilize AI in their decision-making. As strategic interactions between humans and algorithms are a worthwhile topic (March, 2021), in this paper, we investigate the extent to which individuals rely on inputs from AI (an algorithm) in forecasting stock prices.

The literature disagrees about people's tendency to rely on algorithms in making decisions in various domains, such as medical recommendations (Promberger & Baron, 2006), predicting joke funniness (Yeomans, Shah, Mullainathan, & Kleinberg, 2019), and forecasting future stock prices (Önkal, Goodwin, Thomson, Gönül, & Pollock, 2009). On the one hand, Dietvorst, Simmons, and Massey (2015, 2018) coined the term "algorithm aversion" to describe people's tendency not to rely on an algorithm's output after

[☆] We gratefully acknowledge financial support from the Joint Usage/Research Center at the Institute of Social and Economic Research, Osaka University and Japan Society for the Promotion of Science KAKENHI Grant Numbers JP18K19954 , JP20H05631. The experiment reported in this paper was approved by the Research Ethics Committee at the Institute of Social and Economic Research, Osaka University. The datasets and codes used to perform analyses are available at https://osf.io/jhqus/.

[*] Correspondence to: Institute of Social and Economic Research, Osaka University, 6-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan.
E-mail addresses: tiffany.econ@gmail.com (T.T.K. Tse), nobuyuki.hanaki@iser.osaka-u.ac.jp (N. Hanaki), mao.bolin.5m@kyoto-u.ac.jp (B. Mao).

https://doi.org/10.1016/j.joep.2024.102727
Received 20 October 2022; Received in revised form 8 March 2024; Accepted 9 March 2024
Available online 15 March 2024

learning that they are imperfect. On the other hand, Logg, Minson, and Moore (2019) presented evidence of “algorithm appreciation” in tasks such as human weight estimation, forecasting song rank, and forecasting human face attraction when asked to choose between following the advice from algorithms and that from other people. Logg et al. (2019) noted that the “algorithm aversion” found in prior studies may simply be a manifestation of advice aversion, i.e., people’s general tendency to rely more on their own judgments than those of others, irrespective of whether these others are other people or algorithms (Yaniv & Kleinberger, 2000). Castelo, Bos, and Lehmann (2019) argued that the degree of reliance on algorithms can be task dependent by showing evidence that algorithms are appreciated more for objective tasks that involve cognitive ability than for subjective tasks that involve emotional ability. Schniter, Shields, and Sznycer (2020) suggested that participants’ level of trust between human partners and robot partners can be economically similar but emotionally different. Farjam (2019) proposed that participants exhibit a preference for uncertainty generated by computers over humans, even when the probability and expected outcome remain identical.

In many of these studies, participants in experiments were not given any information about the algorithm performance or opportunities to experience the task themselves before deciding whether to rely on the algorithm. For example, two studies that investigated algorithm reliance in forecasting future stock prices (Castelo et al., 2019; Önkal et al., 2009) did not give participants the opportunity to experience the task and compare their own and the algorithm’s performance before deciding how much to rely on the algorithm. Thus, participants’ reluctance to rely on the algorithm (Önkal et al., 2009) as well as their willingness to rely on it (Castelo et al., 2019) may simply be due to differences in participants’ subjective judgment about their own skills relative to those of the algorithm in the specific tasks studied, as suggested by the task dependency of reliance on algorithms (Castelo et al., 2019).

To our knowledge, one of the few exceptions is Dietvorst et al. (2015) in which participants were given the opportunity to directly compare their own and the algorithm’s performance before deciding on how much to rely on the algorithm. It was found that participants were especially averse to the algorithm after seeing it made errors, even when participants observed that it outperformed humans. However, the degree of algorithm reliance when participants learned that they outperformed the algorithm was not investigated in that study.

This leads to the following questions that we address in this paper.

> **R1**: *Does the degree of reliance on algorithms by participants who have no experience in the specific task vary depending on the information regarding the performance level of the algorithm?*
>
> **R2**: *How does experiencing and learning about their own skill in the given task influence participants’ degree of reliance on algorithms?*

R1 concerns the effect of information regarding the algorithm’s performance on the participants’ algorithm reliance when they are uncertain about their own skill in the specific task. R2 is about the impact on algorithm reliance when participants gain experience and are able to directly compare their own and the algorithm’s performance.

We addressed these questions by conducting a set of experiments in which participants forecast stock prices. Our experiments included both between-subject design and within-subject design. For the between-subject design, participants were given information about the overall performance of the algorithms to control for their subjective beliefs. In addition, we varied the performance level of the algorithms (high vs. low) and whether participants were able to learn about their own performance during the practice stage. We also compared cases where participants learned only about their own performance in the practice stage with cases where they could directly compare their own and the algorithm’s performance during the practice stage.

For the within-subject design, there were two main tasks. In task 1, participants first made a forecast and, after observing the advice (i.e., the forecast) from an algorithm, then decided which forecast, their own or that of the algorithm, to submit as the final forecast. Task 2 was similar to task 1, except that after seeing the algorithm’s forecast, participants could freely adapt their initial forecast, and choose a final forecast, without being constrained to choose between their initial forecast and that of the algorithm (as they did in task 1).

We found that the degree of reliance on the algorithms did not differ depending on the performance level of the algorithm when there was no practice stage (and thus, with little idea about their own skill). Participants who had experienced the task and learned about their own skill in the practice stage relied on the algorithm significantly less than those without entering the practice stage, both when they could infer that they outperformed the algorithm and when they could infer that the algorithm outperformed them.

Interestingly, in terms of average forecasting performance, participants relied just enough on the high performing algorithm in our experiment (where increasing their reliance would not have resulted in significantly better forecasting performance), but they relied too much on the low performing algorithm in that they would have done better without the algorithm. Although recent research has been concerned with how one can mitigate the aversion to algorithms (e.g., Dietvorst et al. (2018)), our results suggest that at least in some domains, one should also be concerned about the excessive reliance on possibly low performing algorithms.

The remainder of the paper is organized as follows. Section 2 summarizes the existing literature on algorithm reliance by considering the way that information regarding the algorithm’s performance is provided, Section 3 presents the experimental design and hypotheses, and Section 4 summarizes the results. Section 5 provides a discussion and Section 6 concludes.

2. Literature review

Algorithms outperform humans in many fields but they can also make mistakes. As noted, in most existing experimental studies related to estimating or forecasting, participants were not provided with information regarding the accuracy of the algorithm’s estimates or forecasts. In some studies in which participants were provided with information about the algorithm’s performance, the algorithms were always designed to outperform humans (Bigman & Gray, 2018; Castelo et al., 2019; Dietvorst et al., 2018); thus, the degree of reliance on those algorithms that are outperformed by humans is an issue that has not been investigated. Furthermore, most studies did not provide the opportunity for participants to learn from their own performance in the specific task, the only exception being Dietvorst et al. (2015, 2018), in which data were collected on participants' own performance levels. (See Appendix Table A.1 for the summary of existing studies related to reliance on algorithms based on how information regarding the algorithm's performance was provided)

The literature on algorithm reliance can be divided into three categories depending on the provision of information on algorithm performance: (1) no information on algorithm performance is provided; (2) only general information on algorithm performance is provided; and (3) feedback about algorithm performance in the practice tasks is provided. While many of these studies consider only one performance level of the algorithm, there are studies that vary it. We consider those studies that vary the performance level of the algorithm as a separate category although it is not strictly about information provision.[1]

The first category does not provide any information on the performance of the algorithms or human advisors. The main purpose of this approach is to reduce the confounding effects of such information on decision-making (Jussupow et al., 2020; Logg et al., 2019). Many studies have reported evidence that participants tended to rely more on inputs from other people than on algorithms (Önkal et al., 2009; Promberger & Baron, 2006; Yeomans et al., 2019). By contrast, Logg et al. (2019) found that participants tended to rely more on algorithms than on other people. Dietvorst et al. (2015) also found that participants relied more on algorithms than other people in their control condition in their Study 4. One of the possible reasons for these mixed results is that participants were uncertain about their own performance and therefore, their reliance on the algorithms depended mainly on their perceptions regarding the relative performance of humans and algorithms.

The second category provides general or overall information on algorithm performance. Numerous studies have reported the percentage error that defined the accuracy of the judgments of each algorithm, and most of these used the same accuracy rate for the advice from both algorithms and humans to test the impact of human nature (Gray, Gray, & Wegner, 2007; Haslam, 2006) on algorithm reliance. Some evidence has been reported that participants preferred to receive advice from humans rather than from algorithms (Bigman & Gray, 2018; Longoni, Bonezzi, & Morewedge, 2019), and (Dietvorst et al., 2018) noted that participants relied more on algorithms when they could slightly adjust the advice given by the algorithm.

The third category provides feedback on algorithm performance in the practice tasks. The main purpose of this approach is to understand the impact of observing the algorithm's failure on the participant's algorithm reliance. Thus, cases were selected with both good and poor performance. Most such studies reported that participants punished the algorithms by relying on them less after seeing them err (Bigman & Gray, 2018; Dietvorst et al., 2015, 2018; Gaudeul, Giannetti, et al., 2021; Prahl & Van Swol, 2017). Bigman and Gray (2018) found that aversion to the algorithms on moral decisions existed even when the participants were informed that the algorithm was successful.

In the fourth category, the performance level of the algorithms is varied; that is, studies designed more than one algorithm, all with different performance levels. Most of these studies did not provide participants with information on the overall algorithm performance but they learned about algorithm performance through observing both good and bad outcomes in the given tasks. Madhavan and Wiegmann (2007) reported that participants relied more frequently on algorithms with higher performance in X-ray luggage-screening tasks. Jussupow et al. (2020) noted that this approach often did not produce clear results on algorithm aversion or algorithm appreciation because participants were not informed about the overall performance of the algorithm.

Our paper is the first study to cover all four approaches in one set of experiments to systematically study what factors most affect the level of reliance on the algorithm. First, we provided participants with information on the overall performance of the algorithm to control for participants' subjective beliefs on algorithm performance. Second, participants could learn about their own performance during the practice stage and compare it with the information on the overall performance level of the algorithm. Third, we included treatments where participants could directly compare their own and the algorithm's performance during the practice stage. Fourth, we varied the performance level of the algorithms.

3. Experimental design and hypotheses

3.1. Main tasks

For each treatment, in main task, participants were asked to play the role of financial advisor to forecast future stock prices. They were told that their company had created an algorithm that was designed to forecast stock prices as follows.

"This algorithm makes future stock price forecasts by learning the historical stock price information from January 1, 2000 to January 1, 2020, of 83 target companies ranked top in their capital market sectors (i.e., basic materials, consumer goods, healthcare, services, utilities, conglomerates, financial, industrial goods, and technology)".

They were also informed about the performance level of the algorithms. Then, they were shown a series of 20 graphs, with 12 months' worth of closing prices of randomly selected stocks from the S&P 500 components, commencing from a randomly selected day between January 1, 2008, and December 1, 2018. The participants were not told the name of the stock or the starting date. Each time series was standardized so that its starting price was equal to 100 (see Fig. 1 for an example).

For each graph, participants were asked to forecast the closing price of the stock 30 days after the last price shown on the graph. This forecasting task followed those used in forecasting experiments reported in Bao et al. (2022), Bao, Corgnet, Hanaki, Riyanto,

[1] Jussupow, Benbasat, and Heinzl (2020) classified the literature based on algorithm performance into three groups: (1) performance information is provided; (2) the performance rate is varied during interaction; and (3) algorithm failures are forced.

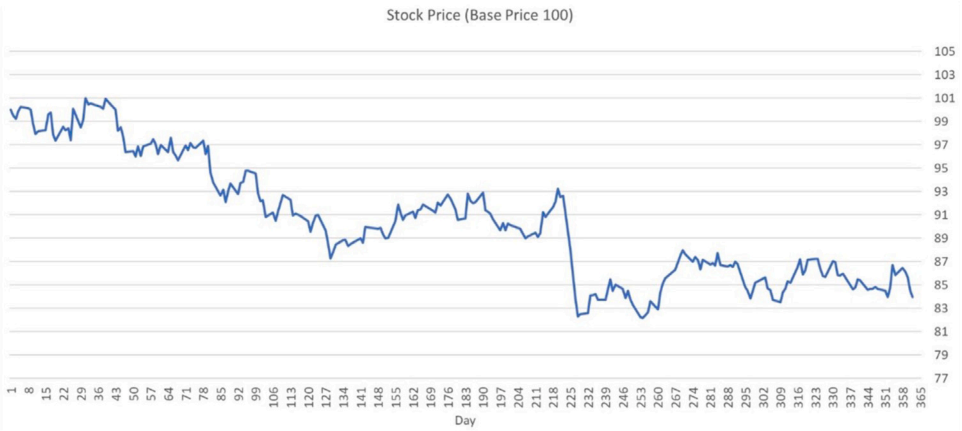


Fig. 1. Sample of the graph.

and Zhu (2023). Participants first entered their forecast for each of the 10 graphs. The order of the display of the 10 graphs was randomized across participants. Then, for the same set of 10 graphs, they were informed of the algorithm's forecast and asked to submit their final forecast, either by selecting between their own forecast and that of the algorithm (task 1), or by freely modifying the forecast (task 2). They were not given feedback about their performance on each graph in the main tasks. The order of the display was different from the order when they entered their initial forecasts and was again randomized across participants. Also, the order of the two tasks was randomized across participants.

At the end of each task, participants were asked to evaluate the accuracy of their forecasts relative to those of the algorithm, based on a scale from –5 (the lowest score, where their forecast was less accurate than the algorithm's forecast to a great extent) and 5 (the highest score, where their forecast was more accurate than the algorithm's forecast to a great extent), with 0 indicating that the participant's forecast had the same accuracy as the algorithm.

Participants were rewarded based on the accuracy of their final forecasts in one randomly chosen graph (out of 20 graphs from two tasks) as follows, where $(\cdot)^+$ denotes $max(\cdot, 0)$.

$$reward = \left(200 - 10 \times \left|\frac{your\ final\ forecast - realized\ price}{realized\ price}\right| \times 100\right)^+$$

If a participant's final forecast in the chosen graph matched the realized price exactly, the participant received 200 points. For each percentage point difference between the participant's final forecast and the realized price, 10 points were subtracted. If the participant's final forecast differed from the realized price by more than 20%, 0 points were awarded. The exchange rate was 1 point = 6 JPY.

3.2. Treatments

3.2.1. Between-subject design

We designed six treatments, varying the performance level of algorithms (high or low) and the opportunity for participants to learn about their own and the algorithms' performance through the practice stage. We refer to the high and low performing algorithms as "good" and "bad" algorithms, respectively.

We measured the performance of the algorithm as well as that of a participant for a particular forecasting task using the absolute percentage error (APE) of their forecast from the realized price using the following equation.

$$APE = \left|\frac{Forecast - realized\ price}{realized\ price}\right| \times 100\%$$

Participants were informed that the mean absolute percentage error (MAPE, defined below) of the algorithm was either around 4.9% (i.e., a good algorithm in Treatments 1, 2, and 3 (hereafter, T1, T2, and T3) or 18.4% (i.e., a bad algorithm in T4, T5, and T6). These MAPEs are based on the performance of the algorithms' test data set which consists of historical stock closing price time series sourced from Yahoo!Finance. (See more details in Online Appendix G)

$$MAPE = \frac{1}{n}\sum\left|\frac{Forecast - realized\ price}{realized\ price}\right| \times 100\%$$

The two types of algorithm, good and bad, were designed to perform, on average, better and worse, respectively, than humans.

To vary the opportunity for participants to learn about their own and the algorithms' performance, we included a practice stage in four of our treatments (T2, T3, T5, and T6). In the practice stage, as in the main task, participants were shown a series of 10 graphs generated in the same way as in the main task and, for each graph, they forecast the closing price for the stock 30 days after

Table 1
Summary of treatments.

Treatment	Algorithms	Practice stage	Number of participants
T1	Good	No practice stage	49
T2	Good	Human	47
T3	Good	Human and algorithm	50
T4	Bad	No practice stage	50
T5	Bad	Human	45
T6	Bad	Human and algorithm	47
Total number of participants			288

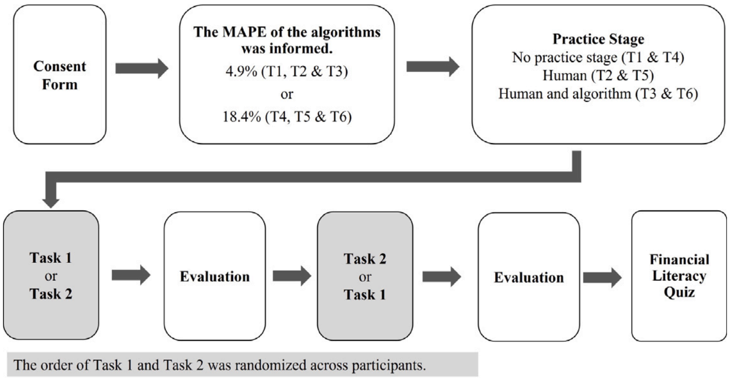

Consent Form
The MAPE of the algorithms was informed.
4.9% (T1, T2 & T3)
or
18.4% (T4, T5 & T6)
Practice Stage
No practice stage (T1 & T4)
Human (T2 & T5)
Human and algorithm (T3 & T6)
Task 1 or Task 2
Evaluation
Task 2 or Task 1
Evaluation
Financial Literacy Quiz
The order of Task 1 and Task 2 was randomized across participants.

Fig. 2. Experiment flow.

the last price shown on the graph. At the end of the practice stage, after participants had finished entering their forecasts for all 10 stocks, we either showed them only their own performance (T2 and T5) or both their own and the algorithm's performance (T3 and T6) for each of the 10 stocks separately, as well as the average across all 10 stocks.

Namely, in T2 and T5, participants were informed of the realized price, their own forecast, and the associated APE for each of 10 stocks, and the MAPE for their own 10 forecasts. In T3 and T6, besides the realized price and their own performance, participants were also informed of the forecast of the algorithm and the associated APE for each of the 10 stocks, and the MAPE of the algorithms' 10 forecasts. There was no practice stage in T1 or T4. See Table 1 for a summary of our six treatments.

3.2.2. Within-subject design

For each treatment, there were two main tasks in which the decision-making methods were different when participants entered their final forecast. In task 1, participants submitted their final forecast either by selecting between their own forecast and that of the algorithm. In task 2, they submitted their final forecast by entering any numbers. The order of the two tasks was randomized across participants.

3.3. Procedure

Fig. 2 demonstrates the flow of the experiment.

First, participants gave online consent by clicking a button, followed by general instructions that informed them of the experimental goals of the main tasks, the information about the algorithms including performance level, and whether they would enter the practice stage or not. In the general instructions, participants were not informed about the sequence of the experiment and the detailed instruction for tasks 1 and 2, but were informed of the MAPE of the algorithm.

After the instructions, participants entered the practice stage in T2, T3, T5 and T6, followed by instructions for the practice stage. There was no practice stage in T1 and T4.

Next, participants entered the main tasks: tasks 1 and 2. The order of the two tasks was randomized across participants. For example, if participants entered task 1 first, then they entered task 2 s and vice versa. Before each task started, participants read the instruction for each task. At the end of each task, participants were asked to evaluate the accuracy of their forecasts relative to those of the algorithm. At the end of the experiment, participants were asked to take the financial literacy quiz. The detailed instructions can be found in Online Appendix I.

3.4. Materials and summary

The experiment was programmed using Qualtrics Survey System. Participants received individual URL links to access the experiment. They can participate in the experiment by using their computers, smartphones and tablets.

The experiment was conducted online from December 1, 2020 to December 7, 2020. We recruited 299 participants who were students of Osaka University registered to the ORSEE (Greiner, 2015) database of the Institute of Social and Economic Research at Osaka University. They received 500 JPY as a participation fee for completing 45 min of experiments, and could earn up to an additional 1200 JPY reward depending on their forecasting performance. We dropped 11 participants (out of 299) from our analyses because they completed the experiment in a very short time (less than 10 min).[2] We also dropped one observation for task 2, Question 9, in which the participant entered a huge number in one forecast possibly due to a typo. In the final sample, 66% of the participants were male, and 81% were undergraduate students, predominantly from the following majors: 37% engineering, 11% economics and management, 10% foreign studies, 9% law, 8% medicine, 7% science, and 8% human science. The final sample had an average financial literacy score of 67% (8 out of 12 questions).

In addition, we gathered information regarding participants' degree of risk aversion and cognitive ability. Participants' characteristics, except for the financial literacy score, were not statistically significantly different across treatments (see Online Appendix B). In the main text, we reported the average treatment effect without controlling for these individual characteristics because we obtained qualitatively similar results even after controlling for them (see Online Appendix B for these additional analyses).

3.5. Hypotheses

We hypothesized that participants did not know their own performance when there was no practice stage. Therefore, they could not compare their own performance with that of the algorithm even when they received information about the overall accuracy of the algorithm in T1 and T4. Their reliance on the algorithm depended on their perception of their own skills relative to that of the algorithm. As a result, the ex ante information about the overall accuracy of the algorithm did not help participants to make decisions on whether to rely on the algorithm. Therefore, we propose the following hypothesis.

Hypothesis 1. There is no difference in the reliance level on the algorithm between T1 and T4.

Participants can learn about their own performance in T2 and T5. They can compare their own performance with the good algorithm in T2 and the bad algorithm in T5. They learn that the algorithm performs better than they do in T2, and worse than they do in T5. Therefore, we propose the following hypothesis.

Hypothesis 2. The reliance level on the algorithm is higher in T2 than in T5.

As noted, the reliance on the bad algorithm depends on participants' perceptions of their own skills and the algorithms in T4. They can learn that they outperform the bad algorithm in T5. Therefore, we propose the following hypothesis.

Hypothesis 3. The reliance level on the algorithm is higher in T4 than in T5.

Similarly, the reliance on the good algorithm depends on participants' perceptions of their own skills and the algorithms in T1. They can learn that their performance is worse than the good algorithm in T2. Therefore, we propose the following hypothesis.

Hypothesis 4. The reliance level on the algorithm is higher in T2 than in T1.

Dietvorst et al. (2018) proposed the concept of algorithm aversion, referring to the fact that people often fail to rely on good algorithms after learning that the algorithms are imperfect. In our experiment, participants receive the same ex ante information about the overall accuracy of the good algorithm (i.e., MAPE = 4.9%) in T2 and T3. However, they receive additional information about the MAPE of the good algorithm in the practice stage (which happens to be worse than the ex ante information; MAPE = 5.89%) in T3. Therefore, we propose the following hypothesis.

Hypothesis 5. The reliance level on the algorithm is higher in T2 than in T3.

Similarly, participants receive the same ex ante information about the overall accuracy of the bad algorithm (i.e., MAPE = 18.4%) in T5 and T6. In addition, they receive information about the MAPE of the bad algorithm in the practice stage (which happens to be better than the ex ante information; MAPE = 10.14%) in T6. Therefore, we propose the following hypothesis.

Hypothesis 6. The reliance level on the algorithm is higher in T6 than in T5.

[2] We conducted a robustness check for the results by including all participants. In T5, one participant completed the experiment in 8 min and misunderstood task 2 by inputting small numbers for the final forecast in 10 questions. We omitted these observations, and obtained similar results.

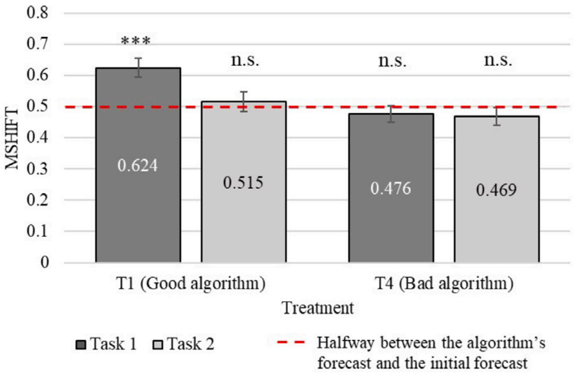


Fig. 3. MSHIFT in T1 and T4. *Notes:* The p values were calculated based on a single-sample t-test. MSHIFTs were compared against the 0.5 level, which is halfway between the algorithm's forecast and the initial forecast. The symbols *, **, and *** indicate significance at the 0.05, 0.01, and 0.001 levels, respectively, and n.s. means that the difference is not statistically significant at the 0.05 level. See Table A1 in Online Appendix A for details.

4. Results

4.1. MSHIFT calculation

We measured the degree of "reliance on algorithms" (Castelo et al., 2019; Logg et al., 2019) by the "shift rate" (Önkal et al., 2009), which is defined for participant i in relation to stock s, as follows.

$$Shift\ Rate_s^i = \frac{Final\ Forecast_s^i - Initial\ Forecast_s^i}{Algorithm's\ Forecast_s - Initial\ Forecast_s^i}$$

A shift rate that is >0.5 indicates that the final forecast is closer to the algorithm's forecast than the participant's own initial forecast. The opposite is true for a shift rate that is <0.5. A shift rate of 1 indicates that the final forecast is exactly the same as the algorithm's forecast, while a shift rate of 0 indicates that the final forecast is exactly the same as the participant's initial forecast. We calculated the mean shift rate (MSHIFT) of 10 graphs in each task in each treatment.

Our discussion is organized as follows. We first compared the degree of reliance on the algorithm when participants were only informed about the average performance level of the algorithm without experiencing the task (T1 vs. T4). We also compared reliance on the algorithm between task 1, when participants had to choose between either their own forecast or that of the algorithm as the final forecast, and task 2, when there was no such restriction regarding the choice of final algorithm. Then, for both types of algorithm, we investigated the effect on participants of experiencing the task and comparing their own performance with the average performance of the algorithm (T2 and T5), or comparing their own and the algorithm's performance side by side (T3 and T6). All the reported results were tested by two-tailed tests, and similar results were obtained by conducting one-tailed tests. Bonferroni-adjusted p-values were reported for the comparison of results among treatments.

4.2. Effect of information on algorithm performance when there is no practice stage

Fig. 3 shows the average MSHIFT in T1 and T4 for task 1 (dark gray) and task 2 (light gray). The error bars correspond to the two standard error range (i.e., average ±one standard error). The average MSHIFTs for task 1 were 0.624 in T1 and 0.476 in T4; for task 2, they were 0.515 in T1 and 0.469 in T4. The MSHIFT was significantly different from 0.5 only in task 1 of T1.

The task 2 results showed that when participants can choose their final forecasts freely, regardless of the average performance level of the algorithm provided (the MAPEs of the algorithm were 4.9% in T1 and 18.4% in T4), on average, they chose a point midway between their own forecast and that provided by the algorithm. When participants had to choose between the two as their final forecasts in task 1, for the bad algorithm they were equally likely to choose the algorithm's or their own initial forecast; for the good algorithm, they were more likely to choose the forecast provided by the good algorithm (on average, 0.15 more likely than was the case for the bad algorithm). This suggests that, when there was no practice stage, for those participants without a good idea about their own performance, information on the performance level of the algorithm did not have a strong effect on their reliance on the algorithm.

Participants considered their forecasts to be slightly less accurate than those of the algorithm in both T1 and T4 (see Fig. 4). The average subjective evaluations of the accuracy of their own forecasts relative to those of the algorithm were –1.041 (task 1) and –0.388 (task 2) in T1, and –0.7 (task 1) and –0.54 (task 2) in T4. As shown in Fig. 4, there was no statistically significant difference between the subjective evaluations between T1 and T4 in either of the two tasks. As implied by the similar degree of reliance on the algorithms in T1 (good algorithm) and T4 (bad algorithm), participants' final forecasts became better than their initial forecasts in T1, but worse in T4, as shown in Fig. 5.

Result 1. *When there was no practice stage, participants relied more on the good algorithm than on the bad algorithm in task 1, but not in task 2. Thus, Hypothesis 1 was rejected in task 1, but not in task 2.*

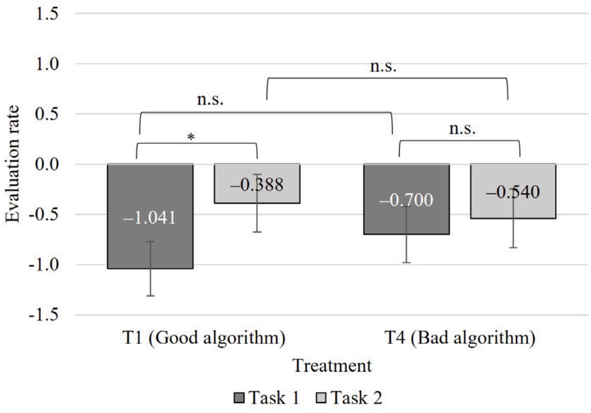


Fig. 4. Evaluation of the accuracy of the initial forecast relative to the algorithm's forecast in T1 and T4. *Notes:* When we compared the evaluation rate between T1 and T4, we regressed the evaluation rate on a treatment dummy (0 if T1 and 1 if T4) by OLS regression model with robust standard errors. When we compared the evaluation rate between task 1 and task 2, we regressed the evaluation rate on a task dummy (0 if task 1 and 1 if task 2) by OLS regression model with robust standard errors clustered by individual participants. The symbols *, **, and *** indicate significance at the 0.05, 0.01, and 0.001 levels, respectively, and n.s. means that the difference is not statistically significant at the 0.05 level. See Tables A2 and A3 in Online Appendix A for details.

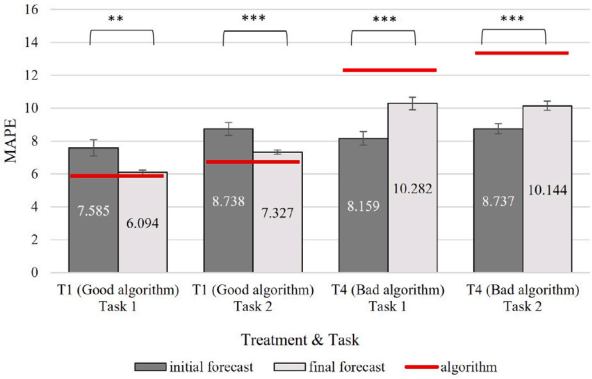


Fig. 5. MAPE in T1 and T4. *Notes:* We regressed the MAPE on final forecast dummies by OLS regression model with robust cluster standard error on participant level. The figure shows the p values for the estimated coefficient on final forecast dummies. The symbols *, **, and *** indicate significance at the 0.05, 0.01, and 0.001 levels, respectively, and n.s. means that the difference is not statistically significant at the 0.05 level. See Tables A2, A4, and A5 in Online Appendix A for details.

4.3. Effect of information on algorithm performance when there is practice stage

Now, we turn to the effect of letting participants experience the task and informing them about their performance in the practice stage. In T2 and T5, participants were only informed about their own performance at the end of the practice stage. The average MAPEs of participants (and the standard errors) during the practice stage were 8.300% (0.578%) and 8.100% (0.386%) in T2 and T5, respectively. Therefore, participants in T2 were aware that the algorithm (with a MAPE of 4.9%) outperformed them on average, and participants in T5 were aware that they outperformed the algorithm (with a MAPE of 18.4%) on average. Fig. 6 shows the MSHIFT in task 1 (dark gray) and task 2 (light gray) in each treatment. The results of T1 and T4 are included for reference. We found that MSHIFT in T2 was much higher than in T5 in task 1 ($F(1, 282) = 59.547$, $p < 0.001$) and task 2 ($F(1, 282) = 47.027$, $p < 0.001$).

Result 2. *Participants relied more on the good algorithm than on the bad algorithm after they learned that the algorithm outperformed humans in T2 and underperformed humans in T5. Thus, Hypothesis 2 was supported in both tasks*

Regardless of the performance level of the algorithm, we observed that allowing participants to gain experience and learn about their own performance level on the specific task decreased their reliance on the algorithm on average. We found that MSHIFT in T5 was lower than in T4 in task 1 ($F(1, 282) = 38.989$, $p < 0.001$) and task 2 ($F(1, 282) = 33.702$, $p < 0.001$). However, MSHIFT in T2 was lower than in T1 in task 1 ($F(1, 282) = 12.055$, $p = 0.002$), and with no significant difference in task 2 ($F(1, 282) = 3.268$, $p = 0.215$).

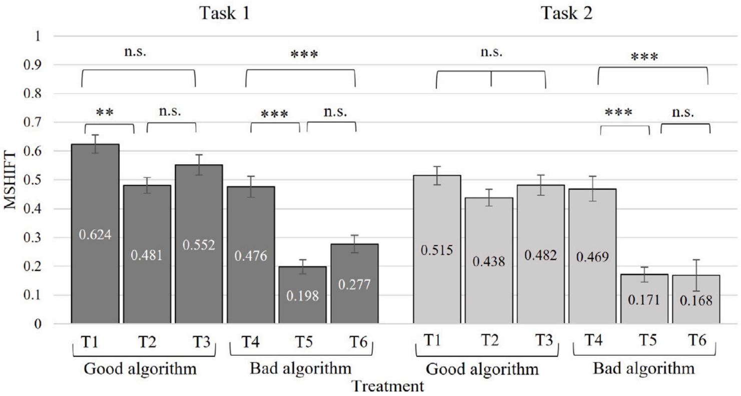


Fig. 6. MSHIFT in tasks 1 and 2. *Notes:* We regressed the MSHIFT on six treatment dummies by OLS regression model with robust standard errors, and compared the estimated dummy coefficients by F test, with comparing result illustrated by Bonferroni-adjusted p values. The symbols *, **, and *** indicate significance at the 0.05, 0.01, and 0.001 levels, respectively, and n.s. means that the difference is not statistically significant at the 0.05 level. See Table A2 in Online Appendix A for details.

Result 3. *Participants relied less on the bad algorithm after they learned that they outperformed the bad algorithm. Thus, Hypothesis 3 was supported in both tasks.*

Result 4. *Participants relied less on the good algorithm after they learned that the good algorithm outperformed them. Thus, Hypothesis 4 was not supported in either task.*

In T3 and T6, participants could directly compare the performance of their own forecasts with those of the algorithm. The average MAPEs (and the standard errors) during the practice stage were 8.064% (0.386%) for the participants and 5.889% for the algorithm in T3, and 7.861% (0.359%) for the participants and 10.144% for the algorithm in T6. Note that the MAPEs of the algorithm in the practice stage of T3 and T6 were both quite different from those seen by participants in the instructions (4.9% and 18.4%). This is because the MAPEs of the algorithms in the instructions were computed based on the large sample of the trials, and not on the small samples of the specific stock periods used in the experiment. However, this discrepancy could have resulted in participants considering the good algorithm to perform poorly in T3 in comparison with T1 and T2 (and thus to rely on the good algorithm less in T3 than in T2), or the bad algorithm to perform better in T6 compared with T4 and T5 (and thus to rely on the bad algorithm more in T6 than in T5).

Regardless of the performance level of the algorithm, on average, in task 1, participants' reliance on the algorithm increased when they were able to directly compare their own forecasts with those of the algorithm. MSHIFT increased, although not significantly, from 0.481 in T2 to 0.552 in T3. Similarly, MSHIFT increased significantly from 0.198 in T5 to 0.277 in T6. However, in task 2, MSHIFTs were similar between T2 and T3 (0.435 and 0.482, respectively) and between T5 and T6 (0.171 and 0.168, respectively).

Result 5. *Participants did not change their reliance level on the good algorithm after observing its performance in the practice stage, which was worse than its overall accuracy. Thus, Hypothesis 5 was not supported in either task.*

Result 6. *Participants relied more on the bad algorithm in task 1 after observing its performance in the practice stage, which was better than its overall accuracy, but this result was not observed in task 2. Thus, Hypothesis 6 was supported in task 1, but not in task 2.*

The significantly lower reliance on the algorithm observed in T2 and T5 compared with T1 and T4, respectively, suggested that, on average, participants who did not experience the task in the practice stage (in T1 and T4) expected their performance to be worse than the 8% MAPE (the average MAPE achieved by participants during the practice stage in T2 and T5). This interpretation was corroborated by their subjective evaluation of the accuracy of their own forecasts relative to those of the algorithm, as shown in Fig. 7. The subjective evaluation of their own forecasts slightly improved from –1.041 in T1 to –0.596 in T2, and there was a much greater improvement from T4 to T5 (–0.7 to 1.178). Indeed, there was a positive (and statistically significant) relationship between MAPE during the practice stage and MSHIFT in T2. That is, those who performed poorly (indicated by a higher MAPE) relied more on the good algorithm. For T5, however, we did not observe such a relationship (see Table D1 in Online Appendix D).

The significant increase in reliance on the algorithm in T6 compared with T5 in task 1 can be understood in terms of the effect of the discrepancy between the MAPE of the algorithm communicated to participants in the instructions (18.4%) and what they observed during the practice stage (10.14%). Recall that in T6, the algorithm performance in the practice stage was higher than it had been introduced to the participants in the beginning. (This was the only information participants received about the algorithm in T5.) In T3, although the algorithm performance in the practice stage was lower (MAPE = 5.89%) than it had been introduced to the participants in the beginning (MAPE = 4.9%), this difference was not sufficient to result in a significant difference in MSHIFT between T2 and T3.

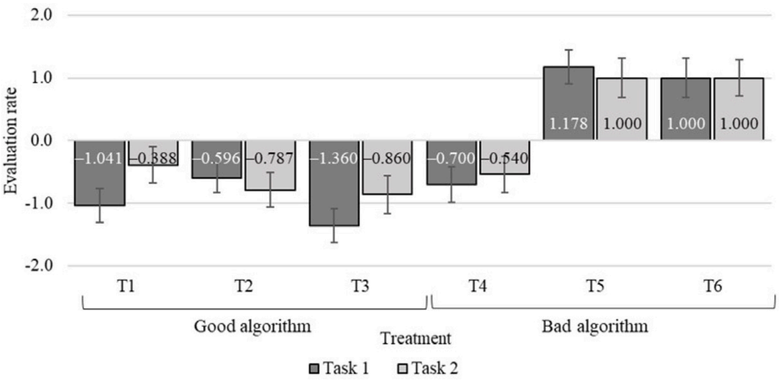


Fig. 7. Evaluation of the accuracy of participants' initial forecast relative to the algorithm's forecast.

Differences in MSHIFT across the treatments that we observed resulted in variations in performance of the final forecasts, measured by MAPE, as shown in Fig. 8(a) for task 1 and Fig. 8(b) for task 2. The figures show the MAPE of the initial forecast, as well as that of the algorithm (the red line). We first discuss the results of task 1, shown in Fig. 8(a)).

We observed some improvement in participants' initial forecasts after the practice stage. The MAPE of the initial forecasts was 7.585% in T1, 6.750% in T2, 6.548% in T3 (although differences were not significantly different), 8.159% in T4, 6.551% in T5, and 7.147% in T6. The difference between T4 and T5 was significant.

The MAPEs of the final forecasts were 6.112% in T2 and 5.806% in T3, which were significantly lower than those of the initial forecasts. Furthermore, the MAPE of final forecasts in T3 was not significantly different from that of the algorithms ($t(49) = 0.500$, $p = 0.619$, see Table E1 in Online Appendix E for details). The significantly lower reliance on the algorithm in T2 compared with T1 did not result in significantly worse forecasts.

By contrast, participants relied too much on the low performing algorithm. The MAPEs of the final forecasts in T5 and T6 were 7.817% and 8.024%, respectively. Although they were significantly lower than in T4 (10.282%) due to both better initial forecasts and lower reliance on the low performing algorithm, they were still significantly higher than participants' initial forecasts. Thus, participants would have been better off without the algorithm.

Similar observations can be made for task 2, as shown in Fig. 8(b). In particular, participants' final forecasts were significantly worse in terms of MAPEs than their initial forecasts in the presence of the low performing algorithm.

5. Discussion

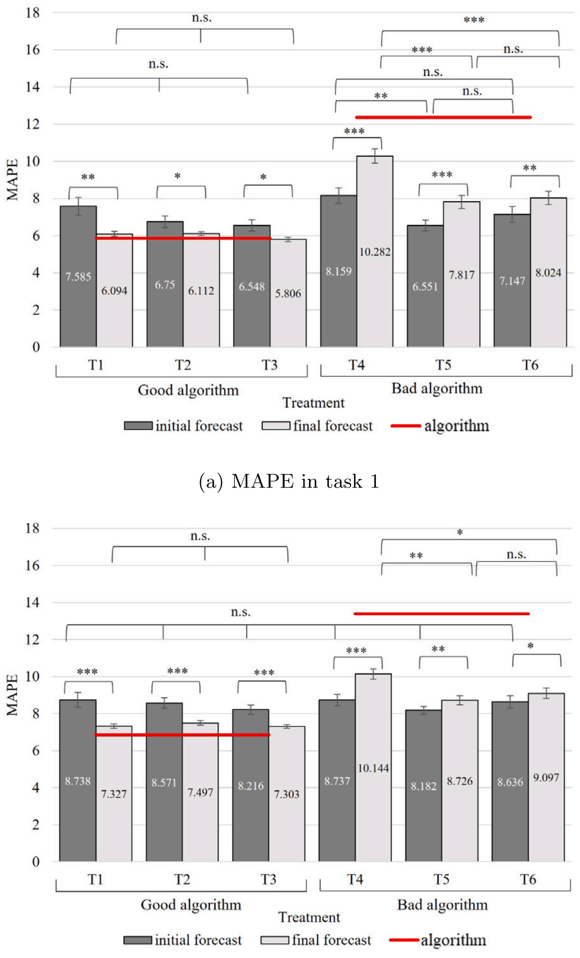
In our experimental design, the decision-making methods as well as the graphs of the stock price time series differ between tasks 1 and 2. Therefore, we focus on testing the hypotheses in tasks 1 and 2 separately, and not comparing the results between tasks 1 and 2. In the following, we discuss the possible reasons why some hypotheses are not supported in either task.

Hypothesis 4 was not supported in either task. Participants were informed about the overall performance of the good algorithm in T1, T2, and T3, for which the MAPE was 4.9%. In T2, when participants gained experience in the practice stage and learned that their own performance level was worse than the overall performance of the good algorithm, they still relied less on the good algorithm, which demonstrates "algorithm aversion".

Hypothesis 5 was also not supported in either task. In T2, participants could compare their own performance in the practice stage (MAPE = 8%) with the overall performance of good algorithms (MAPE = 4.9%). In T3, participants could compare their own performance level (MAPE = 8%) with the performance of good algorithms in the practice stage (MAPE = 5.89%). The performance of the good algorithm in the practice stage was slightly worse than its overall performance. However, during the practice stage, participants observed that the good algorithm outperformed them when they received feedback from each outcome in T3. As a result, reliance on the good algorithm did not significantly differ between T2 and T3.

Someone may think that participants tended to choose the middle between their initial forecast and the algorithm forecast when submitting their final forecast in task 2, due to the compromise effect. If the compromise effect affected the reliance on the algorithm, the MSHIFT should be higher in task 2 than in task 1 when participants receive advice from the bad algorithm. However, there was no significant difference in MSHIFT between task 1 and task 2 in T4 ($t(50) = 0.198$, $p = 0.844$), T5 ($t(45) = 1.025$, $p = 0.311$) and T6 ($t(47) = 1.990$, $p = 0.053$). Therefore, the compromise effect did not influence participants making decision in task 2.

In the experimental instructions, we asked participants to play the role of "financial advisor" and informed them that "their company had created an algorithm". These framings may have induced them to rely more heavily on the algorithm. To address such concerns, we conducted an additional set of experiments without these framings. We found no significant difference between the results of the framed and nonframed experiments for all but one treatment. Even in that treatment, the degree of reliance on



(a) MAPE in task 1

(b) MAPE in task 2

Fig. 8. MAPE in tasks 1 and 2. *Notes:* We regressed the MAPE on six treatment dummies by OLS regression model with robust standard errors, and compared the estimated dummy coefficients by F test, with comparing result illustrated by Bonferroni-adjusted p values. The symbols *, **, and *** indicate significance at the 0.05, 0.01, and 0.001 levels, respectively, and n.s. means that the difference is not statistically significant at the 0.05 level. See Tables A2, A4, and A5 in Online Appendix A for details.

the algorithm was higher in the nonframed experiment than in the framed version. Therefore, we concluded that the results that we report in the main text were not driven by these frames in the experimental instructions. See Online Appendices J and K for details.

6. Conclusion

In this paper, we reported the results of a set of controlled online experiments on forecasting stock prices, exploring (1) whether the degree of reliance on algorithms by participants who had no experience in the specific task varied depending on the performance level of the algorithm, and (2) how participants' gaining experience and learning about their own skill in the given task influenced their degree of reliance on the algorithm.

We found that for those participants without entering the practice stage (and thus, with no idea about their own skill), the degree of reliance on the algorithm did not differ significantly between good and bad algorithms when participants were free to adjust their forecasts after receiving the algorithm's forecast. Those participants who had experienced the task and learned about their own skill in the practice stage relied on the algorithm significantly less than those without entering the practice stage, both when they could infer that they outperformed the algorithm and when they could infer that the algorithm outperformed them. In terms of the average forecasting performance, participants relied on the high performing algorithm in our experiment that indeed brought prediction improvement in many cases. However, they relied too much on the low performing algorithm, even when they could infer that they outperformed the algorithm; in this case, they would have done better without relying on the algorithm at all. While recent research has been concerned with how the aversion to algorithms can be mitigated (e.g., Dietvorst et al. (2018)), our results suggest that at least in some domains, one should also be concerned about the excessive reliance on algorithms.

This study leaves some questions unanswered. First, we did not investigate the dynamics of algorithm reliance. It is possible that if participants learned about the performance of the algorithm relative to their own performance, they might increase their reliance on good algorithms and decrease their reliance on bad ones. Thus, excessive reliance on low performing algorithms may simply be a temporary phenomenon. Second, in our experiment, the advice from the algorithm was provided for free. Yet, in many situations, information has value, and one needs to pay to obtain it. It is possible that if participants have to pay for advice from an algorithm, they may refuse to pay for advice from low performing algorithms, thus solving the problem of excessive reliance on them. Therefore, it is of great interest to investigate how well participants assess the value of the advice coming from algorithms. We plan to investigate these issues in future research.

Data availability

The data is publically available via OSF.

Appendix A. Literature review

See Table A.1.

Table A.1
Existing studies on the algorithm performance information.

Type	Research	Tasks [Source of the advice]	Results	Information about algorithm performance [The algorithms outperformed/underperformed humans]	Participants learnt about their performance [Measured]	Forecasting tasks [Initial forecasts were measured]
1	Logg et al. (2019)	Weight estimation, song rank forecast, attraction forecast, researcher prediction [Algorithms and other people]	Relying more on the algorithms than other people and own estimate	No [Same]	No	Yes [Yes]
1	Önkal et al. (2009)	Stock price forecast [Algorithms and experts]	Relying less on the algorithms than experts	No [Same]	No	Yes [Yes]
1	Promberger and Baron (2006)	To take advice about medical operations [Algorithms and physician]	Relying less on the algorithms than the physician	No [Same]	No	Yes [No]
1	Yeomans et al. (2019)	Joke funniness prediction [Algorithm and other people]	Relying less on the algorithms than other people	No [Outperformed]	No	Yes [No]
2	Bigman and Gray (2018)	To rate algorithms or humans making morally relevant driving, legal, medical, and military decisions [Algorithms and other people]	Averse to the algorithms making moral decisions	Description of the algorithms with positive or negative outcomes (Study 5) [Same] Overall performance with accuracy percentage (Study 9) [Outperformed]	No	No
2	Castelo et al. (2019)	To choose between relying on the algorithms or other people in various tasks (Study 3) [Algorithms and other people] Stock price forecast (Study 6) [Algorithm]	Relying more on the algorithms when the algorithm's performance information was provided (Study 3) Reliance on the algorithms was higher under objective framing than subjective framing (Study 6)	Participants were informed that "the algorithm outperforms humans" (Study 3) [Outperformed] High/low human likeliness and subjective/objective framing (Study 6) [Unknown]	No	No (Study 3) Yes [Yes] (Study 6)
2	Longoni et al. (2019)	To sign up for the stress assessment analyzed by algorithms or physician (Study 1) [Algorithms and physician]	More frequently signing up the stress assessment analyzed by physician than algorithms (Study 1)	Overall performance with accuracy percentage [Same] (Study 1)	No	No

(continued on next page)

Table A.1 (*continued*).

Type	Research	Tasks [Source of the advice]	Results	Information about algorithm performance [The algorithms outperformed/underperformed humans]	Participants learnt about their performance [Measured]	Forecasting tasks [Initial forecasts were measured]
2&3	Dietvorst et al. (2018)	To predict students' performance To choose forecasting processes (Study 3) [Algorithms]	Relying more on the algorithms than humans after adjustment was allowed More frequently choosing the forecasting process in which adjustment was allowed (Study 3)	Overall performance with accuracy percentage [Outperformed] Feedback from the algorithms and the overall performance of the algorithms (Study 3) [Outperformed]	No (Study 1, 2) Yes [Yes] (Study 3)	Yes [Yes]
3	Dietvorst et al. (2015)	To predict students' performance (Studies 1, 2 & 4) To predict the rank of individual US states in terms of the number of airline passengers (Study 3) [Algorithms (Studies 1–3); Algorithms & other people (Study 4)]	Relying less on the algorithms than their own estimates (Studies 1–3) and estimates from other people (Study 4) after seeing the results of the algorithm's forecasts	No information in the control condition. Feedback from the algorithms, but not the overall accuracy percentage [Outperformed]	Yes [Yes]	Yes [Yes]
3	Gaudeul et al. (2021)	To trade in the stock market [Various algorithms]	Only a small minority of participants decided to rely on algorithms after having tried them.	Feedback from the algorithms, but not the overall accuracy percentage [Outperformed]	Yes [Yes]	No
3	Goodyear et al. (2017, 2016)	To detect knives on X-ray luggage screening after receiving advice [Algorithms and experts]	Relying more on the algorithms than experts	Feedback from the algorithm with low accuracy, but not the overall accuracy percentage [Same]	Yes [No]	Yes [No]
3	Prahl and Van Swol (2017)	To predict the number of orthopedic surgeries in the future. [Algorithms and experts]	No significant difference in algorithm utilization between algorithms and experts on average. After receiving severe errors, utilization of algorithms' advice decreased significantly more than experts' advice.	Feedback from the algorithms, but not the overall accuracy percentage [Same]	Yes [Yes]	Yes [Yes]
4	Madhavan and Wiegmann (2007)	To detect concealed weapons on X-ray luggage screening after receiving advice (Study 2) [Experienced algorithms, novice-like algorithms, experts and non-experts]	Relying more on the algorithms with high accuracy than on those with low accuracy	Feedback from the algorithms with high or low accuracy, but not the overall accuracy percentage (Study 2) [Same]	Yes [No]	Yes [No]

Notes: Type 1 indicates that the literature does not provide any information about the performance of the algorithms and human advisors. Type 2 indicates that the literature provides information about the overall performance level of the algorithm. Type 3 indicates that the literature provides feedback about the algorithms' performance in the practice tasks. Type 4 indicates that the literature varies the performance level of the algorithms.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.joep.2024.102727.

References

Bao, T., Corgnet, B., Hanaki, N., Okada, K., Riyanto, Y. E., & Zhu, J. (2022). *Financial forecasting in the lab and the field: qualified professionals vs: Smart Students ISER DP 1156*, Institute of Social and Economic Research, Osaka University.

Bao, T., Corgnet, B., Hanaki, N., Riyanto, Y. E., & Zhu, J. (2023). Predicting the unpredictable: New experimental evidence on forecasting random walksom walks. *Journal of Economic Dynamics & Control*, *146*, Article 104571.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170.

Farjam, M. (2019). On whom would I want to depend; humans or computers? *Journal of Economic Psychology*, *72*, 219–228.

Gaudeul, A., Giannetti, C., et al. (2021). *Fostering the adoption of robo-advisors: a 3-weeks online stock-trading experiment*: *Discussion paper n. 275.*, Dipartimento di Economia e Management (DEM), University of Pisa.

Goodyear, K., Parasuraman, R., Chernyak, S., de Visser, E., Madhavan, P., & Deshpande, F. (2017). An fMRI and effective connectivity study investigating miss errors during advice utilization from human and machine agents and effective connectivity study investigating miss errors during advice utilization from human and machine agents. *Social Neuroscience*, *12*(5), 570–581.

Goodyear, K., Parasuraman, R., Chernyak, S., Madhavan, P., Deshpande, G., & Krueger, F. (2016). Advice taking from humans and machines: An fMRI and effective connectivity study and effective connectivity study. *Frontiers in Human Neuroscience*, *10*(542).

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEErganizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114–125.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, *10*(3), 252–264.

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, *124*, 226–251.

Jung, C., Mueller, H., Pedemonte, S., Plances, S., & Thew, O. (2019). *Machine learning in UK financial services financial services*: *Bank of England and financial conduct authority report*.

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion comprehensive literature review on algorithm aversion. In *Proceedings of the 28th European conference on information systems*.

Kolanovic, M., & Krishnamachari, R. T. (2017). *Big data and AI strategies: machine learning and alternative data approach to investing strategies: machine learning and alternative data approach to investing*: *JP Morgan global quantitative & derivatives strategy report*.

Lewis, M. (2014). *Flash boys: a wall street revolt*. WW Norton & Company.

Liu, X. Y., Yang, H., Chen, Q., Zhang, R., Yang, L., Xiao, B., et al. (2020). FinRL: A deep reinforcement learning library for automated stock trading in quantitative finance: A deep reinforcement learning library for automated stock trading in quantitative finance. arXiv preprint arXiv:2011.09607.

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.

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*(4), 629–650.

Madhavan, P., & Wiegmann, D. A. (2007). Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Human Factors*, *49*(5), 773–785.

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

Meng, T. L., & Khushi, M. (2019). Reinforcement learning in financial markets. *Data*, *4*(3), 110.

OECD (2019). *Artificial intelligence in society*. Paris: OECD Publishing, http://dx.doi.org/10.1787/eedfee77-en.

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.

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

Promberger, M., & Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making*, *19*(5), 455–468.

Schniter, E., Shields, T. W., & Sznycer, D. (2020). Trust in humans and robots: Economically similar but emotionally different. *Journal of Economic Psychology*, *78*, Article 102253.

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, *83*(2), 260–281.

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403–414.