Beware of artificial intelligence's ability before you trust them: Evidence from a stock price forecasting experiment

Tsz Kwan TSE[a] **Nobuyuki HANAKI**[b] **Bolin MAO**[c]

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 could freely decide on 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 was not significantly different. Experienced participants relied less on the algorithm than inexperienced participants, regardless of its performance level. The reliance on the low performing algorithm was positive even when participants could infer 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 of concern.

Keywords: algorithms, financial market, forecasting, modification, technology adoption
JEL classification: C9, D9, G17

[a] Osaka University. E-mail: tszkwantse@iser.osaka-u.ac.jp
[b] Osaka University. E-mail: nobuyuki.hanaki@iser.osaka-u.ac.jp
[c] Kyoto University. E-mail: meta.bolin.mao@gmail.com

1. Introduction

1.1 Background

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 future stock prices and to trade. Such a rise in the use of AI allows investors to utilize advice generated by AIs 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 AIs in their decision-making. In this paper, we investigate the extent to which individuals rely on inputs from AI (an algorithm) in forecasting future stock prices.

1.2 Literature

The literature disagrees about people's tendency to rely on algorithms in making decisions. On the one hand, there are studies reporting evidence of "algorithm aversion" (Dietvorst et al., 2015), i.e., people's tendency to rely more on inputs from humans than from algorithms. Castelo et al. (2019) argue that the degree of algorithm aversion 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.

On the other hand, Logg et al. (2019) present 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. They note that the "algorithm aversion" found in prior studies may simply be a manifestation of "advice aversion" (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). "Advice aversion" may also explain the significant increase in reliance on algorithms when participants can slightly adjust the advice given by the algorithm in making their final decision, compared with situations in which such adjustment is not possible (Dietvorst et al., 2018).

With the exception of Dietvorst et al. (2015), many of the experimental studies that ask participants to choose between their own judgment and the one from the algorithms, including two studies that investigated forecasting future stock prices (Önkal et al., 2009; Castelo et al., 2019), do not give participants the opportunity to experience the task and compare their own performance with that of the algorithms before deciding how much to rely on the algorithm. Thus, participants' reluctance to rely on the algorithm (Önkal et al., 2009) as well as willingness to do so (Castelo et al., 2019) may simply be due to differences in participants' subjective judgment

about their own skills and those of the algorithms in the specific tasks studied, as task dependency of reliance on algorithms (Castelo et al., 2019) suggests.

As already noted, in the experiment by Dietvorst et al. (2015), participants were given opportunities to directly compare their own performance with that of the algorithms before deciding how much to rely on the algorithms. The authors found that participants were especially averse to algorithmic forecasters after seeing them err, even when they saw them outperform a human forecaster. It is not clear, however, how the degree of reliance would depend on the relative performance of the algorithm and participants themselves.

1.3 Research questions

In this study, we aimed to fill this gap in the literature by examining how participants' reliance on algorithms varies depending on how much they know about their own performance and that of the algorithm in the given task. In particular, we proposed the following set of research questions:

R1: Does the degree of reliance on the algorithm by participants who have little experience in the specific task vary depending on the performance level of the algorithm?

R2: How do participants' experiences and learning about their own skill in the given task influence their degree of reliance on the algorithm?

2. Experimental design

In our experiment, participants were shown a series of 20 graphs of 12 months' worth of end-of-day prices of randomly selected stocks from the S&P 500 starting from a randomly selected day between January 1, 2008, and December 1, 2018

For each graph, participants were asked to forecast the closing price of these stocks 30 days after the last price shown on the graph. Participants first entered their forecast for each of the 10 graphs (shown in random order). Then, for the same set of 10 graphs, one by one in a random order, 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 forecasts (task 2). The order of two tasks and the order of the 10 graphics within each task were randomized across participants.

We measured performance of the algorithm as well as that of a participant for a particular forecasting task by the absolute percentage error (APE)[1] of their forecast from the realized price.

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

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

In each treatment, participants were told that an algorithm was designed to forecast stock prices as follows: *"This algorithm makes the future stock price forecast 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, Technology)."*

Participants were also told that the mean absolute percentage error (MAPE)[2] of the algorithm was either around 4.9% (Good algorithm in Treatment 1 (or T1), T2, and T3) or 18.4% (Bad algorithm in T4, T5, and T6).[3]

To vary the opportunity for participants to learn about their own and the algorithms' performance, there was a practice stage in four of our treatments. In the practice stage, just as in the main task, participants were shown a series of 10 graphs generated in the same way as the main task, and for each graph, they forecast the end-of-day price for these stocks 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 only their own performance (T2 and T5) or both their own and the algorithm's performances (T3 and T6) for each of the 10 stocks separately as well as the average across all 10 stocks. That is, 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, 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.

At the end of each task, participants were asked to evaluate the accuracy of their forecasts relative to those of the algorithm between –5 (*the lowest, where your forecast is less accurate than the algorithm's forecast to a great extent*) and 5 (*the highest, where your forecast is 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)

---

[2] $MAPE = \left(\frac{1}{n}\sum\left|\frac{Forecast - realized\ price}{realized\ price}\right|\right) \times 100$, where n = 5311, which is the number of predictions used to measure the performance of the algorithm.

[3] We have created two types of algorithm (good algorithm and bad algorithm). We have designed the good and bad algorithms so that they would perform better and worse, respectively, than humans on average.

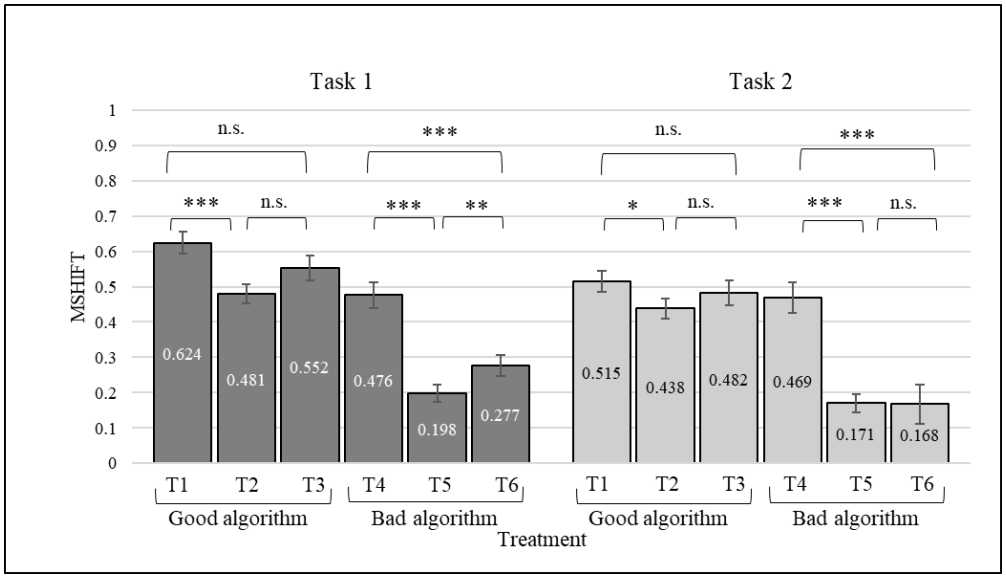
as follows: $Reward\ points = \text{Max}\left[200 - 100 \times \left|\frac{your\ final\ forecast - realized\ price}{realized\ price} \times 100\right|, 0\right]$.

The exchange rate was 1 point = JPY 6.

3. Results

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. Participants received JPY 500 as a participation fee for completing 45 minutes of experiments, and could earn up to an additional JPY 1200 reward depending on their forecasting performance.

We measured the degree of "reliance on algorithms" (Logg et al., 2019; Castelo et al., 2019) by the "shift rate" (Önkal et al., 2009), which is defined for participant i in 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 > 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 shift rate < 0.5. We calculated the mean shift rate (MSHIFT) of 10 graphs in each task in each treatment.



Note: p-values are calculated based on the F test comparing the estimated coefficient on treatment dummies. *, **, and *** indicate significance at the 0.10, 0.05, and 0.01 levels, respectively. n.s. means the difference is not statistically significant at 0.1.

Fig 1. MSHIFT in tasks 1 and 2

Figure 1 shows the MSHIFT in task 1 (dark gray) and task 2 (light gray) in each treatment. We

found that the degree of reliance on the algorithms did not differ depending on the performance level of the algorithm for those participants with little experience in the task (and thus, with little idea about their own skill). Those participants who had experienced the task and learned about their own skill relied on the algorithm significantly less than those without experience, both when they could infer that they had 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 the 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. While some recent research is concerned about 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 be concerned about the excessive reliance on algorithms.

Reference

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: Participants 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: Participants will use imperfect algorithms if they can (even slightly) modify them. Management Science, 64(3), 1155-1170.

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

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

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.