Discussion Paper No. 1194

ISSN (Print) 0473-453X
ISSN (Online) 2435-0982

BEWARE THE PERFORMANCE OF AN ALGORITHM BEFORE RELYING ON IT: EVIDENCE FROM A STOCK PRICE FORECASTING EXPERIMENT

Tiffany Tsz Kwan Tse
Nobuyuki Hanaki
Bolin Mao

Revised March 2024
October 2022

The Institute of Social and Economic Research
Osaka University
6-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan

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

Tiffany Tsz Kwan Tse[†], Nobuyuki Hanaki[‡] and Bolin Mao[§]

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.

Keywords: algorithms, financial market, forecasting, modification, technology adoption

JEL Classification: C90 , G1 , G4 , G17

[*] We gratefully acknowledge financial support from the Joint Usage/Research Center at ISER, 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/.

[†] Corresponding author: Institute of Social and Economic Research, Osaka University, Japan. Address: 6-1 Mihogaoka, Ibaraki, Osaka 567-0047 Japan E-mail: tiffany.econ@gmail.com

[‡] Institute of Social and Economic Research, Osaka University, Japan and University of Limassol, Cyprus. E-mail: nobuyuki.hanaki@iser.osaka-u.ac.jp

[§] Kyoto Institute of Economic Research, Kyoto University, Japan. E-mail: meta.bolin.mao@gmail.com

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 et al., 2020; Henrique et al., 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 et al., 2019), and forecasting future stock prices (Önkal et al., 2009). On the one hand, Dietvorst et al. (2015, 2018) coined the term "algorithm aversion" to describe people's tendency not to rely on an algorithm's output after learning that they are imperfect. On the other hand, Logg et al. (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 et al. (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 et al. (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 A1 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]

[1] Jussupow et al. (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.

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 (Logg et al., 2019; Jussupow et al., 2020). 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 et al., 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 et al., 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, 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

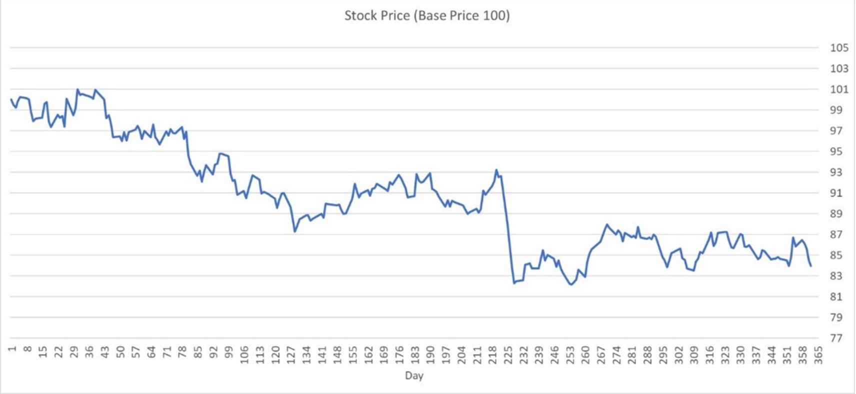
3.1 Main tasks

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 algortihms. 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 Figure 1 for an example).

Fig. 1 Sample of the graph



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, 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 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 2 for a summary of our six treatments.

Table 2 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

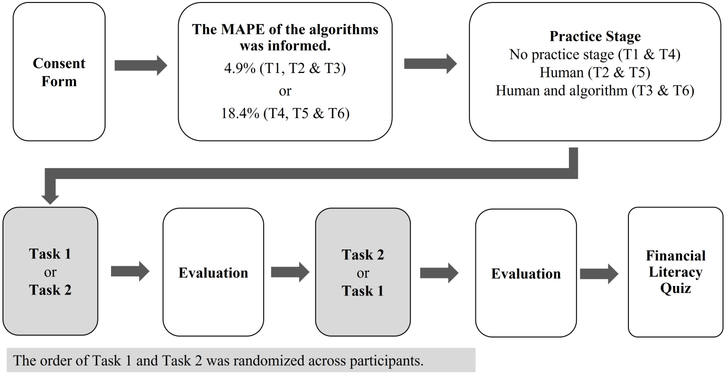
3.2.2 Within-subject design

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

Figure 2 demonstrates the flow of the experiment.

Fig. 2 Experiment flow


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.

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 second 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 minutes of experiments, and could earn up to an additional 1,200 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 minutes).[2] We also dropped one observation for task 2, Question 9, in which the participant entered a huge

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

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.

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.

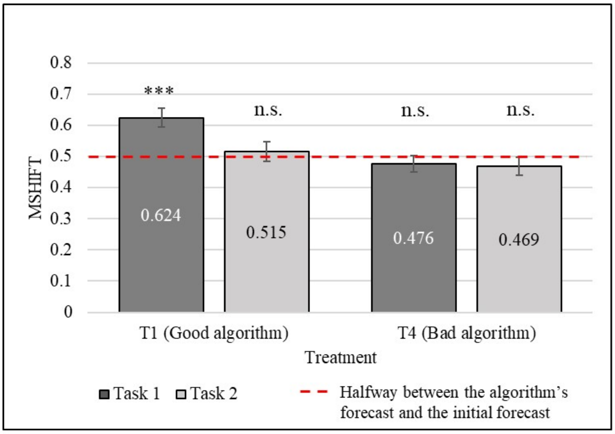
Our discussion is organized as follows. We first compared the degree of reliance on the algorithm when participants were only informed about the average performance level of the algorithm without experiencing the task (T1 vs. T4). We also compared reliance on the algorithm between task 1, when participants 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

Figure 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

Fig. 3 MSHIFT in T1 and T4

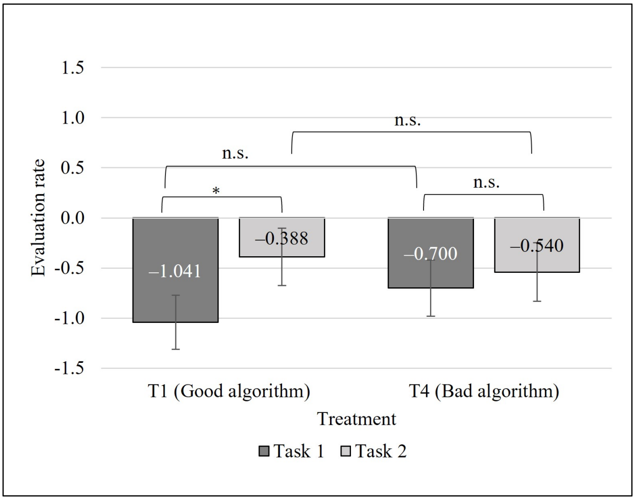


Notes: The p values were calculated based on a single-sample t-test. MSHIFTs were compared against the 0.5 level, which is halfway between the algorithm's forecast and the initial forecast. The symbols *, **, and *** indicate significance at the 0.05, 0.01, and 0.001 levels, respectively, and n.s. means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

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 Figure 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 Figure 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

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.

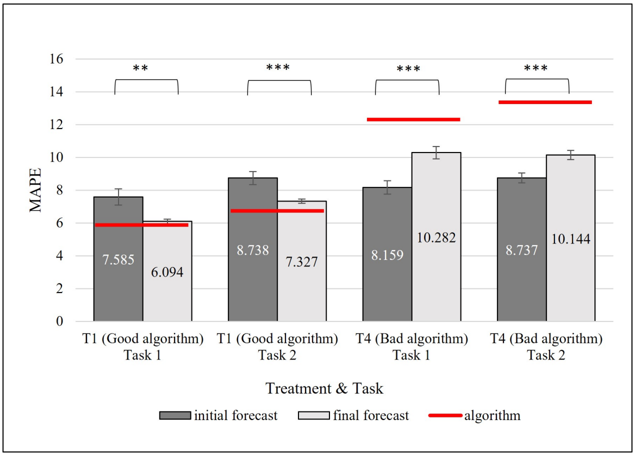
algorithms in T1 (good algorithm) and T4 (bad algorithm), participants' final forecasts became better than their initial forecasts in T1, but worse in T4, as shown in Figure 5.

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.

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,

Fig. 5 MAPE in T1 and T4

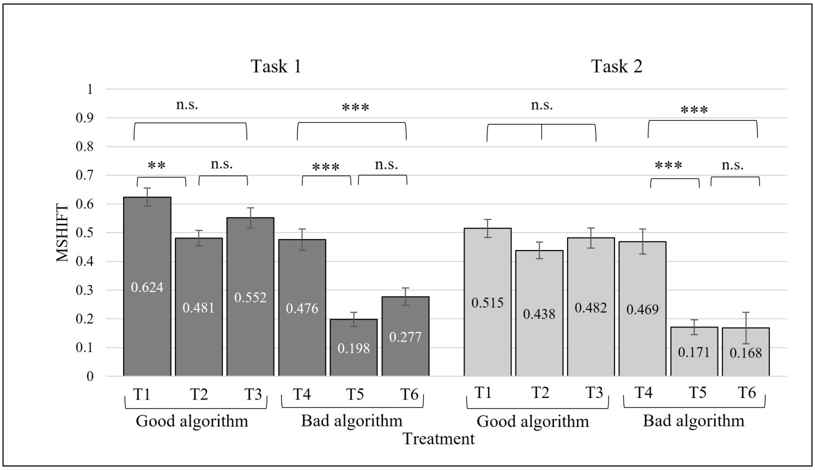


Notes: We regressed the MAPE on final forecast dummies by OLS regression model with robust cluster standard error on participant level. The figure shows the p values for the estimated coefficient on final forecast dummies. The symbols *, **, and *** indicate significance at the 0.05, 0.01, and 0.001 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.

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. Figure 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.

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.

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 ($F(1, 282) = 3.268$, $p = 0.215$).

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.

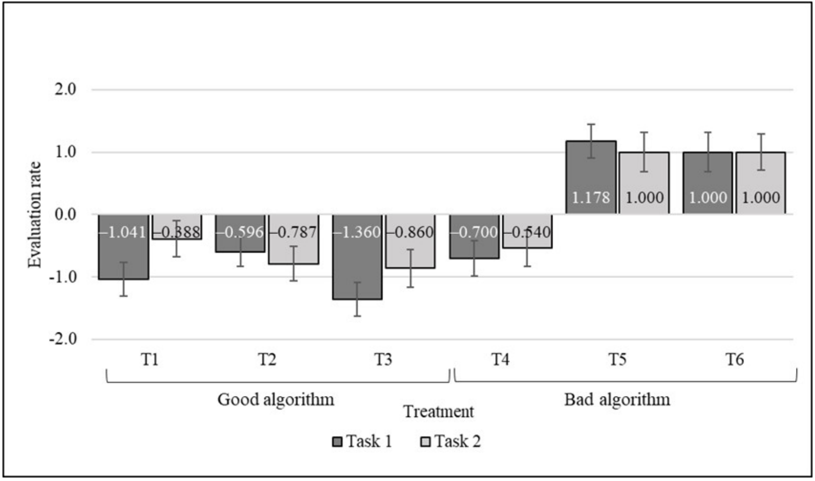
In T3 and T6, participants could directly compare the performance of their own forecasts with those of the algorithm. The average MAPEs (and the standard errors) during the practice stage were 8.064% (0.386%) for the participants and 5.889% for the algorithm in T3, and 7.861% (0.359%) for the participants 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.*

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



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 Figure 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).

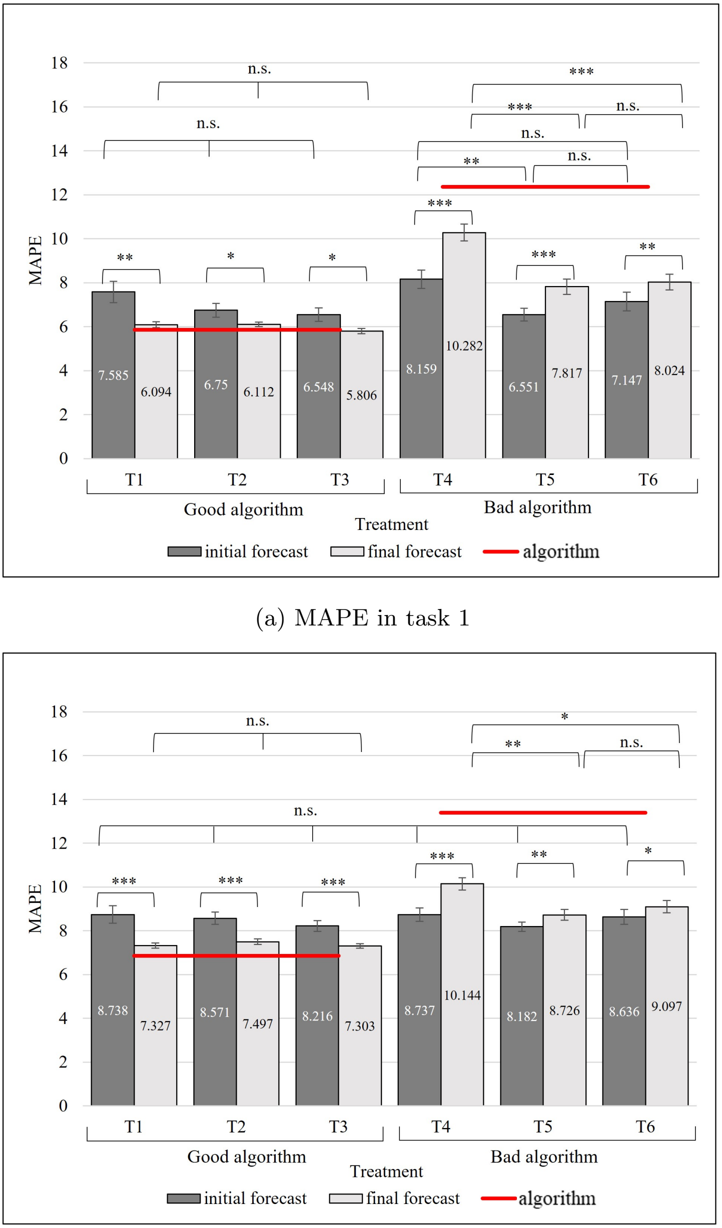
The significant increase in reliance on the algorithm in T6 compared with T5 in task 1 can be understood in terms of the effect of the discrepancy between the MAPE of the algorithm communicated to participants in the instructions (18.4%) and what they observed during the practice stage (10.14%). Recall 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.

Differences in MSHIFT across the treatments that we observed resulted in variations in performance of the final forecasts, measured by MAPE, as shown in Figure 8a for task 1 and Figure 8b 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 Figure 8a).

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

Fig. 8 MAPE in tasks 1 and 2



(a) MAPE in task 1

(b) MAPE in task 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.

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 Figure 8b. 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 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.

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 walks. *Journal of Economic Dynamics and Control*, *146*, 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, G., Krueger, F. (2017). An fMRI 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. *Frontiers in Human Neuroscience*, *10*, 542.

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

Greiner, B. (2015). Subject pool recruitment procedures: Organizing 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. *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. *In proceedings of the 28th european conference on information systems*

(ECIS).

Kolanovic, M., & Krishnamachari, R.T. (2017). Big data and AI 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., Wang, C.D. (2020). FinRL: 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*, 102426.

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

OECD (2019). Artificial intelligence in society. *OECD Publishing, Paris, https://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*, 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.

A Literature Review

Table A1: 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 & 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]

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.

Table A1 Continued: Existing studies on the algorithm performance information

Type	Research	Tasks [Source of the advice]	Results	Information about algorithm performance [The algorithms outperformed/under-performed humans]	Participants learnt about their performance [Measured]	Forecasting tasks [Initial forecasts were measured]
2	Bigman & 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

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.

Table A1 Continued: 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]
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]

Notes: Type A1 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.

Table A1 Continued: 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]
3	Gaudeul & Giannetti (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. (2016, 2017)	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 & 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]

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.

Table A1 Continued: 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]
4	Madhavan & 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.

Online Appendix to "Beware the performance of an algorithm before relying on it: Evidence from a stock price forecasting experiment"

Tiffany Tsz Kwan Tse[*], Nobuyuki Hanaki[†] and Bolin Mao[‡]

[*] Corresponding author: Institute of Social and Economic Research, Osaka University, Japan. Address: 6-1 Mihogaoka, Ibaraki, Osaka 567-0047 Japan E-mail: tiffany.econ@gmail.com

[†] Institute of Social and Economic Research, Osaka University, Japan and University of Limassol, Cyprus. E-mail: nobuyuki.hanaki@iser.osaka-u.ac.jp

[‡] Kyoto Institute of Economic Research, Kyoto University, Japan. E-mail: meta.bolin.mao@gmail.com

Contents

A Results for Figures

Table A1 Comparison between MSHIFT and the halfway point between the algorithm's forecast and the initial forecast using single-sample t-test

Treatment	Task	MSHIFT (Std. Err.)	Obs.	Halfway between algorithm's forecast and initial forecast	t-value (p-value)
T1	1	0.624 (0.031)	49	0.5	4.007 (<0.001)
T1	2	0.515 (0.031)	49	0.5	0.500 (0.619)
T4	1	0.476 (0.037)	50	0.5	–0.645 (0.522)
T4	2	0.469 (0.044)	50	0.5	–0.695 (0.491)

Notes: The number of observations is the number of participants in each treatment.

The ordinary least squares (OLS) linear regression model was used to test the impact of treatment effect on evaluation rate, MSHIFT, MAPE of initial forecast, and MAPE of final forecast in tasks 1 and 2. The dependent variables were evaluation rate in model (1) (2), MSHFT in model (3) (4), MAPE of initial forecast in model (5) (6), and MAPE of final forecast in model (7) (8). The independent variables were treatment dummies. In the estimation, we calculated the robust standard error under heteroskedasticity. Table A2 shows the predicted margin and standard errors estimated by the delta method. Firstly, we performed a joint test that the coefficients on treatment indicators were all equal (i.e., T1=T2=T3=T4=T5=T6) for evaluation rate, MSHIFT, MAPE of initial forecast, and MAPE of final forecast. If the above joint test rejected the null, we performed two joint tests that the coefficients on treatment indicators for the good algorithm were all equal (i.e., T1=T2=T3) and for the bad algorithm were all equal (i.e., T4=T5=T6). When the null was rejected, we performed tests of pair-wise comparison. We reported Bonferroni-adjusted p-values for these pair-wise comparisons.

The dataset was reshaped from wide to long. We generated a task 2 dummy that equaled 0 for task 1 and 1 for task 2. An OLS linear regression model was used to test the impact of task type on evaluation rate in each treatment. The dependent variable was evaluation rate. The independent variable was task

2 dummy. In the estimation, we calculated the robust standard error under heteroskedasticity with participant-level clustering. Results are shown in Table A3.

Table A2 Predicted evaluation rate, predicted MSHIFT, and predicted MAPE for treatment dummies using OLS regression with robust standard error

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Evaluation	Evaluation	MSHIFT	MSHIFT	MAPE	MAPE	MAPE	MAPE
Variables	Task 1	Task 2	Task 1	Task 2	initial	initial	final	final
					forecast	forecast	forecast	forecast
					Task 1	Task 2	Task 1	Task 2
Treatment 1	–1.041	–0.388	0.624	0.515	7.585	8.738	6.094	7.327
	(0.270)	(0.286)	(0.031)	(0.031)	(0.485)	(0.401)	(0.136)	(0.122)
Treatment 2	–0.596	–0.787	0.481	0.438	6.750	8.571	6.112	7.497
	(0.241)	(0.274)	(0.027)	(0.029)	(0.313)	(0.284)	(0.100)	(0.123)
Treatment 3	–1.360	–0.860	0.552	0.482	6.548	8.216	5.806	7.303
	(0.272)	(0.304)	(0.035)	(0.035)	(0.306)	(0.242)	(0.120)	(0.103)
Treatment 4	–0.700	–0.540	0.476	0.469	8.159	8.737	10.282	10.144
	(0.280)	(0.293)	(0.037)	(0.044)	(0.413)	(0.302)	(0.381)	(0.284)
Treatment 5	1.178	1.000	0.198	0.171	6.551	8.182	7.817	8.726
	(0.272)	(0.311)	(0.024)	(0.026)	(0.286)	(0.217)	(0.355)	(0.246)
Treatment 6	1.000	1.000	0.277	0.168	7.147	8.636	8.024	9.097
	(0.317)	(0.285)	(0.030)	(0.055)	(0.420)	(0.340)	(0.357)	(0.281)
	Prob > F	Prob > F	Prob > F	Prob > F	Prob > F	Prob > F	Prob > F	Prob > F
T1=T2=T3								
=T4=T5=T6	0.000	0.000	0.000	0.000	0.014	0.505	0.000	0.000
T1=T2=T3	0.105	0.461	0.003	0.194	0.192	0.446	0.117	0.449
T1 = T2 #	0.660	0.940	0.002	0.215	0.447	1.000	1.000	0.984
T1 = T3 #	1.000	0.775	0.370	1.000	0.216	0.797	0.340	1.000
T2 = T3 #	0.109	1.000	0.333	1.000	1.000	1.000	0.152	0.683
T4=T5=T6	0.000	0.000	0.000	0.000	0.006	0.260	0.000	0.001
T4 = T5 #	0.000	0.001	0.000	0.000	0.005	0.410	0.000	0.001
T4 = T6 #	0.000	0.001	0.000	0.000	0.260	1.000	0.000	0.028
T5 = T6 #	1.000	1.000	0.134	1.000	0.725	0.782	1.000	0.965
Observations	288	288	288	288	288	288	288	288

Notes: (a) Treatment 1 dummy equals 1 for treatment 1, and 0 otherwise. Treatment 2 dummy equals 1 for treatment 2, and 0 otherwise. Treatment 3 dummy equals 1 for treatment 3, and 0 otherwise. Treatment 4 dummy equals 1 for treatment 4, and 0 otherwise. Treatment 5 dummy equals 1 for treatment 5, and 0 otherwise. Treatment 6 dummy equals 1 for treatment 6, and 0 otherwise. (b) The unit of observation is the number of participants. The total number of observations is the number of participants in all treatments. (c) The robust standard errors are in parentheses. (d) # indicates Bonferroni-adjusted p-values.

Table A3 Comparison of evaluation rate between tasks 1 and 2 using OLS regression of evaluation rate on task dummy with robust cluster standard error on participant level

Variables	(1)	(2)	(3)	(4)	(5)	(6)
	T1	T2	T3	T4	T5	T6
Task 2 dummy	0.653*	–0.191	0.500	0.160	–0.178	0.000
	(0.284)	(0.188)	(0.251)	(0.289)	(0.252)	(0.287)
Constant	–1.041***	–0.596*	–1.360***	–0.700*	1.178***	1.000**
	(0.272)	(0.242)	(0.274)	(0.282)	(0.274)	(0.319)
Observations	98	94	100	100	90	94
R-squared	0.028	0.003	0.015	0.002	0.002	0.000
Clusters	49	47	50	50	45	47

Notes: (a) A task dummy equals 0 for task 1 and 1 for task 2. (b) The unit of observation is the number of participants. The total number of observations is the number of participants in each treatment × 2. (c) The robust standard errors clustered by participant levels are in parentheses. $* * *p < 0.001, * * p < 0.01, and * p < 0.05$.

The dataset was reshaped from wide to long. We generated a final forecast dummy that equaled 0 for initial forecast and 1 for final forecast. An OLS linear regression model was used to compare the MAPE of initial forecast and MAPE of final forecast in each treatment. The dependent variable was MAPE. The independent variable was final forecast dummy. In the estimation, we calculated the robust standard error under heteroskedasticity with participant-level clustering. Task 1 results are shown in Table A4, and task 2 results are shown in Table A5.

Table A4 Comparison of MAPE between initial forecast and final forecast in task 1 using OLS regression of MAPE on final forecast dummy with robust cluster standard error on participant level

Variables	(1) T1	(2) T2	(3) T3	(4) T4	(5) T5	(6) T6
Final forecast	–1.492**	–0.637*	–0.742*	2.123***	1.266***	0.877**
	(0.472)	(0.279)	(0.290)	(0.377)	(0.244)	(0.300)
Constant	7.585***	6.750***	6.548***	8.159***	6.551***	7.147***
	(0.488)	(0.315)	(0.308)	(0.415)	(0.288)	(0.422)
Observations	98	94	100	100	90	94
R-squared	0.084	0.039	0.049	0.127	0.081	0.027
Clusters	49	47	50	50	45	47

Notes: (a) The final forecast dummy equals 1 for final forecast and 0 for initial forecast. (b) The unit of observation is the number of participants. The total number of observations is the number of participants in each treatment × 2. (c) The robust standard errors clustered by participant levels are in parentheses. $\ast\ast\ast p < 0.001, \ast\ast p < 0.001, and \ast p < 0.05$.

Table A5 Comparison of MAPE between initial forecast and final forecast in task 2 using OLS regression of MAPE on final forecast dummy with robust cluster standard error on participant level

Variables	(1) T1	(2) T2	(3) T3	(4) T4	(5) T5	(6) T6
Final forecast	–1.411***	–1.074***	–0.913***	1.408***	0.545**	0.461*
	(0.330)	(0.219)	(0.217)	(0.289)	(0.151)	(0.205)
Constant	8.738***	8.571***	8.216***	8.737***	8.182***	8.636***
	(0.403)	(0.286)	(0.243)	(0.303)	(0.219)	(0.342)
Observations	98	94	100	100	90	94
R-squared	0.106	0.115	0.109	0.105	0.030	0.012
Clusters	49	47	50	50	45	47

Notes: (a) The final forecast dummy equals 1 for final forecast and 0 for initial forecast. (b) The unit of observation is the number of participants. The total number of observations is the number of participants in each treatment × 2. (c) The robust standard errors clustered by participant levels are in parentheses. $\ast\ast\ast p < 0.001, \ast\ast p < 0.01, and \ast p < 0.05$.

B Analyses of experimental results conditional on personal characteristics

We used the survey datasets of participants' personal characteristics (Hanaki et al, 2021) measured before the experiment. Personal characteristics include being female, being undergraduate student, financial literacy score, risk aversion score, and cognitive reflection test (CRT) score.

Risk aversion scores were measured using the method used by Masuda and Lee (2019). The elicitation task was originally proposed by Noussair et al (2011). Participants are asked to choose between a risky lottery in which they have a 50% chance of getting JPY650 and a 50% chance of getting JPY50, and a sure payment of JPY X (where X may be 200, 250, 300, 350, or 400). If the two options are indifferent to the respondent, then the X is a certainty equivalent. The larger the risk premium, the more risk averse they are. Usually, we assume that individuals will consistently choose the risky option only when X is less than their certainty equivalent, so the fewer times they choose the risky option, the more risk averse they are.

The CRT is applied following Finucane and Gullion (2010). The three questions were as follows.

1. If it takes 2 nurses 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? (in minutes). [Correct answer: 2 minutes; intuitive answer: 200 minutes]
2. Soup and salad cost 5.50 euros in total. The soup costs 5 euros more than the salad. How much does the salad cost? (in euros). [Correct answer: 0.25 euro; intuitive answer: 0.5 euro]

3. Sally is making some 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 the final concentration? (in hours). [Correct answer: 5 hours; intuitive answer: 3 hours]

The financial literacy scores were measured by following Fernandes et al. (2014). The 12 questions were as follows.

1. Imagine that the interest rate on your savings account was 1% per year and inflation was 2% per year. After 1 year, would you be able to buy:

   1 More than today with the money in this account.

   2 Exactly the same as today with the money in this account.

   3 Less than today with the money in this account.

   4 Don't know.

   5 Refuse to answer.

   [Correct answer: 3]

2. Do you think that the following statement is true or false? "Bonds are normally riskier than stocks."

   1 True.

   2 False.

   3 Don't know.

   4 Refuse to answer.

   [Correct answer: 2]

3. Considering a long time period (for example 10 or 20 years), which asset described below normally gives the highest return?

   1 Savings accounts.

   2 Stocks.

   3 Bonds.

   4 Don't know.

5 Refuse to answer.

[Correct answer:2]

4. Normally, which asset described below displays the highest fluctuations over time?

 1 Savings accounts.

 2 Stocks.

 3 Bonds.

 4 Don't know.

 5 Refuse to answer.

 [Correct answer:2]

5. When an investor spreads his money among different assets, does the risk of losing a lot of money:

 1 Increase.

 2 Decrease.

 3 Stay the same.

 4 Don't know.

 5 Refuse to answer.

 [Correct answer:2]

6. Do you think that the following statement is true or false? "If you were to invest $1000 in a stock mutual fund, it would be possible to have less than $1000 when you withdraw your money?"

 1 Increase.

 2 Decrease.

 3 Stay the same.

 4 Don't know.

 5 Refuse to answer.

 [Correct answer: 1]

7. Do you think that the following statement is true or false? “A stock mutual fund combines the money of many investors to buy a variety of stocks.”

   1 True.

   2 False.

   3 Don’t know.

   4 Refuse to answer.

   [Correct answer: 1]

8. Do you think that the following statement is true or false? “A 15-year mortgage typically requires higher monthly payments than a 30-year mortgage, but the total interest paid over the life of the loan will be less.”

   1 True.

   2 False.

   3 Don’t know.

   4 Refuse to answer.

   [Correct answer: 1]

9. Suppose you had $100 in a savings account and the interest rate is 20% per year and you never withdraw money or interest payments. After 5 years, how much would you have on this account in total?

   1 More than $200.

   2 Exactly $200.

   3 Less than $200.

   4 Don’t know.

   5 Refuse to answer.

   [Correct answer: 1]

10. Which of the following statements is correct?

    1 Once one invests in a mutual fund, one cannot withdraw the money in the first year.

2 Mutual funds can invest in several assets, for example invest in both stocks and bonds.

3 Mutual funds pay a guaranteed rate of return which depends on their past performance.

4 None of the above

5 Don't know.

6 Refuse to answer.

[Correct answer: 2]

11. Which of the following statements is correct? If somebody buys a bond of firm B:

1 He owns a part of firm B.

2 He has lent money to firm B.

3 He is liable for firm B's debts.

4 None of the above

5 Don't know.

6 Refuse to answer.

[Correct answer: 2]

12. Suppose you owe $3,000 on your credit card. You pay a minimum payment of $30 each month. At an Annual Percentage Rate of 12% (or 1% per month), how many years would it take to eliminate your credit card debt if you made no additional new charges?

1 Less than 5 years.

2 Between 5 and 10 years.

3 Between 10 and 15 years.

4 None of the above

5 Don't know.

6 Refuse to answer.

[Correct answer: 4]

Table B1 summarizes participants' personal characteristics. We conducted a one-way ANOVA test to compare personal characteristics among all treatments. There were no statistically significant differences in personal characteristics among treatments, except in the financial literacy score.

Table B1 Summary of participants' personal characteristics

	Treatments						One-way ANOVA	
	T1	T2	T3	T4	T5	T6	F	Prob > F
Female	0.347	0.383	0.380	0.300	0.311	0.319	0.269	0.930
	(0.069)	(0.072)	(0.069)	(0.065)	(0.070)	(0.069)		
Undergraduate student	0.776	0.894	0.860	0.700	0.778	0.766	1.469	0.200
	(0.060)	(0.045)	(0.050)	(0.065)	(0.063)	(0.062)		
Financial literacy score	8.694	7.787	8.180	8.140	8.311	7.170	2.372	0.040
	(0.302)	(0.349)	(0.372)	(0.345)	(0.308)	(0.327)		
Risk aversion score	2.898	3.106	3.380	3.080	3.200	3.340	0.657	0.657
	(0.211)	(0.213)	(0.202)	(0.237)	(0.257)	(0.216)		
CRT score	2.633	2.681	2.540	2.660	2.444	2.681	0.893	0.486
	(0.095)	(0.092)	(0.104)	(0.093)	(0.117)	(0.092)		
Obs.	49	47	50	50	45	47		

Notes: (a) The female dummy equals 1 for female, and 0 otherwise. The undergraduate student dummy equals 1 for undergraduate student, and 0 otherwise. Financial literacy score range = 0–12 (higher score indicates greater financial literacy). Risk aversion score range = 0–5 (higher score indicates a higher level of risk aversion). CRT score range = 0–3 (higher score indicates greater cognitive ability). (b) The unit of observation is the number of participants. The total number of observations is the number of participants in each treatment. (c) The standard errors are in parentheses.

An OLS linear regression model was used to test the impact of treatment effect on evaluation rate, MSHIFT, MAPE of initial forecast, and MAPE of final forecast in tasks 1 and 2, conditional on personal characteristics as described in Table B1. The dependent variables were evaluation rate in model (1) (2), MSHFT in model (3) (4), MAPE of initial forecast in model (5) (6), and MAPE of final forecast in model (7) (8). The independent variables were treatment dummies. The control variables were female, undergraduate student, financial literacy score, risk aversion score, and CRT score. In the estimation, we calculated the robust standard error under heteroskedasticity. In Table B2, we report the predicted margin and standard errors estimated by the delta method. Firstly, we performed a joint test that the coefficients on treatment indicators were all equal (i.e., T1=T2=T3=T4=T5=T6) for evaluation rate, MSHIFT, MAPE of initial forecast, and MAPE of final forecast. If the above joint test rejected the null, we performed two joint tests that the coefficients on treatment indicators for the good algorithm were all equal (i.e., T1=T2=T3) and for the bad algorithm were all equal (i.e., T4=T5=T6). When the null was rejected, we performed tests of pair-wise comparison. We reported Bonferroni-adjusted p-values for these pair-wise comparisons.

Table B2 Predicted evaluation rate, predicted MSHIFT, and predicted MAPE for treatment dummies using OLS regression conditional on personal characteristics with robust standard error

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Evaluation	Evaluation	MSHIFT	MSHIFT	MAPE	MAPE	MAPE	MAPE
Variables	Task 1	Task 2	Task 1	Task 2	initial	initial	final	final
					forecast	forecast	forecast	forecast
					Task 1	Task 2	Task 1	Task 2
Treatment 1	−1.016	−0.346	0.626	0.514	7.612	8.710	6.159	7.363
	(0.287)	(0.296)	(0.032)	(0.031)	(0.489)	(0.393)	(0.146)	(0.136)
Treatment 2	−0.582	−0.806	0.487	0.443	6.796	8.559	6.205	7.528
	(0.246)	(0.279)	(0.029)	(0.031)	(0.319)	(0.288)	(0.128)	(0.142)
Treatment 3	−1.325	−0.848	0.550	0.476	6.532	8.170	5.815	7.270
	(0.268)	(0.304)	(0.037)	(0.036)	(0.316)	(0.247)	(0.127)	(0.114)
Treatment 4	−0.723	−0.544	0.475	0.472	8.162	8.779	10.241	10.153
	(0.280)	(0.298)	(0.035)	(0.042)	(0.398)	(0.303)	(0.354)	(0.264)
Treatment 5	1.172	1.037	0.193	0.163	6.486	8.151	7.762	8.698
	(0.276)	(0.320)	(0.024)	(0.026)	(0.295)	(0.230)	(0.351)	(0.240)
Treatment 6	0.953	0.931	0.277	0.176	7.149	8.711	7.950	9.082
	(0.321)	(0.297)	(0.031)	(0.052)	(0.452)	(0.348)	(0.350)	(0.274)
	Prob > F	Prob > F	Prob > F	Prob > F	Prob > F	Prob > F	Prob > F	Prob > F
T1=T2=T3								
=T4=T5=T6	0.000	0.000	0.000	0.000	0.010	0.415	0.000	0.000
T1=T2=T3	0.120	0.408	0.006	0.269	0.191	0.409	0.061	0.362
T1 = T2 #	0.738	0.745	0.004	0.317	0.439	1.000	1.000	1.000
T1 = T3 #	1.000	0.726	0.361	1.000	0.217	0.744	0.222	1.000
T2 = T3 #	0.128	1.000	0.531	1.000	1.000	0.918	0.086	0.464
T4=T5=T6	0.000	0.000	0.000	0.000	0.004	0.193	0.000	0.000
T4 = T5 #	0.000	0.001	0.000	0.000	0.003	0.313	0.000	0.000
T4 = T6 #	0.000	0.001	0.000	0.000	0.272	1.000	0.000	0.015
T5 = T6 #	1.000	1.000	0.100	1.000	0.657	0.524	1.000	0.865
Observations	288	288	288	288	288	288	288	288

Notes: (a) Treatment 1 dummy equals 1 for treatment 1, and 0 otherwise. Treatment 2 dummy equals 1 for treatment 2, and 0 otherwise. Treatment 3 dummy equals 1 for treatment 3, and 0 otherwise. Treatment 4 dummy equals 1 for treatment 4, and 0 otherwise. Treatment 5 dummy equals 1 for treatment 5, and 0 otherwise. Treatment 6 dummy equals 1 for treatment 6, and 0 otherwise. (b) The unit of observation is the number of participants. The total number of observations is the number of participants in all treatments. (c) The robust standard errors are in parentheses. (d) # indicates Bonferroni-adjusted p-values.

The dataset was reshaped from wide to long. We generated a task 2 dummy that equaled 0 for task 1 and 1 for task 2. An OLS linear regression model was used to test the impact of task type on evaluation rate in each treatment, conditional on personal characteristics. The dependent variable was evaluation rate. The independent variable was task 2 dummy. The control variables were the personal characteristics described in Table B1. In the estimation, we calculated the robust standard error under heteroskedasticity with participant-level clustering. The results are shown in Table B3.

Table B3 Comparison of evaluation rate between tasks 1 and 2 using OLS regression of evaluation rate on task dummy conditional on personal characteristics with robust cluster standard error on participant level

Variables	(1)	(2)	(3)	(4)	(5)	(6)
	T1	T2	T3	T4	T5	T6
Task dummy	0.653*	–0.191	0.500	0.160	–0.178	0.000
	(0.291)	(0.193)	(0.257)	(0.297)	(0.260)	(0.295)
Female	0.268	–0.254	–0.368	–0.243	–0.334	–0.528
	(0.467)	(0.700)	(0.766)	(0.661)	(0.563)	(0.889)
Undergraduate student	–1.008	–0.901	0.734	0.184	0.825	0.192
	(0.602)	(0.494)	(0.639)	(0.578)	(0.576)	(0.638)
Financial literacy score	–0.079	–0.015	–0.219**	–0.039	0.204	–0.080
	(0.094)	(0.129)	(0.077)	(0.103)	(0.125)	(0.123)
Risk aversion score	0.220	0.124	–0.078	0.034	–0.269*	0.052
	(0.180)	(0.202)	(0.231)	(0.173)	(0.131)	(0.206)
CRT score	–0.598*	–0.322	0.484	0.649	0.452	0.012
	(0.296)	(0.359)	(0.454)	(0.366)	(0.361)	(0.584)
Constant	1.273	0.903	-1.030	-2.269	-1.295	1.388
	(1.298)	(1.892)	(1.709)	(1.374)	(1.658)	(2.145)
Observations	98	94	100	100	90	94
R-squared	0.162	0.054	0.140	0.051	0.146	0.019
Clusters	49	47	50	50	45	47

Notes: (a) The task dummy equals 0 for task 1 and 1 for task 2. (b) The unit of observation is the number of participants. The total number of observations is the number of participants in each treatment × 2. (c) The robust standard errors clustered by participant levels are in parentheses. $* * *p < 0.001, * * p < 0.01, and * p < 0.05$

The dataset was reshaped from wide to long. We generated a final forecast dummy that equaled 0 for initial forecast and 1 for final forecast. An OLS linear regression model was used to compare the MAPE of initial forecast and MAPE of final forecast in each treatment, conditional on personal characteristics. The dependent variable was MAPE. The independent variable was final forecast dummy. The control variables were personal characteristics. In the estimation, we calculated the robust standard error under heteroskedasticity with participant-level clustering. The results for task 1 are shown in Table B4, and the results of task 2 are shown in Table B5.

Table B4 Comparison of MAPE between initial forecast and final forecast in task 1 using OLS regression of MAPE on final forecast dummy conditional on personal characteristics with robust cluster standard error on participant level

Variables	(1) T1	(2) T2	(3) T3	(4) T4	(5) T5	(6) T6
Final forecast	–1.492**	–0.637*	–0.742*	2.123***	1.266***	0.877**
	(0.485)	(0.287)	(0.298)	(0.387)	(0.251)	(0.308)
Female	–0.803	–0.033	0.097	–0.543	–1.186	2.174
	(0.623)	(0.453)	(0.353)	(0.843)	(0.623)	(1.360)
Undergraduate student	0.135	0.754	–0.619	–1.174	–1.529	0.738
	(0.560)	(0.479)	(0.391)	(1.072)	(0.759)	(0.848)
Financial literacy score	–0.108	–0.012	0.048	–0.018	–0.276	0.176
	(0.111)	(0.085)	(0.057)	(0.110)	(0.165)	(0.134)
Risk aversion score	0.017	–0.043	0.018	–0.097	0.135	0.382
	(0.156)	(0.119)	(0.156)	(0.194)	(0.165)	(0.246)
CRT score	–0.914	–0.370	–0.043	–1.109**	–0.115	0.592
	(0.666)	(0.237)	(0.251)	(0.365)	(0.376)	(0.804)
Constant	11.060***	7.310***	6.696***	12.536***	10.249***	1.761
	(1.744)	(1.148)	(1.094)	(1.773)	(2.143)	(3.492)
Observations	98	94	100	100	90	94
R-squared	0.152	0.079	0.074	0.196	0.218	0.203
Clusters	49	47	50	50	45	47

Notes: (a) The final forecast dummy equals 1 for final forecast and 0 for initial forecast. (b) The unit of observation is the number of participants. The total number of observations is the number of participants in each treatment × 2. (c) The robust standard errors clustered by participant levels are in parentheses. $* * *p < 0.001, * * p < 0.01, and * p < 0.05$

Table B5 Comparison of MAPE between initial forecast and final forecast in task 2 using OLS regression of MAPE on final forecast dummy conditional on personal characteristics with robust cluster standard error on participant level

Variables	(1) T1	(2) T2	(3) T3	(4) T4	(5) T5	(6) T6
Final forecast	–1.411***	–1.074***	–0.913***	1.408***	0.545**	0.461*
	(0.338)	(0.225)	(0.222)	(0.297)	(0.156)	(0.211)
Female	0.934	0.525	–0.071	0.100	–0.056	0.459
	(0.541)	(0.443)	(0.309)	(0.641)	(0.468)	(1.207)
Undergraduate student	–0.017	–0.150	0.321	–0.865	0.117	0.319
	(0.517)	(0.681)	(0.249)	(0.586)	(0.533)	(0.623)
Financial literacy score	–0.118	0.090	0.057	0.059	0.007	0.046
	(0.162)	(0.073)	(0.054)	(0.088)	(0.105)	(0.133)
Risk aversion score	–0.325	–0.253*	0.110	0.132	0.278*	0.212
	(0.275)	(0.119)	(0.113)	(0.182)	(0.130)	(0.186)
CRT score	–0.544	0.060	0.007	–0.842*	–0.106	–0.419
	(0.606)	(0.214)	(0.200)	(0.386)	(0.325)	(0.729)
Constant	11.824***	8.432***	7.113***	10.665***	7.422***	8.329*
	(2.933)	(1.051)	(0.896)	(1.272)	(1.647)	(3.294)
Observations	98	94	100	100	90	94
R-squared	0.225	0.188	0.138	0.215	0.125	0.089
Clusters	49	47	50	50	45	47

Notes: (a) The final forecast dummy equals 1 for final forecast and 0 for initial forecast. (b) The unit of observation is the number of participants. The total number of observations is the number of participants in each × 2. (c) The robust standard errors clustered by participant levels are in parentheses. $* * *p < 0.001, * * p < 0.01, and * p < 0.05$

Predicted evaluation rate, predicted MSHIFT, and predicted MAPE conditional on personal characteristics are shown in Figures B1-B6

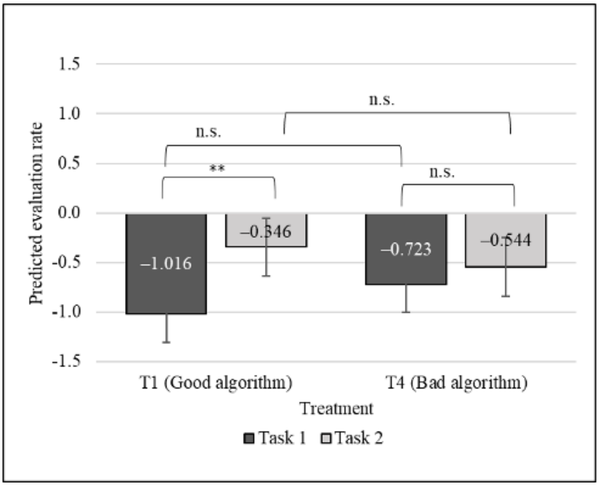


Fig. B1 Predicted evaluation rate in T1 and T4

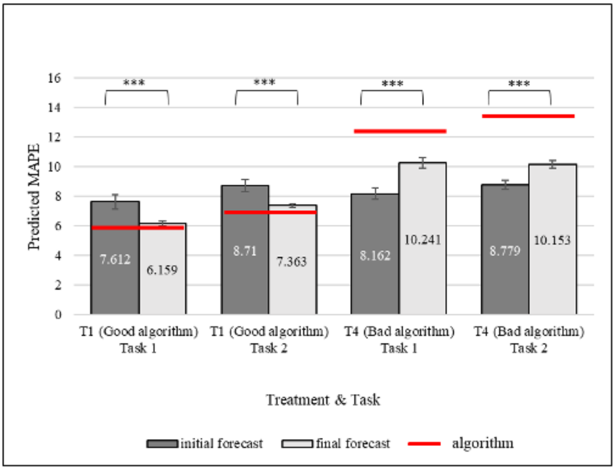


Fig. B2 Predicted MAPE in T1 and T4

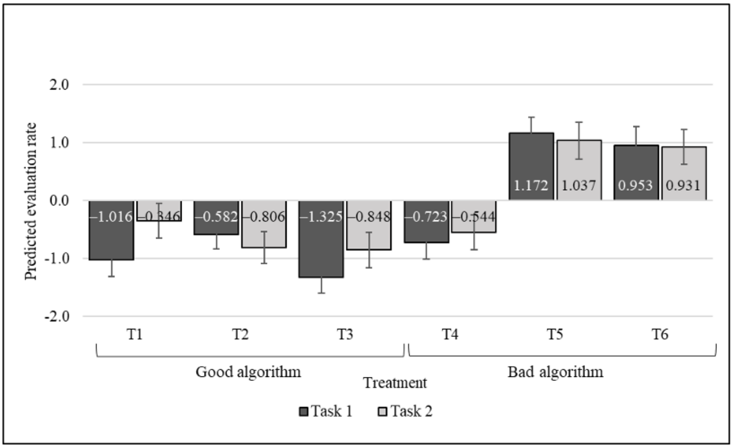


Fig. B3 Predicted evaluation rate in all treatments

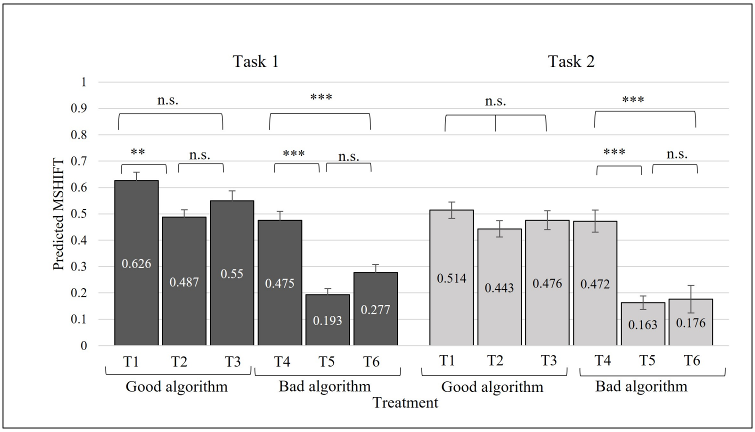


Fig. B4 Predicted MSHIFT in all treatments

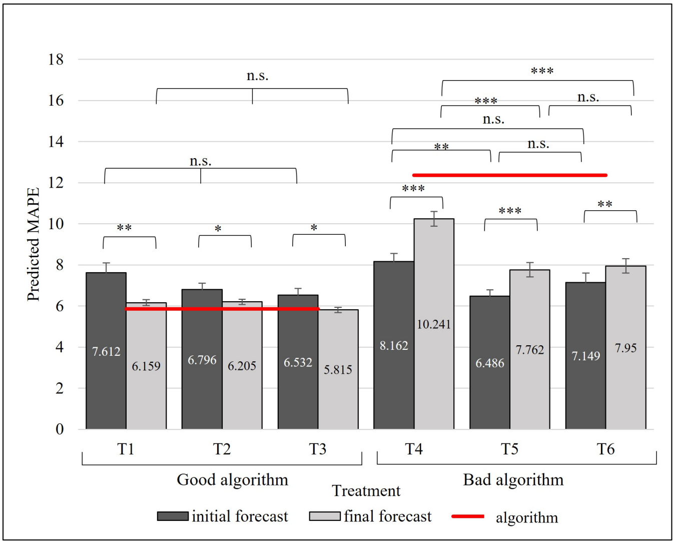


Fig. B5 Predicted MAPE in task 1 in all treatments

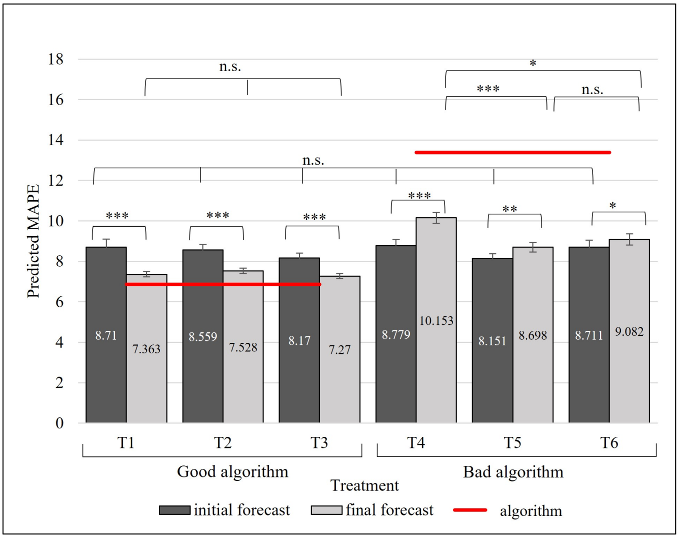


Fig. B6 Predicted MAPE in task 2 in all treatments

C Robustness in experimental design

First, we confirmed that both good and bad algorithms gave unbiased forecasts. We compared the mean percentage error (MPE) (i.e., MAPE without taking the absolute) of the good algorithm and the bad algorithm using a paired t-test. The results are shown in Table C1. We found that the MPE of the good algorithm and the bad algorithm was near zero. MPE did not differ significantly between the good algorithm and the bad algorithm.

Table C1 Comparison of mean percentage error (MPE) between good algorithm and bad algorithm using paired t-test

Task	Good algorithm		Bad algorithm		Diff (Good–Bad)	t-value
	MPE (Std. Err.)	Obs.	MPE (Std. Err.)	Obs.	MPE (Std. Err.)	(p-value)
Practice	–0.026 (0.020)	10	–0.031 (0.039)	10	0.005 (0.040)	0.126 (0.903)
Task 1	–0.058 (0.014)	10	–0.095 (0.038)	10	0.036 (0.029)	1.245 (0.245)
Task 2	0.029 (0.034)	10	–0.018 (0.055)	10	0.047 (0.041)	1.160 (0.276)
All stages	–0.018 (0.015)	30	–0.048 (0.026)	30	0.030 (0.021)	1.415 (0.168)

Notes: The number of observations is the number of questions in each task.

Second, we confirmed that the good algorithm performed better than the participants, and the bad algorithm performed worse than the participants, on average. We compared the MAPE between the algorithm's forecast and initial forecast using a paired t-test. The initial human forecast was the forecast submitted by participants before observing the algorithm's forecast in the practice stage, task 1, and task 2. The results are shown in Table C2. We found that the good algorithm always performed better than the participants, and the bad algorithm always performed worse than the participants.

Table C2 Comparison of MAPE between algorithm forecast and initial forecast using paired t-test

Treatment	Task	Algorithm MAPE	Initial forecast MAPE (Std. Err.)	Diff (Algorithm–Initial) MAPE (Std. Err.)	t-value (p-value)	Obs.
1	Task 1	5.866	7.585 (0.485)	–1.719 (0.485)	–3.544 (<0.001)	49
1	Task 2	6.862	8.738 (0.401)	–1.876 (0.401)	–4.677 (<0.001)	49
2	Practice	5.889	8.300 (0.578)	–2.411 (0.578)	–4.172 (<0.001)	47
2	Task 1	5.866	6.750 (0.314)	–0.884 (0.314)	–2.818 (0.007)	47
2	Task 2	6.862	8.571 (0.284)	–1.709 (0.284)	–6.008 (<0.001)	47
3	Practice	5.889	8.064 (0.386)	–2.175 (0.386)	–5.639 (<0.001)	50
3	Task 1	5.866	6.548 (0.306)	–0.682 (0.306)	–2.227 (0.031)	50
3	Task 2	6.862	8.216 (0.242)	–1.354 (0.242)	–5.591 (<0.001)	50
4	Task 1	12.359	8.159 (0.412)	4.2 (0.412)	10.183 (<0.001)	50
4	Task 2	13.391	8.737 (0.302)	4.654 (0.302)	15.422 (<0.001)	50
5	Practice	10.144	8.100 (0.335)	2.044 (0.335)	6.092 (<0.001)	45
5	Task 1	12.359	6.551 (0.286)	5.808 (0.286)	20.312 (<0.001)	45
5	Task 2	13.391	8.182 (0.218)	5.209 (0.218)	23.948 (<0.001)	45
6	Practice	10.144	7.861 (0.359)	2.283 (0.359)	6.358 (<0.001)	47
6	Task 1	12.359	7.147 (0.420)	5.212 (0.420)	12.409 (<0.001)	47
6	Task 2	13.391	8.636 (0.340)	4.755 (0.340)	13.999 (<0.001)	47

Notes: The number of observations is the number of participants in each treatment.

Third, there was no learning effect within tasks because participants did not receive feedback after providing their forecast in each time series. The order of the 10 graphs was random in tasks 1 and 2 in each treatment. We compared the MAPE of final forecasts between the first five forecasts and the last five forecasts using a paired t-test. The results are shown in Table C3. There was no significant difference between the performance in the first five forecasts and the last five forecasts.

Table C3 Comparison of MAPE between first five human final forecast and last five human final forecasts using paired t-test

Treatment	Task	First five forecasts MAPE (Std. Err.)	Last five forecasts MAPE (Std. Err.)	Diff (First–Last) MAPE (Std. Err.)	t-value (p-value)	Obs.
1	Task 1	6.167 (0.238)	6.020 (0.288)	0.147 (0.454)	0.324 (0.748)	49
1	Task 2	7.100 (0.371)	7.554 (0.349)	–0.454 (0.677)	–0.670 (0.506)	49
2	Task 1	6.295 (0.268)	5.929 (0.235)	0.366 (0.462)	0.792 (0.432)	47
2	Task 2	7.084 (0.387)	7.913 (0.424)	–0.830 (0.773)	–1.074 (0.289)	47
3	Task 1	5.918 (0.198)	5.694 (0.221)	0.224 (0.344)	0.651 (0.518)	50
3	Task 2	7.643 (0.360)	6.963 (0.354)	0.680 (0.683)	0.995 (0.325)	50
4	Task 1	10.414 (0.603)	10.150 (0.447)	0.263 (0.740)	0.356 (0.723)	50
4	Task 2	9.989 (0.503)	10.300 (0.495)	–0.310 (0.822)	–0.378 (0.707)	50
5	Task 1	7.523 (0.447)	8.111 (0.430)	–0.587 (0.516)	–1.139 (0.261)	45
5	Task 2	8.691 (0.397)	8.762 (0.480)	–0.071 (0.730)	–0.097 (0.923)	45
6	Task 1	8.129 (0.472)	7.919 (0.408)	0.211 (0.518)	0.407 (0.686)	47
6	Task 2	9.004 (0.398)	9.190 (0.437)	–0.187 (0.618)	–0.302 (0.764)	47

Notes: The number of observations is the number of participants in each treatment.

D The relationship between MSHIFT and MAPE of human forecast in the practice stage

Table D1 OLS linear regression of MAPE of human forecast in practice stage on mean shift rate in tasks 1 and 2 with the good and bad algorithms, with robust standard errors

Variables	(1) MSHIFT Task1 Good algorithm Treatment 2	(2) MSHIFT Task2 Good algorithm Treatment 2	(3) MSHIFT Task1 Bad algorithm Treatment 5	(4) MSHIFT Task2 Bad algorithm Treatment 5
MAPE of human forecast in practice stage	0.011*	0.019*	–0.003	–0.003
	(0.005)	(0.007)	(0.009)	(0.017)
Constant	0.389***	0.285***	0.226**	0.192
	(0.054)	(0.067)	(0.078)	(0.123)
Observations	47	47	45	45
R-squared	0.055	0.134	0.002	0.001

Notes: (a) The unit of observation is the number of participants. The total number of observations is the number of participants in T2 in model (1) (2) and T5 in model (3) (4). (b) The robust standard errors are in parentheses. $***p < 0.001, **p < 0.01, and*p < 0.05$.

E Comparison of MAPE between algorithm forecast and final forecast

Table E1 Comparison of MAPE between algorithm forecast and final forecast using paired t-test

Treatment	Task	Algorithm MAPE	Final forecast MAPE (Std. Err.)	Diff (Algorithm–Final) MAPE (Std. Err.)	t-value (p-value)	Obs.
1	Task 1	5.866	6.094 (0.136)	–0.228 (0.136)	–1.677 (0.100)	49
1	Task 2	6.862	7.327 (0.122)	–0.465 (0.122)	–3.810 (<0.001)	49
2	Task 1	5.866	6.112 (0.100)	–0.246 (0.100)	–2.462 (0.018)	47
2	Task 2	6.862	7.497 (0.124)	–0.635 (0.124)	–5.144 (<0.001)	47
3	Task 1	5.866	5.806 (0.120)	0.060 (0.120)	0.500 (0.619)	50
3	Task 2	6.862	7.303 (0.103)	–0.441 (0.103)	–4.288 (<0.001)	50
4	Task 1	12.359	10.282 (0.380)	2.077 (0.380)	5.461 (<0.001)	50
4	Task 2	13.391	10.144 (0.284)	3.247 (0.284)	11.451 (<0.001)	50
5	Task 1	12.359	7.817 (0.355)	4.542 (0.355)	12.786 (<0.001)	45
5	Task 2	13.391	8.726 (0.246)	4.665 (0.246)	18.950 (<0.001)	45
6	Task 1	12.359	8.024 (0.357)	4.335 (0.357)	12.144 (<0.001)	47
6	Task 2	13.391	9.097 (0.281)	4.294 (0.281)	15.258 (<0.001)	47

Notes: The number of observations is the number of participants in each treatment.

F Summary of the graphs in the practice stage, task 1, and task 2

Table F1 Performance of good algorithm and bad algorithm in each question in practice stage, task 1, and task 2

Stage	Question	Realized price (Base price)	Good algorithm		Bad algorithm	
			Forecast (Base price)	APE	Forecast (Base price)	APE
Practice	1	118.89	129.74	9.12	132.16	11.16
	2	109.99	100.53	8.6	120.06	9.16
	3	118.01	118.91	0.76	114.27	3.17
	4	134.43	124.78	7.18	117.27	12.76
	5	205.82	186.12	9.57	153.29	25.52
	6	102.51	100.44	2.02	103.39	0.86
	7	143.54	152.76	6.42	131.02	8.72
	8	145.7	142.67	2.08	122.11	16.19
	9	102.32	94.36	7.78	105.42	3.03
	10	111.69	105.71	5.35	123.82	10.86
MAPE in the practice stage				**5.89**		**10.14**
Task 1	1	138.7	129.12	6.91	126.8	8.58
	2	83.97	84.1	0.16	89.63	6.75
	3	99.51	93.28	6.26	87.88	11.69
	4	143.43	141.45	1.38	129.38	9.8
	5	87.04	77.97	10.42	62.07	28.69
	6	348.83	297.4	14.74	276.04	20.87
	7	111.04	107.17	3.48	109.07	1.77
	8	145.82	133.74	8.28	137.5	5.71
	9	183.45	173.6	5.37	143.14	21.97
	10	71.65	70.46	1.66	77.22	7.77
MAPE in task 1				**5.87**		**12.36**
Task 2	1	123.31	123.59	0.22	112.51	8.76
	2	134.63	134.72	0.07	130.77	2.87
	3	140.31	131.88	6.01	138.94	0.98
	4	208.06	180.28	13.35	154.74	25.63
	5	114.65	114.37	0.25	122.42	6.78
	6	134.23	134.12	0.08	100.35	25.24
	7	78.47	84.25	7.37	88.9	13.3
	8	66.62	76.32	14.57	86.6	29.99
	9	110.88	139.65	25.94	119.53	7.8
	10	165.35	166.6	0.75	144.56	12.58
MAPE in task 2				**6.86**		**13.39**

Table F1 Summary of the graphs in the practice stage, task 1, and task 2

Stage	Question	Company		First business day			Last business day			Next 30 business days		
		Name	Ticker	Date	Closing price	Base price	Date	Closing price	Base price	Date	Closing price	Base price
Practice	1	Mettler Toledo	MTD	2017/03/01	483.65	100	2018/02/28	616.22	127.41	2018/03/29	575.03	118.89
	2	Micron Technology	MU	2009/10/01	7.51	100	2010/09/30	7.21	96.01	2010/10/29	8.26	109.99
	3	Cerner	CERN	2011/10/03	32.78	100	2012/09/28	38.70	118.04	2012/10/26	38.69	118.01
	4	Teleflex	TFX	2013/02/01	75.87	100	2014/01/31	93.64	123.42	2014/02/28	101.99	134.43
	5	Domino's Pizza	DPZ	2009/03/02	6.70	100	2010/02/26	12.49	186.42	2010/03/26	13.79	205.82
	6	Lilly (Eli) & Co.	LLY	2010/03/01	34.32	100	2011/02/28	34.56	100.70	2011/03/30	35.18	102.51
	7	Newmont Corporation	NEM	2009/10/01	42.40	100	2010/09/30	62.81	148.14	2010/10/29	60.86	143.54
	8	ONEOK	OKE	2010/03/01	19.81	100	2011/02/28	28.27	142.70	2011/03/30	28.86	145.70
	9	International Flavors & Fragrances	IFF	2011/01/03	55.65	100	2011/12/30	52.42	94.20	2012/01/27	56.94	102.32
	10	Motorola Solutions Inc.	MSI	2009/06/01	25.59	100	2010/05/28	27.69	108.21	2010/06/25	28.58	111.69
Task 1	1	Keysight Technologies	KEYS	2017/11/01	44.57	100	2018/10/31	57.08	128.07	2018/11/30	61.82	138.70
	2	Equifax Inc.	EFX	2017/04/03	136.10	100	2018/03/29	117.81	86.56	2018/04/27	114.28	83.97
	3	Eastman Chemical	EMN	2011/07/01	51.99	100	2012/06/29	50.37	96.87	2012/07/27	51.74	99.51
	4	Ross Stores	ROST	2008/11/03	7.72	100	2009/10/30	11.00	142.52	2009/11/27	11.07	143.43
	5	Ventas Inc	VTR	2008/08/01	52.00	100	2009/07/31	40.31	77.51	2009/08/28	45.27	87.04
	6	Las Vegas Sands	LVS	2009/07/01	7.70	100	2010/06/30	22.14	287.53	2010/07/30	26.86	348.83
	7	Goldman Sachs Group	GS	2013/02/01	149.90	100	2014/01/31	164.12	109.49	2014/02/28	166.45	111.04
	8	Under Armour (Class C)	UA	2018/04/02	13.99	100	2019/03/29	18.87	134.88	2019/04/26	20.40	145.82
	9	Activision Blizzard	ATVI	2014/11/03	20.30	100	2015/10/30	34.76	171.23	2015/11/27	37.24	183.45
	10	Franklin Resources	BEN	2015/05/01	52.14	100	2016/04/29	37.34	71.61	2016/05/27	37.36	71.65
Task 2	1	Genuine Parts	GPC	2010/11/01	47.44	100	2011/10/31	57.43	121.06	2011/11/30	58.50	123.31
	2	Host Hotels & Resorts	HST	2009/10/01	10.84	100	2010/09/30	14.48	133.62	2011/10/28	14.59	134.63
	3	L3Harris Technologies	LHX	2017/07/03	109.67	100	2018/06/29	144.54	131.80	2018/07/27	153.88	140.31
	4	E*Trade	ETFC	2013/02/01	10.80	100	2014/01/31	20.02	185.37	2014/02/28	22.47	208.06
	5	Tapestry, Inc.	TPR	2015/11/02	31.74	100	2016/10/31	35.89	113.07	2016/11/30	36.39	114.65
	6	FedEx Corporation	FDX	2008/12/01	63.45	100	2009/11/30	84.45	133.10	2009/12/30	85.17	134.23
	7	Entergy Corp.	ETR	2010/08/02	79.56	100	2011/07/29	66.80	83.96	2011/08/26	62.43	78.47
	8	Whirlpool Corp.	WHR	2017/07/03	191.97	100	2018/06/29	146.23	76.17	2018/07/27	127.89	66.62
	9	Autodesk Inc.	ADSK	2017/10/02	112.47	100	2018/09/28	156.11	138.80	2018/10/26	124.71	110.88
	10	AutoZone Inc	AZO	2018/05/01	632.16	100	2019/04/30	1028.31	162.67	2019/05/30	1045.29	165.35

Notes: The S&P 500 company list was captured on June 30, 2020.

G Data preparation

G.1 Raw data

We collected the raw stock data from Yahoo!Finance. It consisted of the daily prices (open, high, low, closing, and adjusted closing) and trading volume of 83 NYSE or NASDAQ listed companies which were ranking on the top by the market capitalization in their sectors (i.e., basic materials, consumer goods, healthcare, services, utilities, conglomerates, financial, industrial goods, and technology). We show the stock tickers in Table G2. The stock time series ranged from either 2000/1/1 or the IPO date (if the IPO date is after 2000/1/1), to 2020/1/1.

Table G2 Tickers of the 83 collected stocks

Sector	Stock ticker
Basic materials	XOM, RDS-B, PTR, CVX, TOT, BP, BHP, SNP, SLB, BBL
Consumer goods	AAPL, PG, BUD, KO, PM, TM, PEP, UN, UL, MO
Healthcare	JNJ, PFE, NVS, UNH, MRK, AMGN, MDT, SNY
Services	AMZN, BABA, WMT, CMCSA, HD, DIS, MCD, CHTR, UPS
Utilities	NEE, DUK, D, SO, NGG, AEP, PCG, EXC, SRE, PPL
Conglomerates	IEP, CODI, REX, SPLP, PICO, AGFS, GMRE
Financial	BCH, BSAC, BRK-A, JPM, WFC, BAC, V, C, HSBC, MA
Industrial goods	GE, MMM, BA, HON, LMT, CAT, GD, DHR, ABB
Technology	GOOG, MSFT, FB, T, CHL, ORCL, TSM, VZ, INTC, CSCO

G.2 Technical indicators

To extract more information from the raw data, we calculated the technical indicators based on the raw data by the *ta-lib* package.[1] All the technical indicators used for building stock prediction algorithm are listed in Table G3.

Because some technical indicator calculation will shorten the time series length, therefore the technical indicator time series had different lengths. Consequently, by truncating the overlength, we aligned with all technical indicators such that they had the same length as the shortest one.

[1] https://mrjbq7.github.io/ta-lib/

Table G3 Technical indicator used for our algorithm building

Functions	Technical indicators
Overlap study	Bollinger bands Exponential moving average Double exponential moving average Kaufman adaptive moving average Moving average Midpoint over period Midpoint price over period Parabolic SAR Simple moving average Triangular moving average Weighted moving average
Momentum indicator	Absolute price oscillator Aroon Aroon oscillator balance of power Commodity channel index Moving average convergence/divergence Moving average convergence/divergence with controllable MA type Momentum Percentage price oscillator Rate of change Rate of change ratio Stochastic Stochastic fast Ultimate oscillator Williams' % R
Volume indicator	Chaikin A/D line Chaikin A/D oscillator
Price transform	Average price Median price Typical price Weighted close price
Volatility indicator	True range

For stock i on day t, we further concatenated the technical indicators with the raw data (after alignment) and noted the concatenated data piece as a sample unit $X_{i,t}$, shown in Figure G1. Consequently, the sample unit had 6 features coming from the raw data and 43 features coming from the technical indicators.

G.3 Sampling

We first divided the time span in our dataset into the training period (from 2000/1/1 or IPO to 2019/6/30) and the test period (from 2019/10/1 to 2020/1/1). Data samples whose prediction target lied in the training period

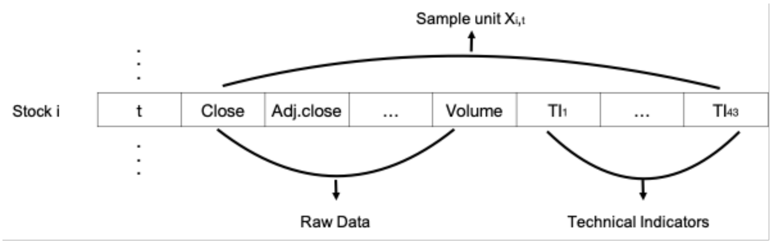


Fig. G1 Sample unit

formed the training dataset, and whose prediction target lied in the test period formed the test dataset.

For each stock and starting from the beginning of the training period, we sampled the sample unit sequence as the model input and we took the sequence's corresponding closing price for prediction. The first and the second time samplings in the training period are illustrated in Figures G2 and G3. Here, L is the sequence length of input; P is the length of prediction gap, and J is the jump size between two consecutive data samples.

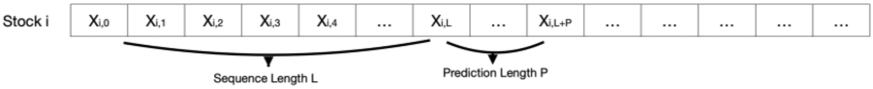


Fig. G2 The first time sampling in training period for stock i

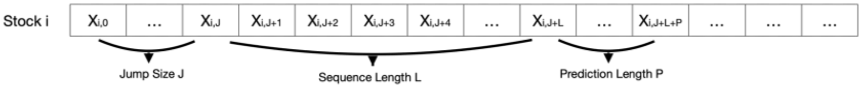


Fig. G3 The second time sampling in training period for stock i

In a similar way, the first time and the second time samplings in the test period are shown in Figures G4 and G5.

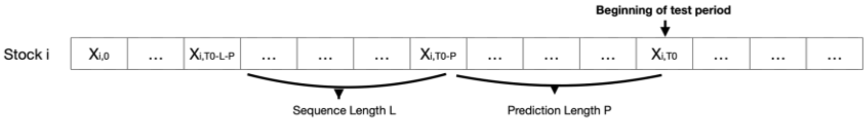


Fig. G4 The first time sampling in test period for stock i

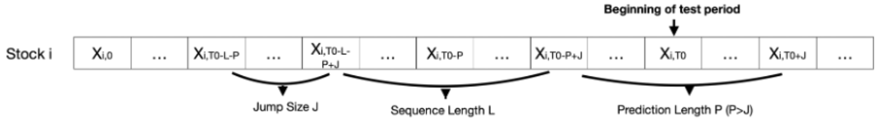


Fig. G5 The second time sampling in test period for stock i

Specifically in our sampling, $L = 253$, which is averagely the number of trading days within a year, and $P = 21$, which is averagely the number of trading days within one month.

G.4 Linear scaling

For each data sample in the training dataset and the test dataset, each feature was scaled by

$$x^*_{i,p,t} = \frac{x_{i,p,t} - x_{i,p,min}}{x_{i,p,max} - x_{i,p,min}}$$

, where $x_{i,p,max}$ and $x_{i,p,min}$ are, respectively, the maximum and minimum value of stock i's feature p in the entire training period.

G.5 Shuffling and batching

We shuffled all the scaled samples in the training dataset and batched every 32 of the samples in a batch. As a result of shuffling, training samples were fed into the model randomly.

H Model structure and training setting

We used a five–layer fully–connected (FC) neural network as our model. Since the data input had the shape of (*batch size, sequence length, feature size*), we first flattened each data input into the shape of (*batch size, sequence length* × *feature size*) as our input layer. Then we set the dimension of the hidden state in each FC layer as 6198, 3099, 1549, 744, and 1, with the last FC layer's output as the prediction. We set the dropout rate as 0.3 and we applied the *sigmoid* function as the activation function for each FC layer.

We chose mean absolute error (MAE) as the loss function and we used *Adam* as the optimizer. The initial learning rate for *Adam* was 0.0001. Training epoch was set as 15 for the good performance model; 2 for the bad performance model.

I Experiment instructions

Instructions (English Translation)

[Screen 1]

Thank you for your participation in this experiment.

This experiment takes around 45 minutes.

You will receive 500 yen participation fee and the rewards depending on your performance in the experiment.

Please go to the next page to start the experiment.

[Screen 2]

GENERAL EXPERIMENTAL INSTRUCTION

In this experiment, you are asked to play a role as **financial advisors** who **forecast the future stock price** based on historical price information.

Your company has created a **robot** that is designed to forecast future stock prices.

This robot makes the future stock price forecast by learning the historical stock price information, from January 1st, 2000 / Initial Public Offering (IPO) day to January 1st, 2020, of 83 target companies rank top in their capital market sectors (i.e. Basic Materials, Consumer Goods, Healthcare, Services, Utilities, Conglomerates, Financial, Industrial Goods, Technology).

The performance of the robot is measured by the percentage error of its forecasts. The percentage error is calculated as follows.

$$\left|\frac{forecast - realized\ price}{realized\ price}\right| \times 100$$

The smaller the percentage error, the higher the accuracy. 0% indicates the forecast exactly the same as the realized price.

The mean percentage error of the robot is around 4.9%. *(shown in Treatment 1, 2 and 3)*
The mean percentage error of the robot is around 18.4%. *(shown in Treatment 4, 5, and 6)*

The mean percentage error is calculated as follows (i.e. n=5311, which is the number of predictions used to measure the performance of the robot).

$$\left(\frac{1}{n}\sum\left|\frac{Forecast - realized\ price}{realized\ price}\right|\right) \times 100$$

You are asked to decide whether you use your own forecast or the robot's forecast to predict the future stock price.

There are a **practice stage** and **2 tasks**. *(shown in Treatment 2, 3, 5 and 6)*
There are **2 tasks**. *(shown in Treatment 1 and 4)*

Firstly, you enter the practice stage to learn the performance of your forecast. *(shown in Treatment 2 and 5)*
Firstly, you enter the practice stage to learn the performance of your forecast and the robot's forecast. *(shown in Treatment 3 and 6)*

Then, you enter Task 1 and Task 2. *(shown in Treatment 2, 3, 5 and 6)*

The information about Task 1 and Task 2 will be displayed later.

Please go to the next page to enter the Task 1 and Task 2. *(shown in Treatment 1 and 4)*

Please go to the next page to enter the practice stage. *(shown in Treatment 2, 3, 5 and 6)*

[Screen 3] *(shown in Treatment 2, 3, 5 and 6)*

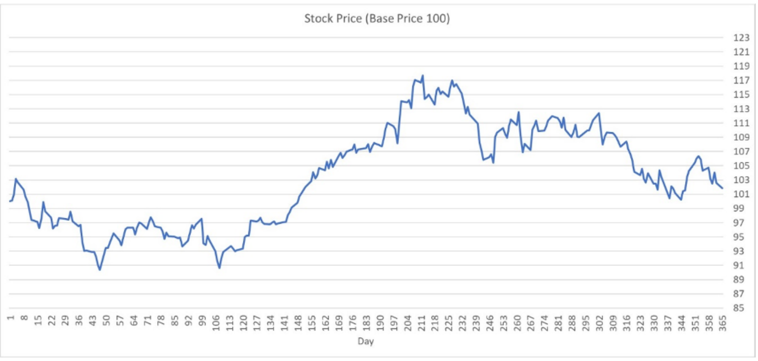
Practice Stage

The following 10 graphs are the 12 months of end-of-day prices of randomly selected stocks from the S&P 500 starting from a randomly selected day between January 1st 2008 and December 1st 2018. You will not be told about the name of the stock or the starting date which was randomly selected. Please note that end-of-day prices have been rescaled so that all starting prices will be equal to 100.

For each graph, please forecast what will be the end-of-day price for this stock 30 days after the last price shown on the graph.

After you finish entering your forecast for 10 graphs, we will show you the performance of your forecast and the robot's forecast.

The following shows the example of the graph.



The X-axis indicates the days of one year (from day 1 to day 365).

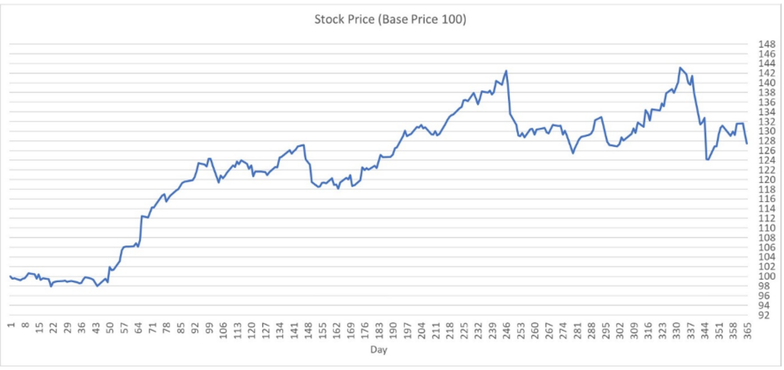
The Y-axis indicates the rescaled stock price starting from 100.

The graph shows the stock price of working days, skipping weekends and holidays.

[Screen 4] 10 questions in practice stage *(shown in Treatment 2, 3, 5 and 6)*

Practice Stage

Q1. What will be the end-of-day price for this stock 30 days after the last price shown on the graph? (The last price is 127.41.)



Please enter your forecast.

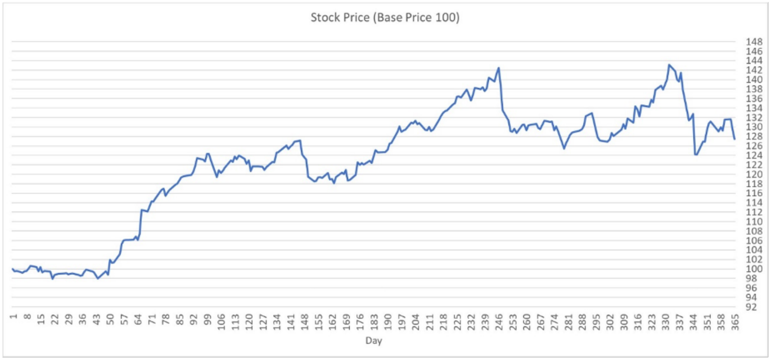
[Screen 5] 10 questions in practice stage *(shown in Treatment 2, 3, 5 and 6)*

Results of Practice Stage

The percentage error of your original forecasts is calculated as follows.

$$\left|\frac{your\ forecast - realized\ price}{realized\ price}\right| \times 100$$

PracticeQ1



The realized price:118.89

Your forecast:100

The robot's forecast: 129.74 *(shown in Treatment 2 and 3)*

The robot's forecast: 132.16 *(shown in Treatment 5 and 6)*

The percentage error of your forecast:15.89%

The percentage error of the robot's forecast: 9.12% *(shown in Treatment 3)*

The percentage error of the robot's forecast:11.16% *(shown in Treatment 6)*

[Screen 6]

Results of Practice Stage

The mean percentage error in the Practice Stage is calculated as follows. (i.e. n=10, which is the number of predictions in the practice stage)

$$\left(\frac{1}{10}\sum\left|\frac{Forecast - realized\ price}{realized\ price}\right|\right) \times 100$$

Mean percentage error of your forecast:19.41%

Mean percentage error of the robot's forecast: 5.89% *(shown in Treatment 3)*

Mean percentage error of the robot's forecast: 10.14% *(shown in Treatment 6)*

In Task 1 and 2, you will perform similar stock price forecasting task.

You will earn points according to the accuracy of your forecast (measured by percentage error).

Your final reward will be based on your performance in one prediction of the chosen task.

Please go to the next page to enter the Task 1 and Task 2.

After you go to the next page, you cannot go back to this page.

[Screen 7]

TASK 1

You will be shown **10** graphs showing **12 months of end-of-day prices** of randomly selected stocks from the S&P 500 starting from a randomly selected day between January 1st 2008 and December 1st 2018. You will not be told about the name of the stock or the starting date which was randomly selected.

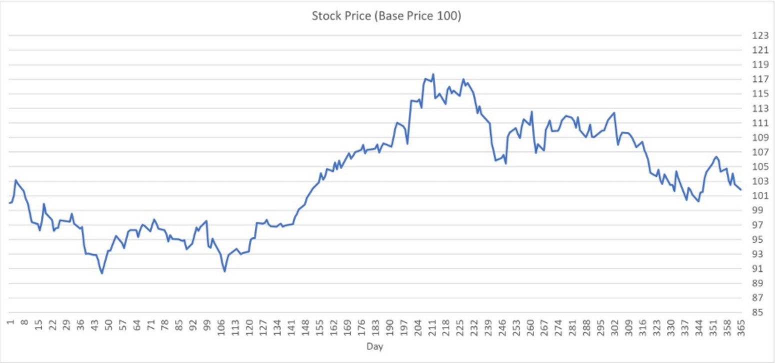
Please note that end-of-day prices have been rescaled so that all starting prices will be equal to 100.

For each graph, you will be asked to forecast what will be the end-of-day price for this stock **30 days after the last price shown on the graph**.

After you finish submitting your forecast, you will receive the forecast by the robot.

Then you can choose between using your own forecast or the robot's forecast as your final forecast to submit.

The following shows the example of the graph.



The X-axis indicates the days of one year (from day 1 to day 365).

The Y-axis indicates the rescaled stock price starting from 100.

The graph shows the stock price of working days, skipping weekends and holidays.

You will be rewarded based on the accuracy of your final forecasts as follows.

$$\text{Max}\,[\,200 - 10 \times |\frac{your\ final\ forecast\ - realized\ price}{realized\ price} \times 100|\,,0\,]$$

If your final forecast is exactly at the price observed, then you will receive 200 points. For each percentage point difference between your final forecast and the observed price, 10 points will be subtracted. If your final forecast differs from the observed price by more than 20%, you will receive 0 points.

If Task 1 is chosen for your final payment, one of the 10 series will be randomly chosen. You will be rewarded based on the point you earned in the chosen series. Your reward will be calculated with 1 point = 6 yen.

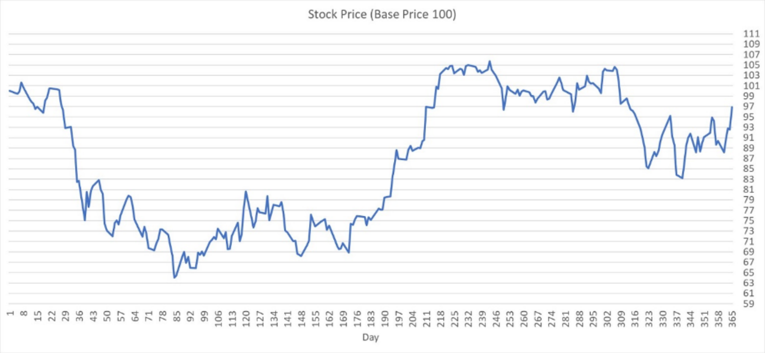
You will not be informed about the accuracy of your forecast until the experiment ends.

Evaluation

After you finish submitting your final forecast, you are asked to evaluate the accuracy of your forecast relative to the robot's forecast.

[Screen 8] 10 questions in Task 1

Task1Q3. What will be the end-of-day price for this stock 30 days after the last price shown on the graph? (The last price is 96.87.)



Please enter your forecast.

[Screen 9]

After you go to the next page, you cannot go back to this page.

[Screen 10]

TASK 1

You now receive the forecast by the robot.

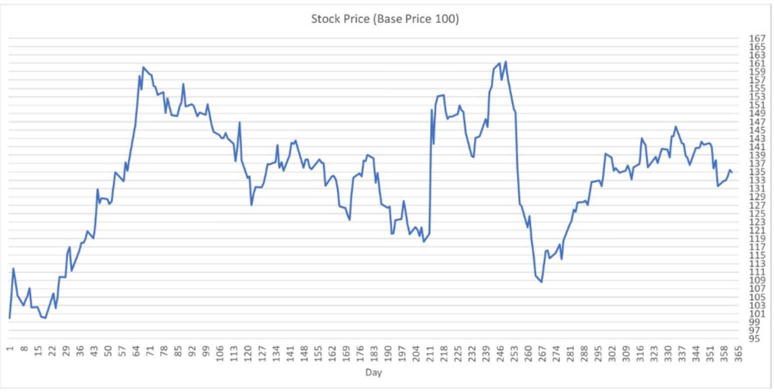
The mean percentage error of the robot is around 4.9%. *(shown in Treatment 1, 2 and 3)*

The mean percentage error of the robot is around 18.4%. *(shown in Treatment 4, 5 and 6)*

Please choose between using your own forecast or the robot's forecast as your final forecast to submit.

[Screen 11] 10 questions in Task 1

Task1Q8. We show your forecast and the robot's forecast for the end-of-day price for this stock 30 days after the last price shown on the graph.



The mean percentage error of the robot is around 4.9%. *(shown in Treatment 1, 2 and 3)*

The mean percentage error of the robot is around 18.4%. *(shown in Treatment 4, 5 and 6)*

Your forecast: 100

The robot's forecast: 133.74 *(shown in Treatment 1, 2 and 3)*

The robot's forecast: 137.50 *(shown in Treatment 4, 5 and 6)*

Please choose between using your own or the robot's forecast as your final forecast to submit.

● 100

○ 133.74 *(shown in Treatment 1, 2 and 3)*

○ 137.50 *(shown in Treatment 4, 5 and 6)*

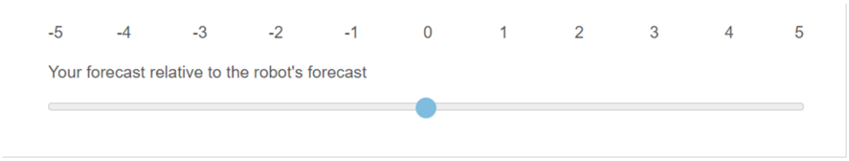
[Screen 12]

After you go to the next page, you cannot go back to this page.

[Screen 13]

Please evaluate the accuracy of your forecast relative to the robot's forecast in this task.

i.e. **from -5 (the lowest, your forecast is <u>less accurate</u> than the robot's forecast to a great extent.) to 5 (the highest, your forecast is <u>more accurate</u> than the robot's forecast to a great extent.)** 0 indicates that your forecast has the same accuracy as the robot's forecast's.



[Screen 14]

<u>TASK 2</u>

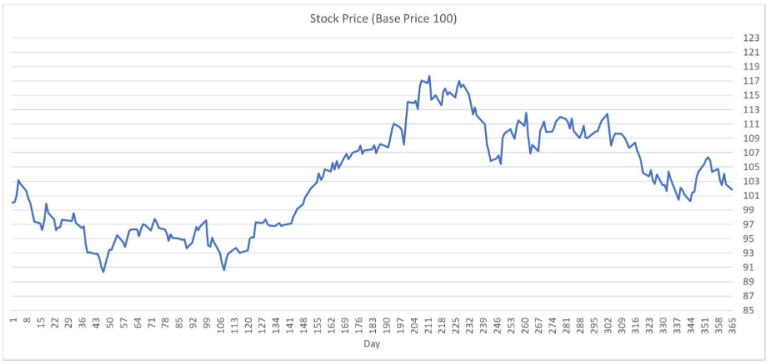
You will be shown **10** graphs showing **12 months of end-of-day prices** of randomly selected stocks from the S&P 500 starting from a randomly selected day between January 1st 2008 and Dec 1st 2018. You will not be told about the name of the stock or the starting date which was randomly selected. **Please note that end-of-day prices have been rescaled so that all starting prices will be equal to 100**.

For each graph, you will be asked to forecast what will be the end-of-day price for this stock **30 days after the last price shown on the graph**.

After you finish submitting your forecast, you will receive the forecast by the robot.

By observing your original forecast and the robot's forecast, you can modify and submit your final forecast.

The following shows the example of the graph.



The X-axis indicates the days of one year (from day 1 to day 365).

The Y-axis indicates the rescaled stock price starting from 100.

The graph shows the stock price of working days, skipping weekends and holidays.

You will be rewarded based on the accuracy of your final forecasts as follows.

$$\text{Max}\,[\,200 - 10 \times |\frac{your\ final\ forecast\ - realized\ price}{realized\ price} \times 100|\,,0\,]$$

If your final forecast is exactly at the price observed, then you will receive 200 points. For each percentage point difference between your final forecast and the observed price, 10 points will be subtracted. If your final forecast differs from the observed price by more than 20%, you will receive 0 points.

If Task 2 is chosen for your final payment, one of the 10 series will be randomly chosen. You will be rewarded based on the point you earned in the chosen series. Your reward will be calculated with 1 point = 6 yen.

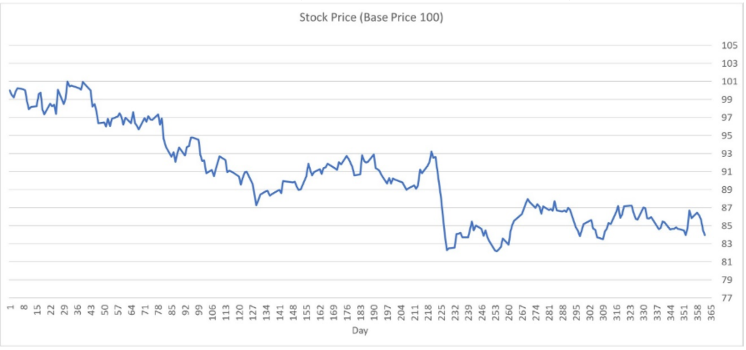
You will not be informed about the accuracy of your forecast until the experiment ends.

<u>Evaluation</u>

After you finish submitting your final forecast, you are asked to evaluate the accuracy of your original forecast relative to the robot's forecast, and also the accuracy of your final forecast relative to the robot's forecast.

[Screen 15] 10 questions in Task 2

Task2Q7. What will be the end-of-day price for this stock 30 days after the last price shown on the graph? (The last price is 83.96.)



Please enter your forecast

[Screen 16]

After you go to the next page, you cannot go back to this page.

[Screen 17]

<u>TASK 2</u>

You now receive the forecast by the robot.

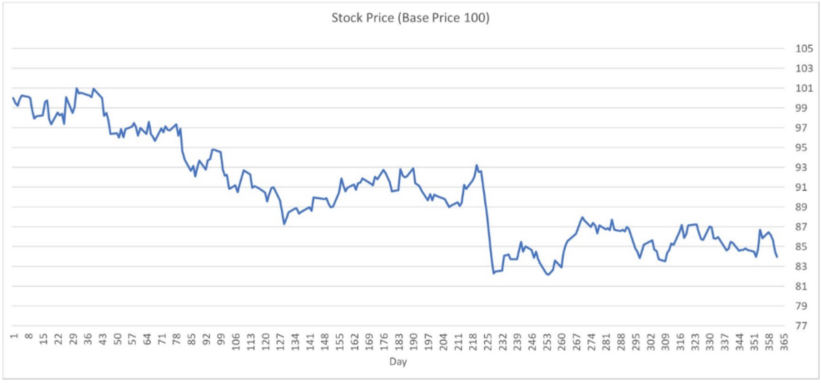
The mean percentage error of the robot is around 4.9%. *(shown in Treatment 1, 2 and 3)*

The mean percentage error of the robot is around 18.4%. *(shown in Treatment 4, 5 and 6)*

<u>**By observing your original forecast and the robot's forecast, you can modify and then submit your final forecast.**</u>

[Screen 18] 10 questions in Task 2

Task2Q7. We show your forecast and the robot's forecast for the end-of-day price for this stock 30 days after the last price shown on the graph.



The mean percentage error of the robot is around 4.9%. *(shown in Treatment 1, 2 and 3)*

The mean percentage error of the robot is around 18.4%. *(shown in Treatment 4, 5 and 6)*

Your forecast: 100

The robot's forecast: 84.25 *(shown in Treatment 1, 2 and 3)*

The robot's forecast: 88.90 *(shown in Treatment 4, 5 and 6)*

Please enter your final forecast.

[Screen 19]

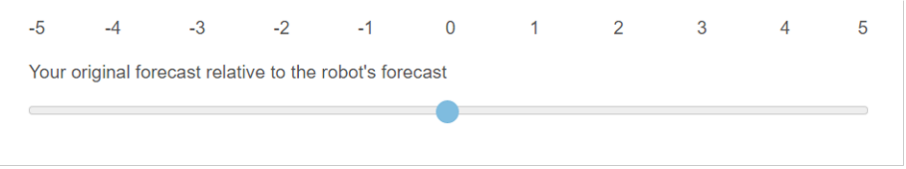
After you go to the next page, you cannot go back to this page.

[Screen 20]

Please evaluate the accuracy of your original forecast relative to the robot's forecast in this task.

i.e. **from -5 (the lowest, your original forecast is less accurate than the robot's forecast to a great extent.) to 5 (the highest, your original forecast is more accurate than the robot's forecast to a great extent.)**

0 indicates that your original forecast has the same accuracy as the robot's forecast's.



Your original forecast relative to the robot's forecast

We now finish the experiment. Please complete the following questionnaire. Thank you.

After you finish the questionnaire, we will show you the experiment results and your rewards.

J Comparison between framed experiments and non-framed experiments

In our original experiment (reported in the main text), we asked participants to play a role as financial advisors who forecast the future stock price based on historical price information. The participants were also told that their company created an algorithm that was designed to forecast future stock prices. These two aspects may have increased participants' reliance on the algorithm even when its performance is low.

To investigate the impact of this framing on the reliance on algorithms, we have conducted a new set of experiments without such a framing. In this new set of non-framed experiments, we remove the framing about the role of financial advisors and the developer (i.e. their company) of the algorithms. Namely, participants are told that "In this experiment, you are asked to forecast the future stock price based on historical price information. A robot has been created to forecast future stock prices." The other aspects of the experimental design as well as procedure of the non-framed experiments are identical to the original framed experiments. The additional non-framed experiment was conducted online between August and September 2021. Total of 252 participants, who have never participated to similar experiments, have participated from the same pool of participants.

We compare MSHIFT between framed experiments and non-framed experiments to investigate the effect of framing on participants' their reliance on the algorithms in framed experiments. Results are shown in Table J1. We found that MSHIFTs are not statistically significantly different between the framed experiments and the non-framed experiments in all treatments and tasks, except Task 2 in Treatment 5. However, the MSHIFT is significantly higher in the non-framed experiments than in the framed experiments in task 2 in Treatment 5. Therefore, the reliance on the algorithm in the framed experiments are due to the wording of the instruction regarding the financial advisor and their company developing in the algorithm.

The comparisons of MAPE of the initial (Table J2) and the final (Table J3) forecasts show no significant difference between the framed and non-framed experiments, except task 2 in Treatment 2. The MAPE of the initial forecast is significantly higher in the framed experiments than in the non-framed experiments in task2 in Treatment 2.

Table J1 Comparison of MSHIFT between framed experiments and non-framed experiments using two-sample t-test

Treatment	Task	Framed MSHIFT (Std. Err.)	Obs.	Non-framed MSHIFT (Std. Err.)	Obs.	t-value (p-value)
T1	1	0.624 (0.031)	49	0.562 (0.027)	39	1.479 (0.143)
T1	2	0.515 (0.031)	49	0.514 (0.029)	39	0.024 (0.981)
T2	1	0.481 (0.027)	47	0.517 (0.045)	42	-0.701 (0.485)
T2	2	0.438 (0.029)	47	0.439 (0.036)	42	-0.022 (0.983)
T3	1	0.552 (0.035)	50	0.503 (0.038)	40	0.955 (0.342)
T3	2	0.482 (0.035)	50	0.451 (0.033)	40	0.634 (0.528)
T4	1	0.476 (0.037)	50	0.419 (0.030)	43	1.173 (0.244)
T4	2	0.469 (0.044)	50	0.421 (0.030)	43	0.878 (0.382)
T5	1	0.198 (0.025)	45	0.212 (0.031)	42	-0.362 (0.718)
T5	2	0.171 (0.026)	45	0.294 (0.052)	42	-2.155 (0.034)
T6	1	0.277 (0.030)	47	0.265 (0.034)	46	0.249 (0.804)
T6	2	0.168 (0.055)	47	0.236 (0.029)	46	-1.085 (0.281)

Note: The number of observations is the number of participants in each treatment.

Table J2 Comparison of MAPE of initial forecast between framed experiments and non-framed experiments using two-sample t-test

Treatment	Task	Framed MAPE (Std. Err.)	Obs.	Non-framed MAPE (Std. Err.)	Obs.	t-value (p-value)
T1	1	7.585 (0.485)	49	6.952 (0.422)	39	0.957 (0.341)
T1	2	8.738 (0.401)	49	8.130 (0.338)	39	1.124 (0.264)
T2	Practice	8.300 (0.578)	47	7.866 (0.447)	42	0.584 (0.561)
T2	1	6.750 (0.314)	47	6.575 (0.342)	42	0.378 (0.706)
T2	2	8.571 (0.284)	47	7.761 (0.184)	42	2.329 (0.022)
T3	Practice	8.064 (0.386)	50	7.980 (0.442)	40	0.144 (0.886)
T3	1	6.548 (0.306)	50	6.473 (0.235)	40	0.187 (0.852)
T3	2	8.216 (0.242)	50	8.495 (0.405)	40	-0.617 (0.539)
T4	1	8.159 (0.412)	50	7.480 (0.382)	43	1.194 (0.236)
T4	2	8.737 (0.302)	50	8.633 (0.302)	43	0.242 (0.810)
T5	Practice	8.100 (0.335)	45	7.754 (0.350)	42	0.715 (0.477)
T5	1	6.551 (0.286)	45	6.669 (0.339)	42	-0.267 (0.790)
T5	2	8.182 (0.218)	45	8.081 (0.245)	42	0.308 (0.759)
T6	Practice	7.861 (0.359)	47	8.195 (0.458)	46	-0.577 (0.566)
T6	1	7.147 (0.420)	47	6.901 (0.501)	46	0.377 (0.707)
T6	2	8.636 (0.340)	47	8.557 (0.317)	46	0.170 (0.865)

Note: The number of observations is the number of participants in each treatment.

Table J3 Comparison of MAPE of final forecast between framed experiments and non-framed experiments using two-sample t-test

Treatment	Task	Framed MAPE (Std. Err.)	Obs.	Non-framed MAPE (Std. Err.)	Obs.	t-value (p-value)
T1	1	6.094 (0.136)	49	6.170 (0.142)	39	-0.383 (0.703)
T1	2	7.327 (0.122)	49	7.253 (0.104)	39	0.450 (0.654)
T2	1	6.112 (0.100)	47	6.309 (0.205)	42	-0.890 (0.376)
T2	2	7.497 (0.124)	47	7.293 (0.088)	42	1.324 (0.189)
T3	1	5.806 (0.120)	50	6.009 (0.130)	40	-1.144 (0.256)
T3	2	7.303 (0.103)	50	7.410 (0.145)	40	-0.619 (0.538)
T4	1	10.282 (0.380)	50	10.861 (0.362)	43	-1.092 (0.278)
T4	2	10.144 (0.284)	50	10.070 (0.245)	43	0.195 (0.846)
T5	1	7.817 (0.355)	45	7.977 (0.400)	42	-0.300 (0.765)
T5	2	8.726 (0.246)	45	8.956 (0.286)	42	-0.613 (0.542)
T6	1	8.024 (0.357)	47	8.240 (0.424)	46	-0.391 (0.697)
T6	2	9.097 (0.281)	47	9.100 (0.316)	46	-0.007 (0.995)

Note: The number of observations is the number of participants in each treatment.

K Results of non-framed experiments

In this appendix, we report the results of the same set of analyses as in the framed experiment for the non-framed experiment. The results are qualitatively the same as in the framed experiment.

Table K1 Comparison between MSHIFT and the halfway point between the algorithm's forecast and the initial forecast in non-framed experiments using single-sample t-test

Treatment	Task	MSHIFT (Std. Err.)	Obs.	Halfway between algorithm's forecast and initial forecast	t-value (p-value)
T1	1	0.562 (0.027)	39	0.5	2.246 (0.031)
T1	2	0.514 (0.029)	39	0.5	0.495 (0.623)
T4	1	0.419 (0.030)	43	0.5	-2.697 (0.010)
T4	2	0.421 (0.030)	43	0.5	-2.649 (0.011)

Note: The number of observations is the number of participants in each treatment.

Table K2 Predicted evaluation rate, predicted MSHIFT, and predicted MAPE for treatment dummies in non-framed experiments using OLS regression with robust standard error

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Evaluation	Evaluation	MSHIFT	MSHIFT	MAPE	MAPE	MAPE	MAPE
Variables	Task 1	Task 2	Task 1	Task 2	initial	initial	final	final
					forecast	forecast	forecast	forecast
					Task 1	Task 2	Task 1	Task 2
Treatment 1	-1.154	-0.744	0.562	0.514	6.952	8.130	6.170	7.253
	(0.319)	(0.351)	(0.027)	(0.029)	(0.422)	(0.338)	(0.142)	(0.104)
Treatment 2	-1.762	-0.810	0.517	0.439	6.575	7.761	6.309	7.293
	(0.274)	(0.326)	(0.045)	(0.036)	(0.342)	(0.184)	(0.205)	(0.088)
Treatment 3	-1.375	-1.250	0.503	0.451	6.473	8.495	6.009	7.410
	(0.272)	(0.260)	(0.038)	(0.032)	(0.234)	(0.405)	(0.130)	(0.144)
Treatment 4	-1.116	-1.116	0.419	0.421	7.480	8.633	10.861	10.070
	(0.316)	(0.298)	(0.030)	(0.030)	(0.383)	(0.302)	(0.362)	(0.245)
Treatment 5	1.333	1.500	0.212	0.294	6.669	8.081	7.977	8.956
	(0.289)	(0.290)	(0.031)	(0.052)	(0.339)	(0.245)	(0.400)	(0.286)
Treatment 6	1.239	1.022	0.265	0.236	6.901	8.557	8.240	9.100
	(0.302)	(0.315)	(0.034)	(0.029)	(0.501)	(0.317)	(0.425)	(0.316)
	Prob > F	Prob > F	Prob > F	Prob > F	Prob > F	Prob > F	Prob > F	Prob > F
T1=T2=T3								
=T4=T5=T6	0.000	0.000	0.000	0.000	0.343	0.097	0.000	0.000
T1=T2=T3	0.330	0.408	0.399	0.187	0.611	0.211	0.427	0.673
T1 = T2 #	0.449	1.000	1.000	0.314	1.000	1.000	1.000	1.000
T1 = T3 #	1.000	0.743	0.621	0.438	0.965	1.000	1.000	1.000
T2 = T3 #	0.952	0.874	1.000	1.000	1.000	0.301	0.651	1.000
T4=T5=T6	0.000	0.000	0.000	0.000	0.279	0.288	0.000	0.006
T4 = T5 #	0.000	0.000	0.000	0.110	0.342	0.470	0.000	0.010
T4 = T6 #	0.000	0.000	0.003	0.000	1.000	1.000	0.000	0.048
T5 = T6 #	1.000	0.795	0.740	0.997	1.000	0.708	1.000	1.000
Observations	252	252	252	252	252	252	252	252

Note: (a) Treatment 1 dummy equals 1 for treatment 1, and 0 otherwise. Treatment 2 dummy equals 1 for treatment 2, and 0 otherwise. Treatment 3 dummy equals 1 for treatment 3, and 0 otherwise. Treatment 4 dummy equals 1 for treatment 4, and 0 otherwise. Treatment 5 dummy equals 1 for treatment 5, and 0 otherwise. Treatment 6 dummy equals 1 for treatment 6, and 0 otherwise. (b) The unit of observation is the number of participants. The total number of observations is the number of participants in all treatments. (c) The robust standard errors are in parentheses. (d) # indicates Bonferroni-adjusted p-values.

Table K3 Comparison of evaluation rate between tasks 1 and 2 using OLS regression of evaluation rate on task dummy in non-framed experiments with robust cluster standard error on participant level

Variables	(1) T1	(2) T2	(3) T3	(4) T4	(5) T5	(6) T6
Task 2 dummy	0.410	0.952**	0.125	0.000	0.167	-0.217
	(0.330)	(0.256)	(0.231)	(0.255)	(0.227)	(0.179)
Constant	-1.154**	-1.762***	-1.375***	-1.116**	1.333***	1.239***
	(0.321)	(0.276)	(0.274)	(0.318)	(0.290)	(0.303)
Observations	78	84	80	86	84	92
R-squared	0.010	0.058	0.001	0.000	0.002	0.003
Clusters	39	42	40	43	42	46

Notes: (a) A task dummy equals 0 for task 1 and 1 for task 2. (b) The unit of observation is the number of participants. The total number of observations is the number of participants in each treatment × 2. (c) The robust standard errors clustered by participant levels are in parentheses. $* * *p < 0.001, * * p < 0.01, and * p < 0.05$.

Table K4 Comparison of evaluation rate between tasks 1 and 2 using OLS regression of evaluation rate on task dummy in non-framed experiments with robust cluster standard error on participant level

Variables	(1) T1	(2) T2	(3) T3	(4) T4	(5) T5	(6) T6
Final forecast	-0.783*	-0.266	-0.465	3.381***	1.308***	1.339***
	(0.359)	(0.252)	(0.238)	(0.341)	(0.311)	(0.355)
Constant	6.952***	6.575***	6.473***	7.480***	6.669***	6.900***
	(0.425)	(0.344)	(0.236)	(0.385)	(0.341)	(0.504)
Observations	78	84	80	86	84	92
R-squared	0.039	0.005	0.037	0.329	0.071	0.044
Clusters	39	42	40	43	42	46

Notes: (a) The final forecast dummy equals 1 for final forecast and 0 for initial forecast. (b) The unit of observation is the number of participants. The total number of observations is the number of participants in each treatment × 2. (c) The robust standard errors clustered by participant levels are in parentheses. $* * *p < 0.001, * * p < 0.01, and * p < 0.05$.

Table K5 Comparison of MAPE between initial forecast and final forecast in task 2 using OLS regression of MAPE on final forecast dummy in non-framed experiments with robust cluster standard error on participant level

Variables	(1) T1	(2) T2	(3) T3	(4) T4	(5) T5	(6) T6
Final forecast	-0.877**	-0.469**	-1.085**	1.437***	0.876***	0.543*
	(0.300)	(0.136)	(0.332)	(0.254)	(0.187)	(0.229)
Constant	8.130***	7.761***	8.495***	8.633***	8.081***	8.557***
	(0.340)	(0.185)	(0.408)	(0.304)	(0.246)	(0.319)
Observations	78	84	80	86	84	92
R-squared	0.075	0.061	0.075	0.140	0.062	0.016
Clusters	39	42	40	43	42	46

Notes: (a) The final forecast dummy equals 1 for final forecast and 0 for initial forecast. (b) The unit of observation is the number of participants. The total number of observations is the number of participants in each treatment × 2. (c) The robust standard errors clustered by participant levels are in parentheses. $* * *p < 0.001, * * p < 0.01, and * p < 0.05$.

Table K6 Summary of participants' personal characteristics in non-framed experiments

	Treatments						One-way ANOVA	
	T1	T2	T3	T4	T5	T6	F	Prob > F
Female	0.421	0.366	0.436	0.488	0.405	0.341	0.48	0.793
	(0.081)	(0.076)	(0.080)	(0.881)	(0.221)	(0.033)		
Undergraduate student	0.718	0.762	0.625	0.738	0.690	0.609	0.76	0.582
	(0.073)	(0.067)	(0.078)	(0.069)	(0.072)	(0.142)		
Financial literacy score	8.128	7.310	7.750	7.581	7.595	8.000	0.73	0.602
	(0.341)	(0.388)	(0.429)	(0.277)	(0.358)	(0.297)		
Risk aversion score	3.974	3.310	3.225	3.395	3.452	3.413	1.35	0.245
	(0.209)	(0.247)	(0.216)	(0.238)	(0.219)	(0.191)		
CRT score	2.308	2.571	2.500	2.581	2.667	2.478	1.1	0.362
	(0.138)	(0.103)	(0.119)	(0.112)	(0.111)	(0.106)		
Obs.	39	42	40	43	42	46		

Notes: The female dummy equals 1 for female, and 0 otherwise. The undergraduate student dummy equals 1 for undergraduate student, and 0 otherwise. Financial literacy score range = 0–12 (higher score indicates greater financial literacy). Risk aversion score range = 0–5 (higher score indicates a higher level of risk aversion). CRT score range = 0–3 (higher score indicates greater cognitive ability).

Table K7 Predicted evaluation rate, predicted MSHIFT, and predicted MAPE for treatment dummies in non-framed experiments using OLS regression conditional on personal characteristics with robust standard error

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Evaluation	Evaluation	MSHIFT	MSHIFT	MAPE	MAPE	MAPE	MAPE
Variables	Task 1	Task 2	Task 1	Task 2	initial	initial	final	final
					forecast	forecast	forecast	forecast
					Task 1	Task 2	Task 1	Task 2
Treatment 1	-1.176	-0.668	0.565	0.513	6.808	8.085	6.076	7.228
	(0.341)	(0.364)	(0.030)	(0.031)	(0.395)	(0.324)	(0.154)	(0.114)
Treatment 2	-1.786	-0.787	0.512	0.438	6.621	7.756	6.370	7.312
	(0.283)	(0.327)	(0.046)	(0.037)	(0.350)	(0.190)	(0.210)	(0.094)
Treatment 3	-1.370	-1.270	0.497	0.448	6.462	8.507	6.022	7.429
	(0.282)	(0.259)	(0.038)	(0.034)	(0.254)	(0.433)	(0.138)	(0.149)
Treatment 4	-0.990	-1.020	0.422	0.414	7.518	8.661	10.900	10.054
	(0.304)	(0.304)	(0.031)	(0.030)	(0.395)	(0.312)	(0.367)	(0.253)
Treatment 5	1.344	1.459	0.213	0.294	6.754	8.104	8.016	8.958
	(0.310)	(0.288)	(0.032)	(0.052)	(0.346)	(0.248)	(0.396)	(0.289)
Treatment 6	1.079	0.813	0.270	0.242	6.977	8.503	8.268	9.059
	(0.293)	(0.295)	(0.036)	(0.031)	(0.515)	(0.328)	(0.441)	(0.328)
	Prob > F	Prob > F	Prob > F	Prob > F	Prob > F	Prob > F	Prob > F	Prob > F
T1=T2=T3=								
T4=T5=T6	0.000	0.000	0.000	0.000	0.382	0.128	0.000	0.000
T1=T2=T3	0.351	0.320	0.338	0.227	0.765	0.245	0.376	0.571
T1 = T2 #	0.530	1.000	0.997	0.385	1.000	1.000	0.775	1.000
T1 = T3 #	1.000	0.549	0.510	0.485	1.000	1.000	1.000	0.874
T2 = T3 #	0.877	0.750	1.000	1.000	1.000	0.347	0.516	1.000
T4=T5=T6	0.000	0.000	0.000	0.000	0.335	0.339	0.000	0.007
T4 = T5 #	0.000	0.000	0.000	0.137	0.430	0.493	0.000	0.014
T4 = T6 #	0.000	0.000	0.005	0.000	1.000	1.000	0.000	0.051
T5 = T6 #	1.000	0.357	0.713	1.000	1.000	1.000	1.000	1.000
Observations	246	246	246	246	246	246	246	246

Note: (a) Treatment 1 dummy equals 1 for treatment 1, and 0 otherwise. Treatment 2 dummy equals 1 for treatment 2, and 0 otherwise. Treatment 3 dummy equals 1 for treatment 3, and 0 otherwise. Treatment 4 dummy equals 1 for treatment 4, and 0 otherwise. Treatment 5 dummy equals 1 for treatment 5, and 0 otherwise. Treatment 6 dummy equals 1 for treatment 6, and 0 otherwise. (b) The unit of observation is the number of participants. The total number of observations is the number of participants in all treatments. (c) The robust standard errors are in parentheses. (d) # indicates Bonferroni-adjusted p-values.

Table K8 Comparison of evaluation rate between tasks 1 and 2 using OLS regression of evaluation rate on task dummy conditional on personal characteristics in non-framed experiments with robust cluster standard error on participant level

Variables	(1) T1	(2) T2	(3) T3	(4) T4	(5) T5	(6) T6
Task dummy	0.421	1.000**	0.103	-0.048	0.167	-0.227
	(0.351)	(0.267)	(0.244)	(0.265)	(0.234)	(0.190)
Female	0.478	-0.363	-1.266*	-0.838	0.175	-1.473*
	(0.644)	(0.646)	(0.500)	(0.482)	(0.474)	(0.600)
Undergraduate student	-0.617	0.183	-0.254	-0.040	-1.154	-0.353
	(0.790)	(0.603)	(0.534)	(0.562)	(0.616)	(0.645)
Financial literacy score	-0.036	-0.126	0.025	0.024	0.067	-0.091
	(0.130)	(0.098)	(0.083)	(0.144)	(0.103)	(0.130)
Risk aversion score	-0.064	-0.058	0.216	-0.231	0.382	-0.085
	(0.216)	(0.156)	(0.159)	(0.178)	(0.252)	(0.195)
CRT score	0.531	-0.061	-0.303	-0.719	1.624***	0.049
	(0.338)	(0.508)	(0.371)	(0.469)	(0.282)	(0.379)
Constant	-1.659	-0.500	-0.780	1.861	-4.099*	2.717
	(1.505)	(1.903)	(1.500)	(1.704)	(1.516)	(1.714)
Observations	76	82	78	84	84	88
R-squared	0.073	0.105	0.133	0.165	0.282	0.152
Clusters	38	41	39	42	42	44

Notes: (a) The task dummy equals 0 for task 1 and 1 for task 2. (b) The unit of observation is the number of participants. The total number of observations is the number of participants in each treatment × 2. (c) The robust standard errors clustered by participant levels are in parentheses. $* * *p < 0.001, * * p < 0.01, and * p < 0.05$.

Table K9 Comparison of MAPE between initial forecast and final forecast in task 1 using OLS regression of MAPE on final forecast dummy conditional on personal characteristics in non-framed experiments with robust cluster standard error on participant level

Variables	(1) T1	(2) T2	(3) T3	(4) T4	(5) T5	(6) T6
Final forecast	-0.743	-0.259	-0.477	3.399***	1.308***	1.28**
	(0.379)	(0.267)	(0.252)	(0.360)	(0.321)	(0.378)
Female	0.483	-0.110	0.022	-0.237	0.597	0.394
	(0.508)	(0.570)	(0.309)	(0.642)	(0.811)	(0.858)
Undergraduate student	0.530	-0.476	0.034	1.086	1.795*	-0.925
	(0.556)	(0.573)	(0.329)	(0.732)	(0.704)	(1.029)
Financial literacy score	0.085	0.064	0.057	0.064	0.119	-0.088
	(0.111)	(0.102)	(0.059)	(0.190)	(0.154)	(0.227)
Risk aversion score	-0.082	0.047	0.106	0.420	-0.310	-0.351
	(0.221)	(0.141)	(0.105)	(0.216)	(0.395)	(0.272)
CRT score	-0.853	-0.387	-0.099	-0.547	-0.568	-0.310
	(0.459)	(0.388)	(0.229)	(0.639)	(0.446)	(0.598)
Constant	7.970***	7.362***	5.910***	6.295*	6.864**	10.075**
	(1.354)	(1.519)	(0.949)	(2.381)	(2.215)	(2.827)
Observations	76	82	78	84	84	88
R-squared	0.188	0.051	0.069	0.402	0.157	0.085
Clusters	38	41	39	42	42	44

Notes: (a) The final forecast dummy equals 1 for final forecast and 0 for initial forecast. (b) The unit of observation is the number of participants. The total number of observations is the number of participants in each treatment × 2. (c) The robust standard errors clustered by participant levels are in parentheses. $* * *p < 0.001, * * p < 0.01, and * p < 0.05$.

Table K10 Comparison of MAPE between initial forecast and final forecast in task 2 using OLS regression of MAPE on final forecast dummy conditional on personal characteristics in non-framed experiments with robust cluster standard error on participant level

Variables	(1)	(2)	(3)	(4)	(5)	(6)
	T1	T2	T3	T4	T5	T6
Final forecast	-0.856*	-0.465**	-1.092**	1.394***	0.875***	0.565*
	(0.318)	(0.144)	(0.352)	(0.264)	(0.193)	(0.246)
Female	0.702	-0.129	0.764	0.253	-0.245	-0.657
	(0.452)	(0.300)	(0.523)	(0.499)	(0.549)	(0.585)
Undergraduate student	0.062	0.434	0.539	0.119	0.394	-0.687
	(0.455)	(0.274)	(0.426)	(0.524)	(0.617)	(0.642)
Financial literacy score	-0.062	-0.015	0.065	0.148	-0.034	0.027
	(0.092)	(0.063)	(0.062)	(0.128)	(0.093)	(0.176)
Risk aversion score	-0.137	-0.040	0.212	-0.113	0.099	0.029
	(0.161)	(0.088)	(0.167)	(0.190)	(0.248)	(0.194)
CRT score	-0.352	-0.092	0.643	-0.033	-0.405	0.086
	(0.389)	(0.214)	(0.319)	(0.263)	(0.325)	(0.369)
Constant	9.639***	7.950***	5.064***	7.811***	8.906***	8.587***
	(1.266)	(0.978)	(1.300)	(1.701)	(1.553)	(2.169)
Observations	76	82	78	84	84	88
R-squared	0.179	0.106	0.194	0.165	0.118	0.074
Clusters	38	41	39	42	42	44

Notes: (a) The final forecast dummy equals 1 for final forecast and 0 for initial forecast. (b) The unit of observation is the number of participants. The total number of observations is the number of participants in each treatment × 2. (c) The robust standard errors clustered by participant levels are in parentheses. $* * *p < 0.001, * * p < 0.01, and * p < 0.05$.

Table K11 Comparison of MAPE between algorithm forecast and initial forecast using paired t-test

Treatment	Task	Algorithm MAPE	Initial forecast MAPE (Std. Err.)	Diff (Algorithm–Initial) MAPE (Std. Err.)	t-value (p-value)	Obs.
1	Task 1	5.866	6.952 (0.422)	-1.086 (0.422)	-2.574 (0.014)	39
1	Task 2	6.862	8.130 (0.338)	-1.268 (0.338)	-3.749 (<0.001)	39
2	Practice	5.889	7.866 (0.447)	-1.977 (0.447)	-4.418 (<0.001)	42
2	Task 1	5.866	6.575 (0.342)	-0.709 (0.342)	-2.074 (0.044)	42
2	Task 2	6.862	7.761 (0.184)	-0.899 (0.184)	-4.885 (<0.001)	42
3	Practice	5.889	7.980 (0.442)	-2.091 (0.442)	-4.733 (<0.001)	40
3	Task 1	5.866	6.473 (0.235)	-0.607 (0.235)	-2.589 (0.014)	40
3	Task 2	6.862	8.495 (0.405)	-1.633 (0.405)	-4.031 (<0.001)	40
4	Task 1	12.359	7.480 (0.382)	4.879 (0.382)	12.759 (<0.001)	43
4	Task 2	13.391	8.633 (0.302)	4.758 (0.302)	15.762 (<0.001)	43
5	Practice	10.144	7.754 (0.350)	2.390 (0.350)	6.831 (<0.001)	42
5	Task 1	12.359	6.669 (0.339)	5.690 (0.339)	16.779 (<0.001)	42
5	Task 2	13.391	8.081 (0.245)	5.310 (0.245)	21.711 (<0.001)	42
6	Practice	10.144	8.195 (0.458)	1.949 (0.458)	4.259 (<0.001)	46
6	Task 1	12.359	6.901 (0.501)	5.458 (0.501)	10.896 (<0.001)	46
6	Task 2	13.391	8.557 (0.317)	4.834 (0.317)	15.250 (<0.001)	46

Notes: The number of observations is the number of participants in each treatment.

Table K12 Comparison of MAPE between first five human final forecast and last five human final forecasts in non-framed experiments using paired t-test

Treatment	Task	First five forecasts MAPE (Std. Err.)	Last five forecasts MAPE (Std. Err.)	Diff (First–Last) MAPE (Std. Err.)	t-value (p-value)	Obs.
1	Task 1	6.592 (0.334)	5.747 (0.243)	0.844 (0.510)	1.654 (0.106)	39
1	Task 2	7.819 (0.371)	6.687 (0.388)	1.132 (0.731)	1.550 (0.130)	39
2	Task 1	6.317 (0.345)	6.301 (0.271)	0.016 (0.465)	0.034 (0.973)	42
2	Task 2	7.293 (0.445)	7.292 (0.405)	0.001 (0.832)	0.001 (0.999)	42
3	Task 1	6.117 (0.266)	5.900 (0.241)	0.217 (0.437)	0.497 (0.622)	40
3	Task 2	7.397 (0.416)	7.423 (0.438)	-0.026 (0.804)	-0.033 (0.974)	40
4	Task 1	10.976 (0.510)	10.746 (0.610)	0.231 (0.860)	0.268 (0.790)	43
4	Task 2	9.532 (0.546)	10.608 (0.381)	-1.076 (0.805)	-1.337 (0.188)	43
5	Task 1	7.678 (0.566)	8.276 (0.557)	-0.599 (0.787)	-0.761 (0.451)	42
5	Task 2	9.263 (0.417)	8.650 (0.440)	0.612 (0.639)	0.958 (0.344)	42
6	Task 1	8.669 (0.575)	7.812 (0.421)	0.856 (0.545)	1.571 (0.123)	46
6	Task 2	9.077 (0.528)	9.123 (0.567)	-0.046 (0.894)	-0.052 (0.959)	46

Notes: The number of observations is the number of participants in each treatment.

Table K13 OLS linear regression of MAPE of human forecast in practice stage on MSHIFT in tasks 1 and 2 with the good and bad algorithms in non-framed experiments, with robust standard errors

Variables	(1) MSHIFT Task1 Good algorithm Treatment 2	(2) MSHIFT Task2 Good algorithm Treatment 2	(3) MSHIFT Task1 Bad algorithm Treatment 5	(4) MSHIFT Task2 Bad algorithm Treatment 5
MAPE of human forecast in practice stage	-0.002	0.005	0.037**	0.007
	(0.012)	(0.008)	(0.013)	(0.019)
Constant	0.531***	0.398***	-0.075	0.240
	(0.108)	(0.074)	(0.105)	(0.190)
Observations	42	42	42	42
R-squared	0.000	0.004	0.177	0.002

Notes: (a) The unit of observation is the number of participants. The total number of observations is the number of participants in T2 in model (1) (2) and T5 in model (3) (4). (b) The robust standard errors are in parentheses. $* * *p < 0.001, * * p < 0.01, and * p < 0.05.$

Table K14 Comparison of MAPE between algorithm forecast and final forecast using paired t-test

Treatment	Task	Algorithm MAPE	Final forecast MAPE (Std. Err.)	Diff (Algorithm–Final) MAPE (Std. Err.)	t-value (p-value)	Obs.
1	Task 1	5.866	6.170 (0.142)	-0.304 (0.142)	-2.140 (0.039)	39
1	Task 2	6.862	7.253 (0.104)	-0.391 (0.104)	-3.756 (<0.001)	39
2	Task 1	5.866	6.309 (0.205)	-0.443 (0.205)	-2.159 (0.037)	42
2	Task 2	6.862	7.293 (0.088)	-0.431 (0.088)	-4.913 (<0.001)	42
3	Task 1	5.866	6.009 (0.130)	-0.143 (0.130)	-1.100 (0.278)	40
3	Task 2	6.862	7.410 (0.145)	-0.548 (0.145)	-3.792 (<0.001)	40
4	Task 1	12.359	10.861 (0.362)	1.498 (0.362)	4.135 (<0.001)	43
4	Task 2	13.391	10.070 (0.245)	3.321 (0.245)	13.569 (<0.001)	43
5	Task 1	12.359	7.977 (0.400)	4.382 (0.400)	10.951 (<0.001)	42
5	Task 2	13.391	8.956 (0.286)	4.435 (0.286)	15.520 (<0.001)	42
6	Task 1	12.359	8.240 (0.424)	4.119 (0.424)	9.706 (<0.001)	46
6	Task 2	13.391	9.100 (0.316)	4.291 (0.316)	13.584 (<0.001)	46

Notes: The number of observations is the number of participants in each treatment.

References

Finucane ML, Gullion CM (2010) Developing a tool for measuring the decision-making competence of older adults. Psychology and aging 25(2):271—288

Hanaki N, Inukai K, Masuda T, et al (2021) Participants' characteristics at ISER-lab in 2020. ISER Discussion Paper 1141, ISER, Osaka University

Masuda T, Lee E (2019) Higher order risk attitudes and prevention under different timings of loss. Experimental Economics 22(1):197–215

Noussair C, Trautmann S, Kuilen G (2011) Higher order risk attitudes, demographics, and financial decisions. Review of Economic Studies 81:325–355