Algorithmic Decision-Making, Economic Behavior and Predictability in Financial Markets

A thesis submitted in fulfillment
of the requirements for the degree of Dr. rer. pol.
from the Faculty of Economic Sciences
University of Göttingen

by

Jan René Judek
born in Hildesheim

Göttingen, 2023

First Supervisor:	Prof. Dr. Kilian Bizer
Second Supervisor:	Prof. Dr. Markus Spiwoks
Third Supervisor:	Prof. Dr. Holger A. Rau

Abstract

Digital transformation is producing a growing number of technological innovations that have an impact on our daily lives. In a variety of areas, economic agents increasingly have the opportunity to interact with algorithms, as shown, for example, by the offering of robo-advisors, and thus also to influence events on financial markets. This thesis aims to examine the behavior of economic agents when interacting with algorithms and their willingness to use them in order to contribute to a better understanding of Algorithm Aversion. Algorithm Aversion describes the negative attitude towards the use of algorithms that economic agents often exhibit once they realize that algorithms are superior but not error-free. The first part of this thesis consists of five experimental studies in this regard. The first contribution shows that Algorithm Aversion in repeated tasks can be partially reduced by increasing experience over time. The second contribution addresses the scope of a decision and shows that the use of algorithms is often rejected in situations where the consequences of an error are serious, even though their use has a higher chance of success. The third contribution shows that possible user interventions in the prediction generation process reduce Algorithm Aversion more reliably if they are granted on the prediction result (output of the algorithm) instead of on the configuration (input of an algorithm). The fourth contribution examines the impact of proxy decisions on Algorithm Aversion. However, making decisions for third parties does not reduce the extent of Algorithm Aversion. The fifth contribution shows that the decision behavior to use an algorithm varies with the prior adoption rate of other economic agents, and that a prior high adoption rate leads to more frequent use of an algorithm than a prior low adoption rate. Overall, Algorithm Aversion proves to be highly robust and can contribute to suboptimal decisions. Overcoming Algorithm Aversion is essential to exploit the great potential that technological innovation brings to forecasting. The second part of this thesis consists of two more papers that contribute to the literature on the quality of capital market forecasts. While the sixth contribution examines the quality of interest rate forecasts in the Latin American region, the seventh contribution focuses on stock market forecasts for three major indices. Overall, the capital market forecasts examined are in most cases inadequate. While forecasts in the Latin American region largely reflect current rather than future interest rate developments, stock index forecasts show that most stock market analysts underestimate the variability of reality and tend toward conservatism. Therefore, it is crucial to improve forecasting models and to react more flexibly to new developments.

Zusammenfassung

Die digitale Transformation bringt immer mehr technische Innovationen hervor, die Auswirkungen auf unser tägliches Leben haben. In einer Vielzahl von Bereichen haben Wirtschaftsakteure zunehmend die Möglichkeit zur Interaktion mit Algorithmen, wie beispielsweise das Angebot von Robo-Advisors zeigt, und damit auch das Geschehen auf den Finanzmärkten zu beeinflussen. Die vorliegende Arbeit hat zum Ziel, das Verhalten von Wirtschaftsakteuren im Umgang mit Algorithmen und deren Bereitschaft zur Nutzung zu untersuchen, um zu einem besseren Verständnis der Algorithm Aversion beizutragen. Die Algorithm Aversion beschreibt die ablehnende Haltung gegenüber dem Einsatz von Algorithmen, welche Wirtschaftsakteure häufig entwickeln, sobald sie erkennen, dass Algorithmen zwar überlegen, aber nicht fehlerfrei sind. Der erste Teil dieser Arbeit umfasst hierzu fünf experimentelle Studien. Der erste Beitrag zeigt, dass die Algorithm Aversion bei wiederholten Aufgaben durch eine zunehmende Erfahrung im Laufe der Zeit teilweise reduziert werden kann. Der zweite Beitrag befasst sich mit der Tragweite einer Entscheidung und zeigt, dass insbesondere in Situationen, die im Fehlerfall schwerwiegende Konsequenzen nach sich ziehen können, häufig auf den Einsatz von Algorithmen verzichtet wird, obwohl deren Nutzung eine höhere Erfolgschance aufweist. Der dritte Beitrag zeigt, dass mögliche Eingriffe eines Nutzers im Prozess der Prognoseerstellung die Algorithm Aversion zuverlässiger reduzieren, wenn diese auf das Prognoseergebnis (Output des Algorithmus), statt auf die Konfiguration (Input eines Algorithmus) gewährt werden. Im vierten Beitrag wird der Einfluss von Stellvertreterentscheidungen auf die Algorithm Aversion untersucht. Das Treffen von Entscheidungen für Dritte führt jedoch nicht zu einer Verringerung des Ausmaßes der Algorithm Aversion. Der fünfte Beitrag zeigt, dass das Entscheidungsverhalten zur Verwendung eines Algorithmus mit der vorherigen Nutzungsrate anderer Wirtschaftsakteure variiert und eine vorherige hohe Akzeptanz eine häufigere Nutzung eines Algorithmus zur Folge hat als eine vorherige schwache Akzeptanz. Gesamtheitlich betrachtet erweist sich die Algorithm Aversion als äußerst robust und kann zu suboptimalen Entscheidungen beitragen. Die Überwindung der Algorithm Aversion ist essenziell, um die großen Potenziale, die technische Innovationen für Prognosen mit sich bringen auszuschöpfen. Zwei weitere Studien bilden den zweiten Teil dieser Arbeit, welche sich in die Literatur zur Güte von Kapitalmarktprognosen einfügen. Während der sechste Beitrag die Güte von Zinsprognosen im lateinamerikanischen Raum untersucht, befasst sich der siebte Beitrag mit Aktienmarktprognosen für drei wichtige Indizes. Insgesamt sind die untersuchten Kapitalmarktprognosen in den meisten Fällen unzureichend. Während die Prognosen im lateinamerikanischen Raum zu einem großen Teil eher die gegenwärtige, statt der zukünftigen Zinsentwicklung widerspiegeln, zeigt sich bei den Aktienindexprognosen, dass Aktienmarktanalysten mehrheitlich die Variabilität der Wirklichkeit unterschätzen und zum Konservatismus neigen. Daher ist es von entscheidender Bedeutung, die Vorhersagemodelle zu verbessern und flexibler auf neue Entwicklungen zu reagieren.

Contents

Chapter I

Introduction and Summaries.....1

Chapter II

Reducing Algorithm Aversion through Experience.....17

Co-authored by Ibrahim Filiz, Marco Lorenz, and Markus Spiwoks

Article in Journal of Behavioral and Experimental Finance, 31(5), 100524, 1-8. (Sep 2021)

Also published as:

Discussion Paper in Sonderforschungsgruppe Institutionenanalyse (sofia), No. 21-1, Darmstadt, January 2021.

Working Paper in Wolfsburg Working Papers (WWP), Ostfalia University of Applied Sciences, No. 21-01, Wolfsburg, January 2021.

Chapter III

The Extent of Algorithm Aversion in Decision-making Situations with Varying Gravity.....39

Co-authored by Ibrahim Filiz, Marco Lorenz, and Markus Spiwoks

Article in PLoS ONE, 18(2), e0278751, 1-21. (Feb 2023)

Also published as:

Discussion Paper in Sonderforschungsgruppe Institutionenanalyse (sofia), No. 21-2, Darmstadt, January 2021.

Working Paper in Wolfsburg Working Papers (WWP), Ostfalia University of Applied Sciences, No. 21-02, Wolfsburg, February 2021.

Chapter IV

Comparing different kinds of influence on an algorithm in its forecasting process and their impact on Algorithm Aversion 69

Co-authored by Zulia Gubaydullina, Marco Lorenz, and Markus Spiwoks

Article in Businesses, 2(4), 448-470. (Oct 2022)

Also published as:

Discussion Paper in Sonderforschungsgruppe Institutionenanalyse (sofia), No. 21-6, Darmstadt, June 2021.

Working Paper in Wolfsburg Working Papers (WWP), Ostfalia University of Applied Sciences, No. 21-04, Wolfsburg, June 2021.

Chapter V

Algorithm Aversion as an Obstacle in the Establishment of Robo Advisors 99

Co-authored by Ibrahim Filiz, Marco Lorenz, and Markus Spiwoks

Article in Journal of Risk and Financial Management, 15(8), 353, 1-25. (Aug 2022)

Also published as:

Discussion Paper in Sonderforschungsgruppe Institutionenanalyse (sofia), No. 22-2, Darmstadt, July 2022.

Working Paper in Wolfsburg Working Papers (WWP), Ostfalia University of Applied Sciences, No. 22-01, Wolfsburg, July 2022.

Chapter VI

Willingness to Use Algorithms Varies with Social Information on Weak vs. Strong Adoption: An Experimental Study on Algorithm Aversion 135

Submitted to Journal of Behavioral Decision Making

Also published as:

Discussion Paper in Sonderforschungsgruppe Institutionenanalyse (sofia), No. 22-5, Darmstadt, December 2022.

Working Paper in Wolfsburg Working Papers (WWP), Ostfalia University of Applied Sciences, No. 23-01, Wolfsburg, January 2023.

Chapter VII

Interest Rate Forecasts in Latin America 153

Co-authored by Ibrahim Filiz, Marco Lorenz, and Markus Spiwoks

Article in Journal of Economic Studies, 49(5), 920-936. (July 2022)

Also published as:

Discussion Paper in Sonderforschungsgruppe Institutionenanalyse (sofia), No. 20-5, Darmstadt, September 2020.

Working Paper in Wolfsburg Working Papers (WWP), Ostfalia University of Applied Sciences, No. 20-02, Wolfsburg, October 2020.

Chapter VIII

Sticky Stock Market Analysts 179

Co-authored by Ibrahim Filiz, Marco Lorenz, and Markus Spiwoks

Article in Journal of Risk and Financial Management, 14(12), 593, 1-27. (Dec 2021)

Also published as:

Discussion Paper in Sonderforschungsgruppe Institutionenanalyse (sofia), No. 21-3, Darmstadt, February 2021.

Working Paper in Wolfsburg Working Papers (WWP), Ostfalia University of Applied Sciences, No. 21-03, Wolfsburg, April 2021.

Chapter I

Introduction and Summaries

In our fast-moving world characterized by technological progress and digital transformation, more and more technical tools are being used. Among other things, the advance of artificial intelligence is producing powerful algorithms that are also having an impact on activities in financial markets. Technical innovations, such as the establishment of robo advisors, have the effect of opening new possibilities for economic agents in their activities on financial markets. As a result, economic agents must also focus on how they deal with technical innovations that have emerged, such as algorithms. How do economic agents react to these new technologies and how do they use them?

The focus of research on human decision behavior has long been on the rationally acting, always seeking the maximum utility and fully informed economic agent. However, in reality economic agents exhibit behavior that is not in line with these approaches (for example, Tversky & Kahneman, 1974). Subsequently, the research streams of behavioral economics and behavioral finance developed to provide a more realistic account of the behavior of economic agents who are subject to judgment biases and use heuristics (for example, Kahneman & Tversky, 1979). Behavioral financial market research incorporates emotional behaviors and cognitive limitations of economic agents to explore, among other things, financial market operations. In this way, factors are identified that cause economic agents to evaluate information inappropriately, leading, for example, to biased perceptions of profits and losses.

In the digital age, technical innovations are constantly being created, resulting in, among other things, increasing availability of algorithms to the general public. As a result, people also have increasing points of contact with algorithmic decision-making systems at work and in everyday life, such as asset management (Niszczota & Kaszás, 2020; Méndez-Suárez, García-Fernández & Gallardo, 2019), medicine (Beck et al., 2011; Ægisdóttir et al., 2006; Grove et al., 2000), justice (Ireland, 2020; Simpson, 2016), or sports (Pérez-Toledano et al., 2019). Technological advances are also opening entirely new possibilities, such as determining the likelihood of ex-offenders to recidivate (Wormith & Goldstone, 1984) or predictive policing (Mohler et al., 2015). However, the integration of artificial intelligence or powerful algorithms can only succeed in organizations if people have trust in the new technologies (Glikson & Woolley, 2020). Current research efforts therefore focus, among other things, on how economic agents interact with algorithmic decision-making systems, or algorithms in general, and examine the willingness to use new technologies in business and financial markets.

Where humans are slowed down in decision making by their cognitive limitations, algorithms are often able to identify complex relationships in large data sets and make more accurate predictions about future developments (Youyou, Kosinski & Stillwell, 2015; Dawes, Faust & Meehl, 1989; Meehl, 1954). However, despite the fact that powerful algorithms have long been available that can perform tasks faster, more accurately, and more cost-effectively than humans (Upadhyay & Khandelwal, 2018), humans often avoid using them, preferring to trust their own judgment or that of a human expert (Prahl & Van Swol, 2017; Dietvorst, Simmons & Massey, 2015). The accuracy of stochastic models is underestimated by economic agents and human predictions are preferred (Önkal et al., 2009).

The term algorithm aversion, established in 2015 by Dietvorst, Simmons & Massey, describes the negative attitude of people towards algorithms. As soon as economic agents realize that an algorithm is superior but not free of errors, they often refrain from using it (Dietvorst, Simmons & Massey, 2015). Forecasting errors committed by an algorithm lead to a stronger rejection attitude than forecast errors committed by humans. That is, following bad advice, willingness to use algorithmic advice decreases more than willingness to use advice from humans (Prahl & Van Swol, 2017). While humans are forgiven for occasional mistakes, algorithmic systems are expected to always make error-free predictions (Alvarado-Valencia & Barrero, 2014).

Forecasting errors of automated systems can at the same time be a possible cause of algorithm aversion and contribute to the fact that economic agents refrain from using algorithmic advice especially after observing forecast errors. The timing of when economic agents are confronted with a forecasting error by an algorithm influences their trust in it. Forecast errors that occur early during interaction with an algorithm have a stronger negative impact on trust than forecast errors that occur later in the interaction (Manzey, Reichenbach & Onnasch, 2012). In the case of prediction errors, task complexity is also relevant. Errors made by automated systems have a greater negative impact on trust when task complexity is perceived to be low than when task complexity is perceived to be high (Madhavan, Wiegmann & Lacson, 2006). If an algorithm is unable to solve tasks with low complexity, economic agents assume that it cannot successfully perform tasks with high complexity either (Hoff & Bashir, 2015). In contrast, if an algorithm makes a forecasting error and economic agents perceive that an algorithm is capable of learning from committed errors, trust in the algorithm's capabilities in turn increases, leading to more frequent use (Reich, Kaju & Maglio, 2022).

The phenomenon known as algorithm aversion is becoming increasingly relevant in research on human decision-making behavior in the context of automated decision-making (for detailed literature reviews on algorithm aversion, see Mahmud et al., 2022; Burton, Stein & Jensen, 2020; Jussupow, Benbasat & Heinzl, 2020). All the efforts to advance technological progress in decision making with algorithms can only positively support humans in their decision-making behavior if the resulting technological tools, such as algorithms, are accepted and used by humans. However, due to reservations about algorithms that are not completely free of errors, there is often a lack of acceptance

for their actual use (Jussupow, Benbasat & Heinzl, 2020; Prahl & Van Swol, 2017; Dietvorst, Simmons & Massey, 2015). Research in the field of behavioral economics addresses not only the causes of algorithm aversion, but also how to reduce it so that people can improve their decisions in cooperation with algorithmic decision systems.

One way to increase the willingness to use algorithms, i.e., to reduce algorithm aversion, is through allowing users to influence the process of algorithmic forecasting. If economic agents gain partial control over an algorithm, this may increase the likelihood of its use (Kawaguchi, 2021; Dietvorst, Simmons & Massey, 2018). In this context, the sense of exerting control may result from a genuine understanding of an algorithm's performance or from adjustments in the process of algorithmic forecasting. It is not relevant whether the control or opportunities for influence affect the operation or performance of an algorithm (Burton, Stein & Jensen, 2020). A real or at least perceived decision control leads to the satisfaction of users' psychological needs and self-interests (Colarelli & Thompson, 2008).

Economic agents want to exert influence on algorithms and participate in making predictions, rather than leave this entirely to an algorithm. For example, being able to provide feedback to an algorithm about its performance increases trust in the algorithmic system and perceptions of its accuracy, even if system accuracy does not improve in response to the feedback (Honeycutt, Nourani & Ragan, 2020). Taking one's expectations into account, such as integrating one's own forecasts into the algorithmic forecasting process, also lead to an increased willingness to use an algorithm (Kawaguchi, 2021). On the other hand, if economic agents are allowed to adjust the algorithmic forecast, they are more likely to use an algorithm. Even if an algorithm commits a forecasting error, users do not withdraw trust from it but persist in its use. This is true even when economic agents are not allowed a free, but only a severely limited, ability to adjust the algorithmic forecast (Dietvorst, Simmons & Massey, 2018). However, adjustments in the process of algorithmic forecasting simultaneously create a conflict of interest: the possibility to make adjustments increases the acceptance to use algorithms, but at the same time the adjustments lead to a decrease in the quality of the final decisions (Sele & Chugunova, 2022).

The perceived objectivity of a task also has an influence on the willingness to use an algorithm. Economic agents who perceive a task to be performed as rather objective are more likely to have it performed by an algorithm than if they perceive the task as rather subjective. However, how objective a task is perceived by economic agents can be changed by means of description (for example, that stock prices would be determined by numerical indicators (objective) or by feelings and intuition (subjective)). The higher the perceived objectivity of a task, the higher the willingness to use an algorithm (Castelo, Bos & Lehman, 2019). Some decision-making situations have a high level of inherent uncertainty. It turns out that economic agents reject the processing of a task by an algorithm especially in these uncertain decision domains (such as medical decisions or consumer demand forecasts) and prefer human judgments (Dietvorst & Bharti, 2020).

In addition, more human-like algorithms can contribute to frequent algorithm use. If an algorithm is more likely to be attributed abilities that are highly human-like by means of description, such as creating music and art or understanding emotions, algorithm aversion decreases. However, this is only effective for subjectively perceived tasks (Castelo, Bos & Lehman, 2019). Humanizing an algorithm by labeling it with a name leads to more frequent use when task complexity is low than when the algorithm is not named. At high task complexity, this effect reverses (Hodge, Mendoza & Sinha, 2021). Even in tasks that require empathy and are therefore presumed to give humans an advantage (for example, judging whether someone else finds a joke funny), an algorithm performs better, and economic agents are still unwilling to use the algorithm (Yeomans et al., 2019).

Another factor in the processing of a task by an algorithm is the reaction time. Economic agents perceive slowly generated forecasts from algorithms as less accurate and are less willing to use them. For human-generated forecasts, this effect is reversed, and slowly generated forecasts are perceived as more accurate (Efendić, van de Calseyde & Evans, 2020). Time pressure when processing a task also reduces algorithm aversion, as economic agents lose confidence in their own forecasts when they are under time pressure (Jung & Seiter, 2021).

Explanations that provide information about the accuracy or success rate of an algorithm increases the willingness of its use. In the case of prediction errors, these explanations are a suitable means of reducing the decline in willingness to use an algorithm (Ben David, Resheff & Tron, 2021). In this context, however, it is also apparent that a user's expectations of an algorithm are relevant. Economic agents are less likely to use an algorithm if the expectations of an algorithm are not met. However, a transparent explanation of how an algorithm works can counteract this. On the other hand, if too much information is provided, this weakens trust in an algorithm (Kizilcec, 2016).

New data sources and new techniques in data analysis have a high potential to increase predictive accuracy (Jung & Seiter, 2021). However, the potentials enabled by technological innovations and algorithms remain unused if economic agents are not willing to use them (Reich, Kaju & Maglio, 2022; Dietvorst & Bharti, 2020). By using algorithmic decision systems, an overall increase in forecast quality or improved accuracy of forecasts can be achieved. Therefore, it is essential to consider the decision-making behavior of economic agents in the context of automated decision making. This dissertation contributes to the scientific debate by considering, among other things, the behavior of economic agents when interacting with algorithms in the context of algorithm aversion. The contributions from chapters 2 to 6 therefore examine the phenomenon of algorithm aversion in more detail and consider, among other things, ways to reduce it.

In **Chapter 2** – *Reducing Algorithm Aversion through Experience (with Filiz, I., Lorenz, M. & Spiwoks, M.)* – algorithm aversion is examined with regard to learning processes. The occurrence of algorithm aversion might be related to the overestimation of one's own abilities (overconfidence). Proeger and Meub (2014) show that economic incentives, repeated feedback, and the resulting increase in experience can be appropriate means to

help economic agents better assess their own abilities. Repeated decision situations, clear feedbacks and economic incentives could thus also contribute to reduce algorithm aversion.

This is examined in an economic experiment in which 143 subjects are given the task of forecasting the price development of a stock in 40 periods. They do not have to predict the exact stock price, but only the trend (rising or falling price) of the stock. The subjects can make their own stock price forecast or use a forecasting computer (algorithm), which forecasts the future stock price development correctly in 70% of cases. A mathematical equation is available for submitting one's own forecasts. In each period, four influencing factors are announced, overlaid by a random influence. In this way, the subjects can mathematically approximate the stock price in the next period. In each of the 40 periods, participants can thus choose between the three options (1) Own forecast: stock price rises, (2) Own forecast: stock price falls and (3) Use of the forecasting computer. A performance-related bonus is paid: each correct forecast is rewarded with a bonus of 0.50 euros. After each individual period, the subjects receive feedback on the success of their decision and the success of the forecasting computer by showing the event that occurred and the current state of bonus. Intuitive stock price forecasts by the subjects are generally inferior to the forecasts of the forecasting computer. It is examined whether an overestimation of one's own forecasting abilities leads to a rejection of the forecasting computer. If subjects learn to better estimate their own abilities through repeated tasks, regular feedback, and economic incentives, algorithm aversion can be reduced through a learning process over the course of 40 periods.

As a result, it is found that a total of 45.9% of the decisions are in favor of the algorithm and 54.1% of the decisions are against the algorithm. Algorithm aversion is thus evident to a considerable extent. Further, the decisions in the first 5 (10/15/20) periods are contrasted with the decisions in the last 5 (10/15/20) periods. This shows that the willingness to use the forecasting computer is relatively low, especially at the beginning of the economic experiment in periods 1 to 8, and that there are major reservations about the algorithm. After that, however, the willingness to use the forecasting computer increases and the subjects seem to realize that their own forecasts are inferior to the algorithm. In the first 5 (10/15/20) periods, the usage rate of the algorithm increases continuously. Looking at the last 20 (15/10/5) periods, however, there are no longer large differences in the usage rate of the algorithm. The proportion of decisions in favor of the algorithm is always around 50%. Regardless of whether the first 5 periods are compared to the last 5 periods or the first 10/15/20 periods to the last 10/15/20 periods, there are significant differences in the subjects' decision behavior: in the respective last periods, the algorithm is used significantly more often than in the respective first periods.

However, only some of the subjects give up their reservations about the forecasting computer. Even at the end of the 40 periods, just under 50% of the subjects still refrain from using the forecasting computer. Although by this time they may already have realized that they cannot achieve better results with their own stock price forecasts than

by using the algorithm. Nevertheless, economic incentives, repeated tasks and continuous feedback seem to be suitable to reduce algorithm aversion at least partially.

In **Chapter 3** – *The Extent of Algorithm Aversion in Decision-making Situations with Varying Gravity (with Filiz, I., Lorenz, M. & Spiwoks, M.)* – algorithm aversion is examined in the context of the possible consequences of a decision-making situation. While an error in some decision situations may prove to be trivial, an error in other decision situations may result in serious consequences. In view of the possible consequences of a decision, one can therefore weigh up whether it should be made oneself or entrusted to an algorithm. However, if the probability of success of a decision can be increased by using an algorithm, it would be reasonable to consult an algorithm especially in decision situations that may have serious consequences. Does algorithm aversion vary in decision contexts with different gravity of possible consequences?

It turns out that framing in decision making with algorithms can be suitable to exert an influence on willingness to use algorithms (Hou & Jung, 2021; Castelo, Bos & Lehmann, 2020). However, research on algorithm aversion does not yet show a consistent picture with regard to the varying severity of the consequences of a decision situation (Utz, Wolfers & Göritz, 2021; Renier, Schmid Mast & Bekbergenova, 2021).

In this study, therefore, six decision situations, which are based on identical mathematical ratios for the successful performance of a task, are placed in different contexts. Three decision situations (driving a car, evaluating MRI scans, and evaluating files in criminal proceedings) can have potentially serious consequences if performed incorrectly. In three other decision situations (finding a partner, selecting cooking recipes, and making weather forecasts), on the other hand, the consequences of incorrect performance may be less severe. In all decision situations, subjects have the choice whether a specialized computer program (algorithm) or trained employees (human experts) should perform the task. The algorithm always has a 70% probability of success and the human experts a 60% probability of success in performing the task. Before the selection is made, the subjects give an assessment of the gravity of the respective decision situation on a scale from 0 (not serious) to 10 (very serious).

Each subject is presented with one of the six decision situations. The subjects are aware that the algorithm has a 10% higher probability of success than the human experts. The subjects receive a performance-based payment of 4.00 euros, which is linked to the successful completion of the task. The successful completion of a task is determined with the help of a deck of cards, in which 7 out of 10 cards lead to the payment of the success-dependent bonus in the case of delegation to the algorithm and 6 out of 10 cards in the case of delegation to the human experts.

The assessment of the gravity shows that the possible consequences of the decision-making situations are perceived differently by the subjects. In three decision situations with potentially serious consequences, the mean value of the perceived gravity is 9.00. For three decision situations with potentially less serious consequences, the mean value is 6.54. In the non-serious decision situations, 70.83% of the subjects decide to delegate the

task to the algorithm. In the serious decision situations, on the other hand, only 50.70% of the subjects decide to delegate the task to the algorithm. Thus, significantly fewer subjects decide in favor of the algorithm. However, by deciding against the algorithm, the subjects reduce the probability of successfully completing the task. Particularly in situations with potentially serious consequences, decision-makers should be interested in choosing the alternative that has the highest probability of success to avert danger to life and limb. It is concluded that algorithm aversion occurs most frequently where it can cause the most harm.

In **Chapter 4** – *Comparing different kinds of influence on an algorithm in its forecasting process and their impact on Algorithm Aversion (with Gubaydullina, Z., Lorenz, M. & Spiwoks, M.)* – it is examined whether algorithm aversion can be reduced by giving decision makers the possibility to influence the configuration of an algorithm (algorithmic input). The fact that an opportunity to influence the result of an algorithm (algorithmic output) contributes to the reduction of algorithm aversion has already been shown in the literature (Dietvorst, Simmons & Massey, 2018). Other studies suggest that an influence on algorithmic input can also reduce the extent of algorithm aversion (Kawaguchi, 2021; Jung & Seiter, 2021; Burton, Stein & Jensen, 2020; Nolan & Highhouse, 2014). In this study, therefore, decision makers are granted the opportunity to participate in the configuration of an algorithm by allowing influence on the weighting of an input factor of an algorithm.

The subjects are asked to forecast the exact price development of a stock in ten consecutive periods. In addition to the option of submitting their own forecast, a forecasting computer (algorithm) is available, which in 7 out of 10 cases delivers a forecast that deviates by less than 15 euros from the stock price that occurred. A performance-related bonus is paid, which is higher the closer the forecast is to the stock price that occurred. The economic experiment is conducted in three treatments: In Treatment 1, a decision is made once before the first forecast is issued as to whether the subjects' own forecasts or the forecasts of an algorithm are to be used to determine the bonus. In Treatment 2, the subjects again choose between their own forecasts and the forecasts of the algorithm, which can be adjusted by up to +/- 5 euros. In Treatment 3, the subjects choose between their own forecasts and the forecasts of the algorithm, whereby the subjects can influence the weighting of an input factor of the algorithm once in advance.

As a result, it is found that the different possibilities of influence cause different decisions to use an algorithm. In Treatment 1 (no possibility to influence) only 44.23% of the subjects decide to use the algorithm. If, on the other hand, the result of the algorithm can be adjusted by up to +/- 5 euros (Treatment 2), 69.23% of the subjects decide to use the algorithm. In Treatment 3 (possibility to influence the algorithmic input), most subjects (58.49%) still decide to use the algorithm. Thus, the possibility to influence the algorithmic output is more likely to reduce algorithm aversion. The possibility to influence the algorithmic input, on the other hand, is only conditionally suitable for reducing algorithm aversion: algorithm aversion decreases, but the effect does not prove to be statistically significant. People want to retain the upper hand over the algorithm in the process of decision-making and have the last word before making a forecast.

In **Chapter 5** – *Algorithm Aversion as an Obstacle in the Establishment of Robo Advisors (with Filiz, I., Lorenz, M. & Spiwoks, M.)* – the influence of proxy decisions on algorithm aversion is examined. That is, one economic agent makes decisions for another economic agent. Some studies suggest that economic agents show altered levels of care and/or risk-taking when making decisions for others (Andersson et al., 2022; Eriksen, Kvaløy & Luzuriaga, 2020; Vieider et al., 2016; Pahlke, Strasser & Vieider, 2015). This could lead economic agents making decisions for others trying their best to make meaningful decisions, potentially contributing to a reduction in algorithm aversion.

The subjects are asked to make investment decisions in an economic experiment. To this end, four tasks are set for the formation of a stock portfolio in each case. In each of the four cases, there are two stocks to choose from, a certain number of which can be included in a portfolio. Subjects are given information about the dividend payment of the stocks. The subjects' task is to compose the portfolio in such a way that the highest possible dividends are achieved with the lowest possible risk. Subjects can either make the diversification decision themselves or delegate it to a robo advisor (algorithm) specialized in making meaningful portfolio decisions. The experiment is conducted in two treatments: in the treatment "Self", a subject makes the portfolio decision and benefits from the success of the decision itself. In the treatment "Representative", one subject makes the portfolio decision and another subject benefits from the success of the decision.

The result shows that overall, only 40.3% of the decisions are in favor of the robo advisor and 59.7% in favor of own diversification decisions, although the robo advisor finds optimal portfolios more often. Algorithm aversion is therefore present. However, a comparison of the treatments "Self" and "Representative" shows no significant difference in the usage of the robo advisor. In the treatment "Self", the proportion of decisions in favor of the algorithm is 40.9%. In the treatment "Representative", the proportion of decisions for the algorithm is 39.7%. Thus, making decisions for others have no effect on algorithm aversion.

However, a significant difference in decision behavior between the two treatments becomes evident when considering the subjects' own diversification decisions: while the optimal portfolio is found in only 30.2% of the decisions in the treatment "Self", this is achieved in 41.5% of the decisions in the treatment "Representative". Thus, in the treatment "Representative", the subjects make a greater effort to find the optimal portfolio with their own diversification decision. However, the stronger efforts are not reflected in a higher usage of the robo advisor. The reservations about using the algorithm are apparently stronger than the effort to make decisions for others with special diligence.

In **Chapter 6** – *Willingness to Use Algorithms Varies with Social Information on Weak vs. Strong Adoption: An Experimental Study on Algorithm Aversion* – the decision behavior to use an algorithm is examined when information about its prior usage rate is provided. Economic agents tend to align their behavior with the behavior of other economic agents, which is also referred to as herd behavior (Spyrou, 2013; Raafat, Chater & Frith, 2009). This can lead economic agents to refrain from incorporating rational aspects into their

decision making and blindly mimic the actions of other economic agents with their decisions (Baddeley et al., 2012). This study therefore provides decision makers with information about the prior usage rate of an algorithm and examines the impact on decision-making behavior as well as algorithm aversion.

In an economic experiment, the subjects are asked to make ten stock price forecasts. They are provided with the price trends of two stocks whose price trends are very similar. The subjects can thus draw conclusions about the development of the price of one stock from the development of the price of the other stock. A bonus is paid depending on the success of the forecasts. The subjects are provided with a forecasting computer to make their stock price forecasts. In 6 out of 10 cases, the forecasts deviate by a maximum of 10 percent from the actual stock price. The subjects must decide once whether their own forecasts or the forecasts of the forecasting computer are to be used to determine the performance-related bonus. The study, which is based on a between-subjects design, consists of two treatments. In the first treatment, the subjects receive information about low social acceptance of the forecasting computer, and in the second treatment about high social acceptance.

As a result, it is found that the subjects' decision behavior to use the forecasting computer differs significantly between the two treatments. The forecasting computer gives stock price forecasts that have, on average, up to 88% lower forecast error than the subjects' own forecasts. When informed about previous low acceptance of the forecasting computer, 51.97% of subjects use the forecasting computer to determine their performance-based bonus. When informed about the previous high acceptance, however, 65.35% of the subjects decide to use the forecasting computer. Furthermore, this effect is mainly since women decide to use the forecasting computer because of a previous high acceptance. In terms of affinity for technology interaction (ATI), subjects who have a low ATI score are more likely to use the forecasting computer. At low acceptance 64.86% and at high acceptance already 84.21% of the less affine subjects use the forecasting computer.

Thus, in addition to providing information about accuracy, information about the previous acceptance of an algorithm also affects algorithm aversion. Economic agents who are informed not only about the accuracy of an algorithm but also about its previous strong acceptance by other economic agents are significantly more willing to use an algorithm than economic agents who are informed not only about its accuracy but also about its previous weak acceptance. Thus, the willingness to use an algorithm varies with the information about its low or high previous usage rate.

In addition to the five contributions on algorithm aversion, two other contributions of the present dissertation focus on the quality of financial market forecasts. Capital market research produces, among other things, tools for predicting future developments in capital markets. The prediction of future developments on financial markets is indispensable for banks and other institutions operating on these markets to always have an overview of the current market situation and to be able to make meaningful economic investment decisions. However, empirical testing of the quality of financial market

forecasts (for example, interest rate and stock market forecasts) often shows that anticipation of future developments in financial markets is not possible or only possible to a limited extent (for a synoptic overview, see Filiz et al., 2019). The contribution in Chapter 7 therefore focuses on the quality of interest rate forecasts and the contribution in Chapter 8 on the quality of stock market forecasts.

In **Chapter 7** – *Interest Rate Forecasts in Latin America (with Filiz, I., Lorenz, M. & Spiwoks, M.)* – the quality of interest rate forecasts in the Latin American region is examined. The quality of forecasts of future interest rate developments is of key importance for banks and investment companies engaged in maturity transformation in the lending business. The study evaluates the quality of short-term interest rate forecasts in Argentina (30-day deposit rate), Brazil (financing overnight rate SELIC), Chile (monetary policy rate), Mexico (28-day closing rate CETES) and Venezuela (30-day deposit rate). The data used are sourced from the journal Latin American Consensus Forecasts. The journal requests forecasts from various banks and institutions on a monthly basis. This study analyzes the forecasts of the aforementioned forecast items published monthly in the Latin American money market in the period from 2001 to 2019. A total of 28,451 interest rate forecasts are available in 209 forecast time series with four- and thirteen-month forecast horizons.

The Diebold-Mariano test, the sign test, the TOTA coefficient and the test for unbiasedness are used to assess the quality of the forecasts. Using the Diebold-Mariano test, a comparison is made to the naïve forecast, allowing an assessment of statistical significance. To perform the test, the mean squared forecast errors for the time series of the expert forecasts and the naïve forecasts are determined. An expert forecast should be expected to perform better in the test than a naïve forecast, which is available free of charge and assumes that no change will occur in the future.

Furthermore, the sign test is carried out, which examines the extent of the forecast change. This checks whether the forecast direction (rising or falling) corresponds to the actual direction in which the interest rate level has developed. In other words, a distinction is made between whether an increase or decrease in the forecast item was forecast and the event that occurred is compared. In this way, it is possible to determine how often the development trend was correctly captured. The results are subjected to a chi-square test to check whether the frequency distribution differs significantly from a random forecast. In this way, a conclusion can be drawn as to whether the forecasts of interest are significantly better or worse than a random forecast.

The TOTA coefficient is used as a third measure of forecast quality to examine the extent to which the forecasts are influenced by the present level of interest rates. To this end, the forecast time series are examined at their respective dates of origin and validity and compared with actual developments. A topically oriented trend adjustment indicates that the forecasts issued reflect the present to a greater extent than the future. For the calculation, both the correlation between the forecasts at their validity dates and the events that occurred and the correlation between the forecasts at their issue dates and the events that occurred are considered. If the TOTA coefficient takes a value of < 1, the

forecasts reflect the present more strongly than the future (topically oriented trend adjustment).

As a result, applying the aforementioned forecast quality measures, it is found that interest rate forecasts in Latin America are relatively often successful in the period from 2001 to 2019. Forecasts of interest rate developments in Brazil, Chile and Mexico can be considered successful. In the totality of the forecasts examined, about 32% of the forecast time series have significantly better expert forecasts than the assumption of naïve forecasts (Diebold-Mariano test). About 78% of the forecast time series can predict the development trend (rising or falling) significantly more accurately than a random forecast (sign test). However, the results on the TOTA coefficient show that the forecasts were strongly influenced by the current interest rate level. About 93% of the forecast time series exhibit the phenomenon of topically oriented trend adjustment and thus reflect present rather than future interest rate developments. Examination of the forecasts for systematic forecast errors also shows that about 98% of the forecast time series are biased (test for unbiasedness).

In **Chapter 8** – *Sticky Stock Market Analysts (with Filiz, I., Lorenz, M. & Spiwoks, M.)* – the forecasting quality of stock market forecasts on the German Stock Market Index (DAX), Dow Jones Industrial Index (DJI) and Euro Stoxx 50 (SX5E) is examined. The forecasts are taken from the German business and daily newspapers Handelsblatt and Frankfurter Allgemeine Zeitung. Once a year, both ask several German private, state, and major banks as well as several international banking houses for stock market forecasts on the aforementioned indices and publish them at the beginning of a year. This study analyzes the stock market forecasts published in the period from 1992 to 2020. A total of 2,761 stock market forecasts are available at six- and twelve-month forecast horizons.

The variability of reality is often underestimated by forecasters, which according to Ogburn (1934) is due to a tendency toward conservatism. This is taken to mean that unusual events are forecast less frequently than they are observed in reality, the standard deviation of forecasts is lower than the standard deviation of actual events, and the average amount of forecast change falls short of the average amount of actual change.

To assess forecast quality, this study follows the methodology of Ogburn (1934) and adds some additional forecast quality measures to the analysis. The prediction-realization diagram compares the predicted percent change and the actual percent change. If the slope of the regression line is < 1, this indicates that the variability of the actual events is underestimated. In addition, the test for unbiasedness is applied, among other things, to examine whether the forecast errors exhibit systematic overestimation or underestimation. If the forecasters tend to be conservative and the forecasts thus turn out to be biased, the regression line in the prediction-realization diagram is too flat (slope < 1). In addition, the Diebold-Mariano test is applied to make a comparison to the naïve forecast. Experts should be more accurate in their forecasts of the future development of the indices than the corresponding naïve forecasts, which are available free of charge.

As a result, for the three forecast items DAX, DJI and SX5E, it is found, irrespective of the forecast horizons, that ordinary events (rising stock index) are overrepresented and unusual events (falling stock index) are underrepresented in the forecasts. Also, the dispersion measured by the standard deviation of the forecasts turns out to be mostly lower than the standard deviation of the actual events. Looking at the slope of the regression lines in the prediction-realization diagrams, the magnitude of the forecast changes lags behind that of the actual changes (slope < 1). The rates of change in the stock indices are significantly underestimated. The test for unbiasedness shows that the forecasts are biased regardless of the forecast horizon. The comparison to the naïve forecast also shows that the accuracy of the analyzed forecasts does not exceed the accuracy of the naïve forecasts.

Ogburn (1934) found that unusual events are forecast less frequently than they are observed in reality, the standard deviation of forecasts is lower than that of actual events, and the magnitude of forecast changes lags behind the magnitude of actual changes. Today's forecasts by stock market analysts exhibit the same findings that Ogburn noted back in 1934. Stock market analysts systematically underestimate the variability of reality and tend toward conservatism. To improve stock market forecasts in the future, it is essential that analysts improve their forecasting models and, above all, react more flexibly and courageously to new developments.

References

Ægisdóttir, S., White, M.J., Spengler, P.M., Maugherman, A.S., Anderson, L.A., Cook, R.S., Nichols, C.N., Lampropoulos, G., Walker, B.S., Cohen, G.R., & Rush, J.D. (2006). The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction. *The Counseling Psychologist*, 34(3), 341-382.

Alvarado-Valencia, J.A., & Barrero, L.H. (2014). Reliance, trust and heuristics in judgmental forecasting. *Computers in Human Behavior*, 36(C), 102-113.

Andersson, O., Holm, H.J., Tyran, J.-R., & Wengström, E.R. (2022). Deciding for Others Reduces Loss Aversion. *Management Science*, 62(1), 29-36.

Baddeley, M., Burke, C.J., Schultz, W., & Tobler, P.N. (2012). Herding in Financial Behaviour: A Behavioural and Neuroeconomic Analysis of Individual Differences. https://doi.org/10.17863/CAM.1041.

Beck, A., Sangoi, A., Leung, S., Marinelli, R. J., Nielsen, T., Vijver, M. J., West, R., Rijn, M.V., & Koller, D. (2011). Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival. *Science Translational Medicine*, 3(108), 108-113.

Ben David, D., Resheff, Y.S., & Tron, T. (2021). Explainable AI and Adoption of Algorithmic Advisors: an Experimental Study. *ArXiv*, *abs/2101.02555*.

Burton, J., Stein, M., & Jensen, T. (2020). A Systematic Review of Algorithm Aversion in Augmented Decision Making. *Journal of Behavioral Decision Making*, 33(2), 220-239.

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

Colarelli, S.M., & Thompson, M.B. (2008). Stubborn Reliance on Human Nature in Employee Selection: Statistical Decision Aids Are Evolutionarily Novel. *Industrial and Organizational Psychology*, 1(3), 347-351.

Dawes, R.M., Faust, D., & Meehl, P.E. (1989). Clinical versus actuarial judgment. *Science*, 243(4899), 1668-1674.

Dietvorst, B.J., & Bharti, S. (2020). People Reject Algorithms in Uncertain Decision Domains Because They Have Diminishing Sensitivity to Forecasting Error. *Psychological Science*, 31(10), 1302-1314.

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.

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), 114-126.

Efendić, E., Van de Calseyde, P.P., & Evans, A.M. (2020). Slow response times undermine trust in algorithmic (but not human) predictions. *Organizational Behavior and Human Decision Processes*, 157, 103-114.

Eriksen, K.W., Kvaløy, O., & Luzuriaga, M. (2020). Risk-taking on behalf of others. *Journal of Behavioral and Experimental Finance*, 26(C), 1-13.

Filiz, I., Nahmer, T., Spiwoks, M., & Bizer, K. (2019). The accuracy of interest rate forecasts in the Asia-Pacific region: opportunities for portfolio management. *Applied Economics*, 51(59), 6309-6332.

Glikson, E., & Woolley, A.W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals*, 14(2), 627-660.

Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E., & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychological Assessment*, 12(1), 19-30.

Hodge, F.D., Mendoza, K.I., & Sinha, R.K. (2021). The effect of humanizing robo-advisors on investor judgments. *Contemporary Accounting Research*, 38(1), 770-792.

Hoff, K.A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors*, 57(3), 407-434.

Honeycutt, D., Nourani, M., & Ragan, E. (2020). Soliciting Human-in-the-Loop User Feedback for Interactive Machine Learning Reduces User Trust and Impressions of Model Accuracy. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1), 63-72.

Hou, Y.T.Y., & Jung, M.F. (2021). Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. *Proceedings of the ACM on Human-Computer Interaction,* 5(CSCW2), 1-25.

Ireland, L. (2020). Who errs? Algorithm aversion, the source of judicial error, and public support for self-help behaviors. *Journal of Crime and Justice*, 43(2), 174-192.

Jung, M., & Seiter, M. (2021). Towards a better understanding on mitigating algorithm aversion in forecasting: an experimental study. *Journal of Management Control*, 32, 495-516.

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

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2), 263-291.

Kawaguchi, K. (2021). When Will Workers Follow an Algorithm? A Field Experiment with a Retail Business. *Management Science,* 67(3), 1670-1695.

Kizilcec, R.F. (2016). How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2390-2395.

Madhavan P., Wiegmann D.A., & Lacson F.C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors*, 48(2), 241-256.

Mahmud, H., Islam, A.N., Ahmed, S.I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, 175, 121390.

Manzey D., Reichenbach J., & Onnasch L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, 6, 57-87.

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

Méndez-Suárez, M., García-Fernández, F., & Gallardo, F. (2019). Artificial Intelligence Modelling Framework for Financial Automated Advising in the Copper Market. *Journal of Open Innovation: Technology, Market, and Complexity*, 5(4), 81.

Mohler, G.O., Short, M.B., Malinowski, S., Johnson, M.E., Tita, G.E., Bertozzi, A., & Brantingham, P.J. (2015). Randomized Controlled Field Trials of Predictive Policing. *Journal of the American Statistical Association*, 110(512), 1399-1411.

Niszczota, P., & Kaszás, D. (2020). Robo-investment aversion. *PLoS ONE*, 15(9), 0239277, 1-19.

Nolan, K.P., & Highhouse, S. (2014). Need for autonomy and resistance to standardized employee selection practices. *Human Performance*, 27(4), 328-346.

Ogburn, W.F. (1934). Studies in Prediction and the Distortion of Reality. *Social Forces*, 13, 224-229.

Önkal, D., Goodwin, P., Thomson, M.E., Gönül, S., & Pollock, A.C. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, 22(4), 390-409.

Pahlke, J., Strasser, S., & Vieider, F.M. (2012). Risk-taking for others under accountability. E*conomics Letters*, 114(1), 102-105.

Pérez-Toledano, M., Rodriguez, F.J., García-Rubio, J., & Ibáñez, S.J. (2019). Players' selection for basketball teams, through Performance Index Rating, using multiobjective evolutionary algorithms. *PLoS ONE,* 14(9), 1-20.

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

Proeger, T., & Meub, L. (2014). Overconfidence as a Social Bias: Experimental Evidence. *Economics Letters*, 122(2), 203-207.

Raafat, R.M., Chater, N., & Frith, C. (2009). Herding in humans. *Trends in Cognitive Sciences*, 13(10), 420-428.

Reich, T., Kaju, A., & Maglio, S.J. (2022). How to overcome algorithm aversion: Learning from mistakes. *Journal of Consumer Psychology,* ahead-of-print, 1-18.

Renier, L.A., Schmid Mast, M., & Bekbergenova, A. (2021). To err is human, not algorithmic – Robust reactions to erring algorithms. *Computers in Human Behavior*, 124, 106879.

Sele, D., & Chugunova, M. (2022). Putting a Human in the Loop: Increasing Uptake, but Decreasing Accuracy of Automated Decision-Making. *Max Planck Institute for Innovation & Competition Research Paper, No. 22-20*. Available at SSRN 4285645.

Simpson, B. (2016). Algorithms or advocacy: does the legal profession have a future in a digital world? *Information & Communications Technology Law*, 25(1), 50-61.

Spyrou, S.I. (2013). Herding in financial markets: a review of the literature. *Review of Behavioral Finance*, 5, 175-194.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Sciences*, 185(4157), 1124-1131.

Upadhyay, A.K., & Khandelwal, K. (2018). Applying artificial intelligence: implications for recruitment. *Strategic HR Review*, 17(5), 255-258.

Utz, S., Wolfers, L.N., & Göritz, A.S. (2021). The effects of situational and individual factors on algorithm acceptance in covid-19-related decision-making: A preregistered online experiment. *Human-Machine Communication*, 3, 27-46.

Vieider, F., Villegas-Palacio, C., Martinsson, P., & Mejía, M. (2016). Risk taking for oneself and others: A structural model approach. *Economic Inquiry*, 2016, 54(2), 879-894.

Wormith, J.S., & Goldstone, C.S. (1984). The Clinical and Statistical Prediction of Recidivism. *Criminal Justice and Behavior*, 11(1), 3-34.

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

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4), 1036-1040.

Chapter II

Reducing Algorithm Aversion through Experience

Co-authored by Ibrahim Filiz, Marco Lorenz, and Markus Spiwoks
Contribution Jan René Judek: 45%

Published:
Journal of Behavioral and Experimental Finance, Vol. 31, Issue 5, 100524, 1-8. (Sep 2021)
https://doi.org/10.1016/j.jbef.2021.100524

Also published as:

Discussion Paper in Sonderforschungsgruppe Institutionenanalyse (sofia), No. 21-1, Darmstadt, January 2021.

Working Paper in Wolfsburg Working Papers (WWP), Ostfalia University of Applied Sciences, No. 21-01, Wolfsburg, January 2021.

Abstract

In the context of an experiment, we examine the persistence of aversion towards algorithms in relation to learning processes. The subjects of the experiment are asked to make one share price forecast (rising or falling) in each of 40 rounds. A forecasting computer (algorithm) is available to them which has a success rate of 70%. Intuitive forecasts made by the subjects usually lead to a significantly poorer success rate. Feedback provided after each round of forecasts and a clear financial incentive led to the subjects becoming better able to estimate their own forecasting abilities. At the same time, their aversion to algorithms also decreases significantly.

Keywords

Algorithm aversion; Overconfidence; Operating experience; Stock market forecasting; Behavioral finance; Experiments.

JEL Classification

D83; D84; D91; G17; G41.

Highlights

- Subjects overestimate their own competence, which can lead to rejection of algorithms.
- Intuitive share price forecasts are clearly inferior to those of the algorithm.
- Over time, subjects begin to use the algorithm more frequently.
- Repeated tasks, constant feedback and financial incentives can reduce algorithm aversion.
- A learning process can significantly weaken a tendency towards algorithm aversion.

1 Introduction

Bank customers are becoming increasingly aware of charges, which is creating considerable cost pressure for banks. Particularly in the high-cost field of asset management, banks are endeavoring to reduce their personnel costs in relation to the provision of services to customers with low to medium amounts of assets. The substantial progress made in the field of artificial intelligence is increasingly leading banks to offer the services of so-called robo advisors which can provide customers with largely automated asset management (see, for example, Rühr et al., 2019; Jung et al., 2018; Singh & Kaur, 2017). There are some typical errors which are frequently made by professional investors as well as amateurs. For example, securities portfolios are often under-diversified (see, for example Dimmock et., 2016; Anderson, 2013; Hibbert, Lawrence & Prakash, 2012; Goetzmann & Kumar, 2008), or portfolios are restructured too frequently (see, for example, Barber & Odean, 2001; Barber & Odean, 2000). Many stock market players tend to see patterns in the trends of prices on the capital markets when in reality there are none (see, for example Zielonka, 2004; Wärneryd, 2001; Gilovich, Vallone & Tversky, 1985; Roberts, 1959). In this way, their gut feeling often entices them into making suboptimal investment decisions (see, for example, Frydman & Camerer, 2016; Kudryavtsev, Cohen & Hon-Snir, 2013). Problematic behavioral tendencies of this kind can easily be avoided with a suitably programmed robo advisor. An offer of reliable and cheap asset management which also has a favorable risk-return profile can thus be made to clients (see, for example, Rossi & Utkus, 2020; Bhatia, Chandani & Chhateja, 2020; D'Acunto, Prabhala & Rossi, 2019; Beketov, Lehmann & Wittke, 2018; Uhl & Rohner, 2018).

However, many people have reservations about automated processes. This frequently also applies even when it is clearly recognizable that an algorithm (such as that in a robo advisor) achieves better results than when an expert has taken on this task. This phenomenon is referred to as algorithm aversion (see, for example Erlei et al., 2020; Ku, 2020; Köbis & Mossink, 2020; Castelo, Bos & Lehmann, 2019; Dietvorst, Simmons & Massey, 2018; Prahl & Van Swol, 2017; Dietvorst, Simmons & Massey, 2015). This problem also occurs when subjects have to decide whether they trust themselves or an algorithm more (see, for example Efendić, Van de Calseyde & Evans, 2020; Rühr et al., 2019; Dietvorst, Simmons & Massey, 2018; Dietvorst, Simmons & Massey, 2015). Even when there are clear indications that it is hardly possible to make better decisions than the algorithm over the longer term, many subjects still tend to trust themselves more.

It seems reasonable to suppose that overestimation of one's own abilities plays a significant role here. Algorithm aversion and overconfidence are thus presumably closely related phenomena. However, there is an opportunity here. Proeger and Meub (2014) show that financial incentives, repeated feedback, and the gradual development of subjects' experience can help them to learn to assess their capabilities better. A learning process can thus lead to a reduction of overconfidence.

It thus seems feasible that algorithm aversion can also be decreased notably when decision-making situations repeat themselves, clear feedback is provided, and there are financial incentives. It is precisely this which is examined in this study on the basis of an experiment using repeated share price forecasts.

2 Experimental design and hypotheses

The subjects have the task of forecasting the price of a stock in 40 periods (see Appendix A). However, they do not have to predict the exact price, only whether it will rise or fall. The price is always moving up or down, so there is never an unchanged price. Either the price rises, or it falls. The price of the share is essentially determined by four fundamental influencing factors (A, B, C and D). However, these fundamental influencing factors are supplemented by a random influence ε (cf. Filiz, Nahmer & Spiwoks, 2019; Meub et al., 2015; Becker, Leitner & Leopold-Wildburger, 2009).

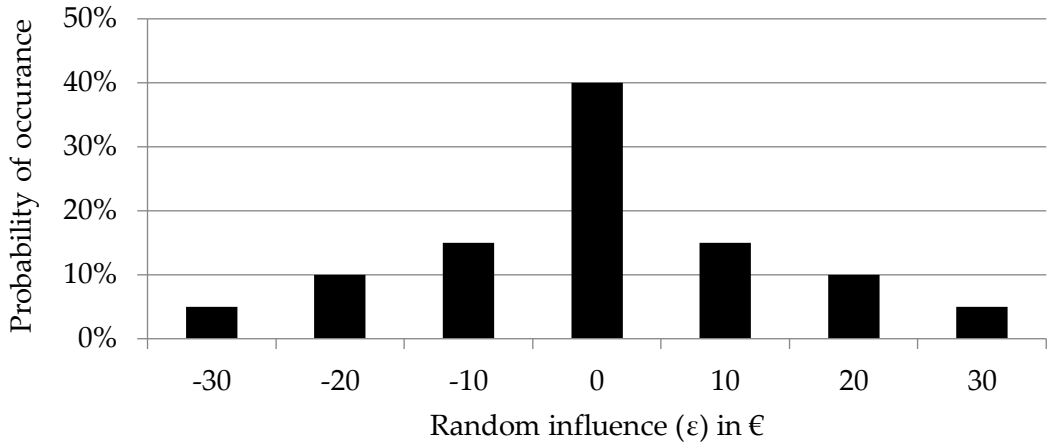
The price (K) of the share comes about as follows:

$$K_t = 32\ A_t + 1\ B_t - 18\ C_t + 44\ D_t + \varepsilon_t$$

The fundamental influencing factors (A, B, C and D) would, without the random influence ε_t, lead to a change in the price of between €0 and €10. If the fundamental influencing factors develop favorably overall, without the random influence ε_t there would always be a price increase between €0 and €10. This means that: €0 < (ΔK_t - ε_t) < +€10. However, if the fundamental influencing factors develop unfavorably overall, without the random influence ε_t there would always be a fall in the price of between €0 and €10. This means that: €0 > (ΔK_t - ε_t) > –€10.

The random influence ε_t has an expected value of 0 and exhibits the following distribution: with a probability of 40%, the random influence εt will not influence the price. With a probability of 15% each, the random influence ε_t will change the price by +€10 or by –€10. With a probability of 10% each, the random influence ε_t will change the price by +€20 or by –€20. And with a probability of 5% each, the random influence ε_t will change the price by +€30 or by –€30 (Figure 1).

Figure 1: Distribution of probability of the random influence ε_t



The fundamental influencing factors are announced to the subjects before each prediction round. In each round they have the opportunity to either make their own assessment (price rises or falls) or to delegate the decision to a forecasting computer (algorithm). In 70% of cases, the forecasting computer estimates the trend of the future share price correctly.

In other words, the algorithm merely exploits the available information about the fundamental influencing factors and the random influence ε_t in an optimal way. As the expected value of the random influence εt is zero, the algorithm calculates as follows: $K_t = 32\ A_t + 1\ B_t - 18\ C_t + 44\ D_t + 0$. Then it compares K_t with K_{t-1}. If $K_t > K_{t-1}$, the algorithm forecasts a rising trend. If $K_t < K_{t-1}$, the algorithm predicts a downward trend.

If the fundamental data suggests a rising trend (+€10 > ΔK_t > €0), this remains true in 70% of cases, also after the random influence ε_t is taken into consideration. A downward trend rather than a rising trend only transpires if the following random influences occur: ε_t = –€10 (15% probability) or ε_t = –€20 (10% probability) or ε_t = –€30 (5% probability). If the fundamental data suggests a downward trend (–€10 < ΔK_t < €0), this remains true in 70% of cases, also after the random influence ε_t is taken into consideration. An upward trend rather than a downward one only transpires when the random influences ε_t = +€10 (15% probability) or ε_t = +€20 (10% probability) or ε_t = +€30 (5% probability) occur.

The algorithm thus uses the existing information optimally, but its forecasts are by no means perfect. It is only right in 70% of cases. The phenomenon of algorithm aversion appears particularly in the case of algorithms which obviously do not function perfectly (see, for example, Dietvorst, Simmons & Massey, 2015).

The subjects are given an insight into 40 periods of historical prices of stock Z before they have to make their first decision (see Appendix C). In these 40 periods of price history, the price has risen exactly 20 times and has fallen exactly 20 times. This pattern remains in the subsequent 40 periods too: the price rises 20 times and falls 20 times. The subjects are not explicitly informed about this. However, by looking at the price history

they can obtain an impression of how the share price has risen just as frequently as it has fallen.

The subjects are aware of the mechanism behind how the price is formed ($K_t = 32\ A_t + 1\ B_t - 18\ C_t + 44\ D_t + \varepsilon_t$) and about the probability distribution of ε_t. In addition, the subjects are made expressly aware of the fact that the forecasting computer (algorithm) makes a correct prediction in 70% of cases. Test questions are used to ensure that the subjects have understood this point of departure (see Appendix B).

For a total of 40 times the subjects now have the choice whether to make their own forecast or to trust the algorithm. For every correct forecast they make or which they let the algorithm make for them (price rises or falls), the subjects receive a payment of 50 cents. They receive no payment if they or the algorithm make an incorrect forecast.

As the sequence of rising and falling price trends has no pattern which would enable a rational forecast (see Appendix D), the subjects have a choice between three strategies, although they in no way have to stick to just one of them. In each round of forecasts they have a free choice as to how they act. It is only in this way that we can observe possible learning effects. These three strategies are basically as follows:

(1) The subjects try to guess the trend of the price intuitively. In this case they would guess correctly in around 50% of cases. The expected value of their payment in this case is €10.

(2) The subjects use all of the information available to them and make forecasts in the same way as the algorithm would. To support them in this strategy they are given a pocket calculator, a pen and paper. In this case they will choose correctly in around 70% of cases. The expected value of their payment is €14.

(3) They delegate the forecasting to the algorithm. In this case they will make a correct forecast in around 70% of cases. The expected value of their payment is €14.

Subjects who act rationally and maximize their utility (*homo oeconomicus*) would have to choose the third strategy. The first strategy leads to a noticeable reduction in the expected value of their payment. The second strategy does not lead to a higher expected value of the payment than the third strategy, but due to the considerable calculation work required (an overall total of 160 multiplications with 320 factors plus 40 additions of 160 summands) it is prone to errors and arduous. A rational subject will therefore undoubtedly choose the third strategy.

However, it is well-known from earlier studies that in many subjects, looking at the price history of a stock triggers a strong feeling of intuition about its possible future trend (see, for example, Zielonka, 2004; Wärneryd, 2001; Roberts, 1959). We therefore presume that by no means all subjects will stay with the third strategy from the first round of the game to the last.

Hypothesis 1 is therefore: Some subjects will – at least sometimes – not choose the third strategy (delegation of forecasting to the algorithm).

Null hypothesis 1 is therefore: All subjects will choose the third strategy (delegation of forecasting to the algorithm) in all forty rounds of the game.

We presume that algorithm aversion and overconfidence are similar behavioral anomalies. The subjects will therefore frequently follow their own intuition instead of the algorithm (first strategy) because they overestimate their own forecasting ability. If one takes into account the results of the research by Proeger and Meub (2014), it can be presumed that the subjects will gradually learn to assess their own forecasting ability more realistically, because after each round of forecasting they are informed about how the price has changed (rising or falling), how successful they have been with their decisions (the current amount of their payment), and how successful they would have been if they had always delegated the forecasts to the algorithm (see Appendix C).

Hypothesis 2 is therefore: In the last 5 (10/15/20) rounds of forecasting, the subjects will trust the algorithm significantly more often than they did in the first 5 (10/15/20) rounds.

Null hypothesis 2 is therefore: In the last 5 (10/15/20) rounds of forecasting, the subjects will not choose the algorithm significantly more often than they did in the first 5 (10/15/20) rounds.

3 Results

The experiment is carried out between 2-14 November 2020 in the Ostfalia Laboratory of Experimental Economic Research (OLEW) of Ostfalia University of Applied Sciences in Wolfsburg. Overall, 143 subjects take part in the experiment. The subjects are students of Ostfalia University of Applied Sciences in Wolfsburg. 65 subjects (45.5%) study at the Faculty of Business, 60 subjects (42%) at the Faculty of Automotive Engineering, and 18 subjects (12.6%) at the Faculty of Public Health Services. 91 subjects (63.6%) are male, 50 subjects (35%) are female, and 2 subjects (1.4%) assign themselves to the category of third gender. The youngest subject is 18. The oldest subject is 35. The average age of the subjects is 23.5 years.

The experiment is programmed with z-Tree (cf. Fischbacher, 2007). In the Ostfalia Laboratory for Experimental Economic Research (OLEW), there are twelve computer workplaces. However, only a maximum of four are used per session. This ensures that a considerable distance can be maintained between the subjects. This is necessary due to the Covid-19 pandemic so that there is no danger to the health of the subjects. The workplaces in the laboratory are also equipped with divider panels, which makes it possible to completely separate the subjects from each other. The experiments are constantly monitored by the experimenter so that communication between the subjects and the use of prohibited aids (such as smartphones) can be ruled out. Overall, a total of 42 sessions are carried out. A session lasts an average of 45 minutes.

Is algorithm aversion exhibited in this experiment, or do all of the subjects consistently select the algorithm? A subject who is fully informed and always looking for their maximum utility (homo oeconomicus) would have to trust the algorithm in each of the 40 rounds of forecasting. This strategy leads to the maximum possible expected value in terms of the payment.

Overall, 143 subjects make 40 decisions each. This makes a total of 5,720 decisions. Of these, only 2,624 decisions (45.9%) are made in favor of the algorithm. In 3,096 decisions (54.1%), the subjects do not trust the algorithm. A clear majority of decisions are thus characterized by reservations in relation to the algorithm (Table 1).

Table 1: Decisions for and against the algorithm

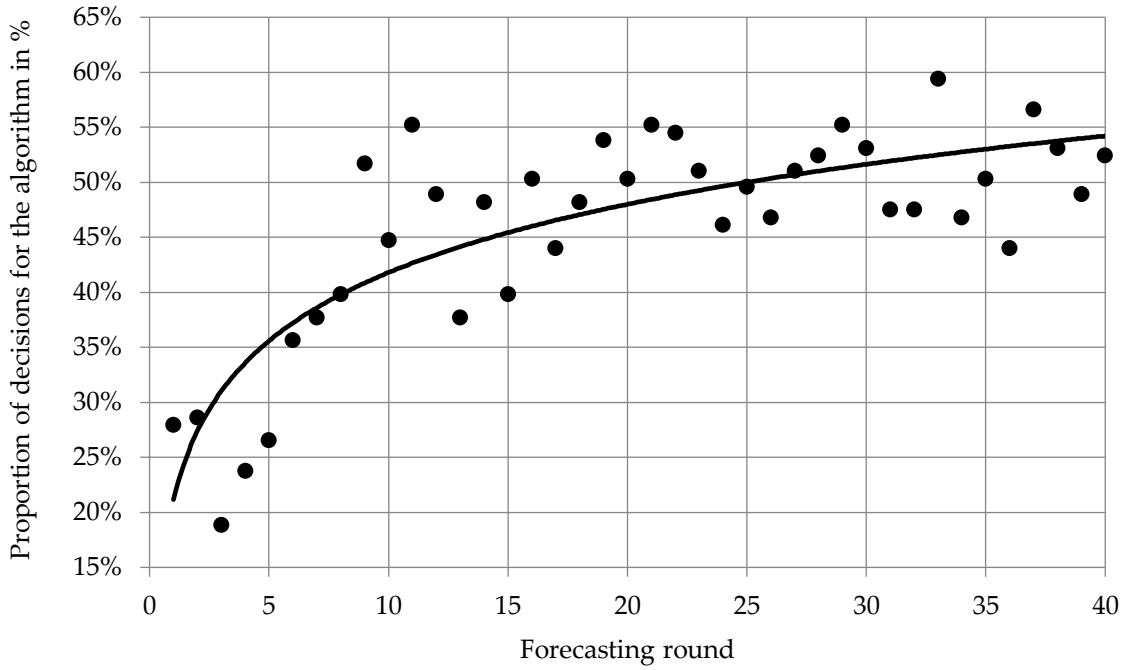
Decisions in favor of the algorithm		Decisions against the algorithm	
Number	%	Number	%
2,624	45.9%	3,096	54.1%

The t-test makes it clear that null hypothesis 1 must be rejected. The p-value ≤ 0.001 underlines the clarity of the results: the presumption is thus confirmed that a considerable amount of algorithm aversion would be revealed and that the subjects by no means always make rational decisions which maximize their utility.

Only very few subjects consistently pursue the strategy in which the fundamental data is used in order to determine the expected value of the next share price and to make a comparison with the last actual price. We can only observe this behavior in five subjects (3.5%). Their decisions against the algorithm must, however, be fully attributed to the phenomenon of algorithm aversion, because for a fully informed subject who wishes to maximize his or her utility, it is clearly recognizable that this strategy does not lead to a higher expected value of the payment. At the same time, it must be feared that errors can creep in given the multitude of calculations required (in 40 rounds a total of 160 multiplications with 320 factors and then 40 additions of 160 summands). That is why this tiresome mathematical recapitulation of the algorithm also reveals objectively unjustified reservations in relation to its reliability (Table 1).

Of particular interest is now whether the aversion to algorithms declines over time. Many subjects begin the experiment with an unjustified confidence that they can forecast the development of the share price (rising or falling) better than the algorithm. However, the sequence of rising and falling share prices is a random process with a probability of occurrence of 50% each for a rise or a fall of the price (see Appendix D). No information about the next movements of the stock can be derived from the price history. In this respect, intuitive decisions lead to a significant reduction in the expected payment in the medium to long-term.

Figure 2: Proportions of decisions in favor of the algorithm in % according to forecasting rounds



After each round of forecasts, the subjects are informed about the success of the algorithm and if applicable about the success of their own diverging forecast. As time passes, it thus becomes increasingly clear to the subjects that trusting their own intuition and not the algorithm is a sub-optimal strategy. As the experiment proceeds, part of the subjects gives up their reservations in relation to the algorithm (Figure 2 and Table 2). If one inserts a logarithmic regression line (Figure 2), the characteristics of a typical learning curve (see, for example Anzanello & Fogliatto, 2011; Wright, 1936) with declining learning progress can be recognized.

It can be seen that the percentage of decisions for the algorithm is initially quite low. On average in the first five rounds of forecasts, only around a quarter of the decisions of the subjects (25.2%) are for the algorithm, but then a swift learning process begins. Many subjects recognize that their intuition is not sufficiently reliable. On average in rounds 6-10 the percentage of decisions in favor of the algorithm already rises to 42%. On average in rounds 11-15 the percentage of decisions in favor of the algorithm continues to rise to 46%.

Table 2: Decisions for and against the algorithm according to forecasting rounds

Forecasting round	Decisions for the algorithm		Decisions against the algorithm	
	Number	Percent	Number	Percent
1	40	27.97%	103	72.03%
2	41	28.67%	102	71.33%
3	27	18.88%	116	81.12%
4	34	23.78%	109	76.22%
5	38	26.57%	105	73.43%
6	51	35.66%	92	64.34%
7	54	37.76%	89	62.24%
8	57	39.86%	86	60.14%
9	74	51.75%	69	48.25%
10	64	44.76%	79	55.24%
11	79	55.24%	64	44.76%
12	70	48.95%	73	51.05%
13	54	37.76%	89	62.24%
14	69	48.25%	74	51.75%
15	57	39.86%	86	60.14%
16	72	50.35%	71	49.65%
17	63	44.06%	80	55.94%
18	69	48.25%	74	51.75%
19	77	53.85%	66	46.15%
20	72	50.35%	71	49.65%
21	79	55.24%	64	44.76%
22	78	54.55%	65	45.45%
23	73	51.05%	70	48.95%
24	66	46.15%	77	53.85%
25	71	49.65%	72	50.35%
26	67	46.85%	76	53.15%
27	73	51.05%	70	48.95%
28	75	52.45%	68	47.55%
29	79	55.24%	64	44.76%
30	76	53.15%	67	46.85%
31	68	47.55%	75	52.45%
32	68	47.55%	75	52.45%
33	85	59.44%	58	40.56%
34	67	46.85%	76	53.15%
35	72	50.35%	71	49.65%
36	63	44.06%	80	55.94%
37	81	56.64%	62	43.36%
38	76	53.15%	67	46.85%
39	70	48.95%	73	51.05%
40	75	52.45%	68	47.55%

The learning process and the gradual fading away of algorithm aversion take place above all in the first 20 rounds of forecasting (Figure 3). In the final 20 rounds of forecasting, however, there is no longer a significant reduction of algorithm aversion (Figure 4). In the first five rounds of forecasting the algorithm is chosen 180 times (25.2%) and in the last five rounds 365 times (51.1%). In the first 10 rounds of forecasting the algorithm is chosen 480 times (33.6%) and in the last 10 rounds 725 times (50.7%). In the first 15 rounds of forecasting the algorithm is chosen 809 times (37.7%) and in the last 15 rounds 1,095 times (51.1%). In the first 20 rounds of forecasting the algorithm is chosen 1,162 times (40.6%) and in the last 20 rounds 1,462 times (51.1%).

Figure 3: Percentage of decisions in favor of the algorithm in the first 5, 10, 15 and 20 rounds of forecasting

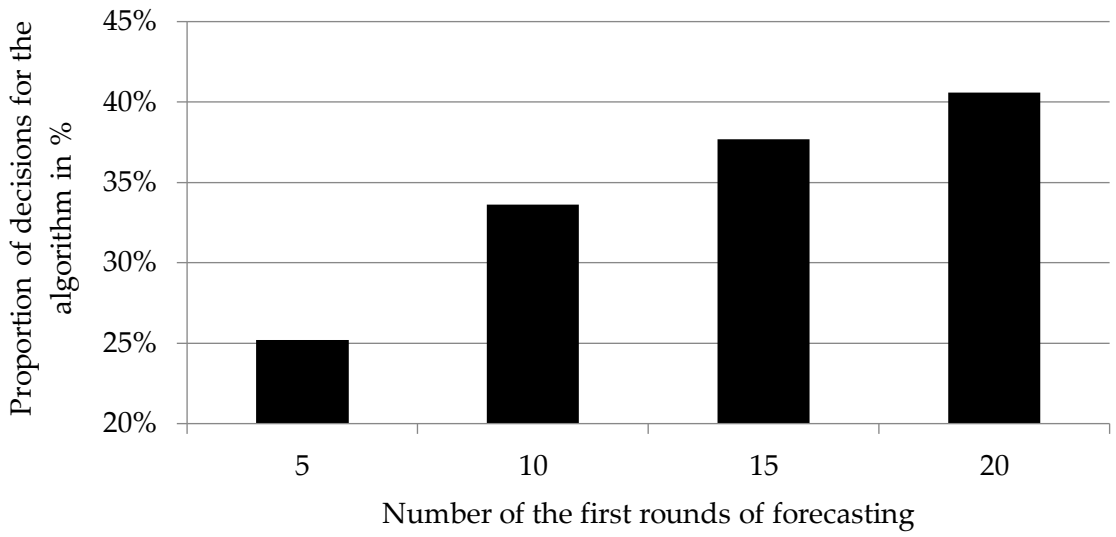
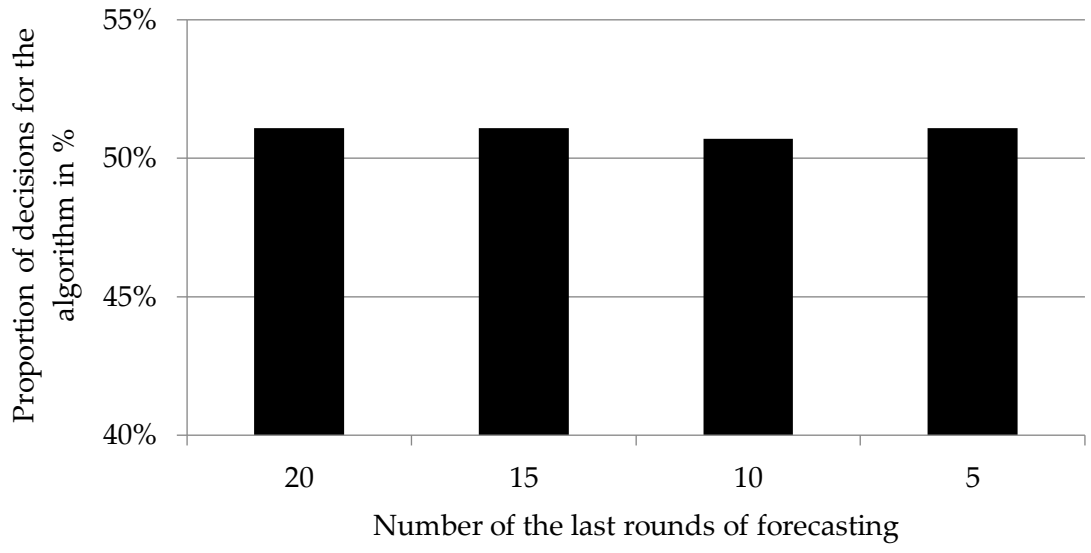


Figure 4: Percentage of decisions in favor of the algorithm in the last 20, 15, 10 and 5 rounds of forecasting



Hypothesis 2 can be checked with the help of a regression analysis. When carrying out a linear regression ($y_t = \beta_0 + \beta_1 \cdot x_t + u_t$) as well as when carrying out a logarithmic regression ($y_t = \beta_0 + \beta_1 \cdot \ln(x_t) + u_t$), it is clearly shown that algorithm aversion recedes over the course of 40 rounds of forecasting. The proportion of decisions in favor of the algorithm thus rises significantly. The p-values of the t-tests are unequivocal (Table 3). It is therefore clear that null hypothesis 2 must be rejected. Algorithm aversion is significantly reduced during a learning process with a declining course.

Table 3: Regression analysis of the increase in decisions in favor of the algorithm

Regression	Regression equation	β_1	t-value	p-value
Linear	$y_t = \beta_0 + \beta_1 \cdot x_t + u_t$	+0.57	5.92	0.000***
Logarithmic	$y_t = \beta_0 + \beta_1 \cdot \ln(x_t) + u_t$	+8.96	8.59	0.000***

*** = significant with an error probability of 1%, ** = significant with an error probability of 5%, * = significant with an error probability of 10%.

Another procedure for examining the significance of the learning process is the Wilcoxon signed-rank test. With the aid of this test, it can also be established whether the gradual increase in the number of decisions in favor of the algorithm is statistically significant (Table 4).

Here, we observe the number of subjects who follow the algorithm in the last 5 (10/15/20) rounds of forecasting more frequently (less frequently/unchanged) than in the first 5 (10/15/20) rounds of forecasting. It has hardly any influence on the results whether one compares the first 5 rounds of forecasting with the last 5 rounds of forecasting, or whether one compares the first 10 rounds of forecasting with the last 10 rounds of forecasting, or whether one compares the first 15 rounds of forecasting with the last 15 rounds of forecasting, or whether one compares the first 20 rounds of forecasting with the last 20 rounds of forecasting. In all four cases, it can be seen that a learning process sets in over the course of the 40 rounds of forecasting. The subjects learn to assess their forecasting abilities more realistically. Algorithm aversion declines notably. In the Wilcoxon signed-rank test, the results prove to be highly significant (Table 4). Null hypothesis 2 must be rejected. Experience with the advantages of algorithms can thus certainly lead to a reduction of algorithm aversion.

Table 4: Decision-making behavior in the first and last rounds of forecasting

Number (x) of forecasting rounds considered (first and last)	Subjects with fewer decisions in favor of the algorithm in the first x forecasting rounds than in the last x forecasting rounds	Subjects with more decisions in favor of the algorithm in the first x forecasting rounds than in the last x forecasting rounds	Subjects with the same number of decisions in favor of the algorithm in the first x forecasting rounds as in the last x forecasting rounds	Σ	p-value Wilcoxon signed-rank test
5	79	17	47	143	0.000***
10	80	21	42	143	0.000***
15	80	28	35	143	0.000***
20	81	30	32	143	0.000***

*** = significant with an error probability of 1%, ** = significant with an error probability of 5%,
* = significant with an error probability of 10%.

The effect sizes of the learning process can be described with either the Pearson correlation coefficient r (Fritz, Morris & Richler, 2012) or using Cohen's d (Cohen, 1992; Cohen, 1988). Pearson's correlation coefficient examines the strength of the correlation between two samples. Cohen's d, on the other hand, considers the expected values of two distributions – the further apart they are, the higher it is. In this way, the first 5 (10/15/20) forecasting rounds can be compared with the last 5 (10/15/20) forecasting rounds. Whereas Pearson's correlation coefficient r corresponds to strong effects according to its categorization by Cohen (1992), Cohen's d shows the average effect sizes of the learning process (Table 5).

Table 5: Effect sizes of the learning processes according to Pearson's r and Cohen's d

Comparison	Pearson's correlation coefficient r	Cohen's d
First 5 forecasting rounds compared to the last 5 rounds	0.57	0.73
First 10 forecasting rounds compared to the last 10 rounds	0.54	0.65
First 15 forecasting rounds compared to the last 15 rounds	0.50	0.58
First 20 forecasting rounds compared to the last 20 rounds	0.49	0.56

However, it is also shown that their experiences only convince a part of the subjects to give up their aversion to algorithms. Even at the end of the 40 rounds of forecasting, just under half of the subjects still show no desire to use the algorithm. On average in rounds 36-40, just below 49% of decisions made are still against the algorithm. At this point in time, the subjects must have realized that their intuitive share price forecasts are far inferior to those of the algorithm, but they decline to use it, nevertheless.

As a phenomenon, algorithm aversion has a certain similarity to overconfidence. Learning effects lead to a more realistic estimation of the subject's abilities and thus to a decrease in algorithm aversion. However, the phenomenon of algorithm aversion obviously contains additional aspects which cannot be rectified by the gradual recognition of the superior performance of an algorithm. Among many subjects, their reservations towards the algorithm remain even when they have learned through their own experience that foregoing the algorithm is not in their financial interests.

4 Summary

We experimentally examine the persistence of aversion towards algorithms in relation to learning processes. When subjects have to decide whether they should let an algorithm do a task for them or whether they would rather do it themselves, a possible overestimation of their own competence can lead to rejection of the algorithm. Overconfidence can, however, be tempered by a learning process. Repeated tasks, constant feedback and financial incentives can contribute towards subjects gradually learning to better estimate their own abilities. We are interested in the question of whether such learning processes can also contribute to a reduction of algorithm aversion.

In the experiment, the subjects are asked to make share price forecasts (the price will rise or the price will fall). In 40 rounds of forecasting they can either trust their own assessment or put their faith in a forecasting computer (algorithm). Intuitive forecasts usually turn out to be less successful than the algorithm. The payments made to the subjects depend on the success of their forecasts – regardless of whether the forecasts are their own or whether they are made by the algorithm. After each round, the forecasting results are presented. The subjects can see how much payment they have received and how much they would have obtained if they had trusted the algorithm from the very beginning.

Many subjects recognize as early as the first ten rounds of forecasts that their intuitive share price forecasts are clearly inferior to those of the algorithm. They exhibit an increasing readiness to trust in the algorithm. Regression analysis and the Wilcoxon signed-rank test both show that a learning process can significantly weaken a tendency towards algorithm aversion. With the aid of Pearson's r and Cohen's d, it can be seen that the learning process exhibits a moderate effect size. However, it is also shown that in a considerable part of the subjects there is no weakening of algorithm aversion even over 40 rounds of forecasts.

References

Anderson, A. (2013), Trading and Under-Diversification, *Review of Finance*, 17(5), 1699-1741.

Anzanello, M. J., & Fogliatto, F. S. (2011), Learning curve models and applications: Literature review and research directions, *International Journal of Industrial Ergonomics*, 41(5), 573-583.

Barber, B. M., & Odean, T. (2001), Boys Will Be Boys: Gender, Overconfidence, and Common Stock Investment, *Quarterly Journal of Economics*, 116(1), 261-292.

Barber, B. M., & Odean, T. (2000), Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors, *The Journal of Finance*, 55(2), 773-806.

Becker, O., Leitner, J., & Leopold-Wildburger, U. (2009), Expectation formation and regime switches, *Experimental Economics*, 12(3), 350-364.

Beketov, M., Lehmann, K., & Wittke, M. (2018), Robo Advisors: quantitative methods inside the robots, *Journal of Asset Management*, 19, 363-370.

Bhatia, A., Chandani, A., & Chhateja, J. (2020), Robo advisory and its potential in addressing the behavioral biases of investors — A qualitative study in Indian context, *Journal of Behavioral and Experimental Finance*, 25.

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

Cohen, J. (1988), Statistical power analysis for the behavioral sciences (2nd ed), Hillsdale, N.J., L. Erlbaum Associates.

Cohen, J. (1992), A power primer, *Psychological Bulletin*, 112(1), 155-159.

D'Acunto, F., Prabhala, N., & Rossi, A. G. (2019), The Promises and Pitfalls of Robo-Advising, *The Review of Financial Studies*, 32(5), 1983-2020.

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.

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), 114-126.

Dimmock, S. G., Kouwenberg, R., Mitchell, O. S., & Peijnenburg, K. (2016), Ambiguity Aversion and Household Portfolio Choice Puzzles: Empirical Evidence, *Journal of Financial Economics*, 119, 559-577.

Efendić, E., Van de Calseyde, P. P., & Evans, A. M. (2020), Slow response times undermine trust in algorithmic (but not human) predictions, *Organizational Behavior and Human Decision Processes*, 157, 103-114.

Erlei, A., Nekdem, F., Meub, L., Anand, A., & Gadiraju, U. (2020), Impact of Algorithmic Decision Making on Human Behavior: Evidence from Ultimatum Bargaining, *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1), 43-52.

Filiz, I., Nahmer, T., & Spiwoks, M. (2019), Herd behavior and mood: An experimental study on the forecasting of share prices, *Journal of Behavioral and Experimental Finance*, 24, 1-10.

Fischbacher, U. (2007), z-Tree: Zurich Toolbox for Ready-made Economic Experiments, *Experimental Economics*, 10(2), 171-178.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012), Effect Size Estimates: Current Use, Calculations, and Interpretation, *Journal of Experimental Psychology: General*, 141(1), 2-18.

Frydman, C., & Camerer, C. F. (2016), The Psychology and Neuroscience of Financial Decision Making, *Trends in Cognitive Sciences*, 20(9), 661-675.

Gilovich, T., Vallone, R., & Tversky, A. (1985), The hot hand in basketball: On the misperception of random sequences, *Cognitive Psychology*, 17(3), 295-314.

Goetzmann, W. N., & Kumar, A. (2008), Equity Portfolio Diversification, *Review of Finance*, 12(3), 433-463.

Hibbert, A. M., Lawrence, E. R., & Prakash, A. J. (2012), Can Diversification Be Learned?, *The Journal of Behavioral Finance*, 13(1), 38-50.

Jung, D., Dorner, V., Glaser, F., & Morana, S. (2018), Robo-Advisory - Digitalization and Automation of Financial Advisory, *Business & Information Systems Engineering*, 60(1), 81-86.

Köbis, N., & Mossink, L. D. (2020), Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry, *Computers in Human Behavior*, 114, 1-13.

Ku, C. Y. (2020), When AIs Say Yes and I Say No: On the Tension between AI's Decision and Human's Decision from the Epistemological Perspectives, *Információs Társadalom*, 19(4), 61-76.

Kudryavtsev, A., Cohen, G., & Hon-Snir, S. (2013), "Rational" or "Intuitive": Are Behavioral Biases Correlated Across Stock Market Investors?, *Contemporary Economics*, 7(2), 31-53.

Meub, L., Proeger, T., Bizer, K., & Spiwoks, M. (2015), Strategic coordination in forecasting - An experimental study, *Finance Research Letters*, 13(1), 155-162.

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

Proeger, T., & Meub, L. (2014), Overconfidence as a Social Bias: Experimental Evidence, *Economics Letters*, 122(2), 203-207.

Roberts, H. V. (1959), Stock market ''patterns'' and financial analysis: Methodological suggestions, *Journal of Finance*, 1(14), 1-10.

Rossi, A. G., & Utkus, S. P. (2020), Who Benefits from Robo-advising? Evidence from Machine Learning, *SSRN*, 3552671, doi.org/10.2139/ssrn.3552671.

Rühr, A., Streich, D., Berger, B., & Hess, T. (2019), A Classification of Decision Automation and Delegation in Digital Investment Systems, *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 1435-1444.

Singh, I., & Kaur, N. (2017), Wealth Management Through Robo Advisory, *International Journal of Research - Granthaalayah*, 5(6), 33-43.

Uhl, M. W., & Rohner, P. (2018), Robo-advisors versus traditional investment advisors: An unequal game, *The Journal of Wealth Management*, 21(1), 44-50.

Wärneryd, K.-E. (2001), Stock-Market Psychology, Cheltenham: Edward Elgar, Cheltenham.

Wright, T. P. (1936), Factors affecting the cost of airplanes, *Journal of the Aeronautical Sciences*, 3(4), 122-128.

Zielonka, P. (2004), Technical analysis as the representation of typical cognitive biases, *International Review of Financial Analysis*, 13, 217-225.

Appendix A: Instructions for the game

The Game

In this game you are requested to make forecasts on the future trend of a share price. You will forecast the price movements of a share (share Z) in 40 periods. However, you do not predict the exact price of the stock, you only forecast whether it will rise or fall. The price of share Z is always moving, it never remains unchanged. It rises or it falls.

The price of share Z in € at the point in time t (K_t) is always determined by four fundamental influencing factors (A_t, B_t, C_t and D_t) and a random influence (ε_t). The fundamental influencing factors are announced before every round of forecasting. The subjects are also aware of the specific influence the fundamental data has on the share price.

The price (K_t) of share Z is formed as follows:

$$K_t = 32 \cdot A_t + 1 \cdot B_t - 18 \cdot C_t + 44 \cdot D_t + \varepsilon_t$$

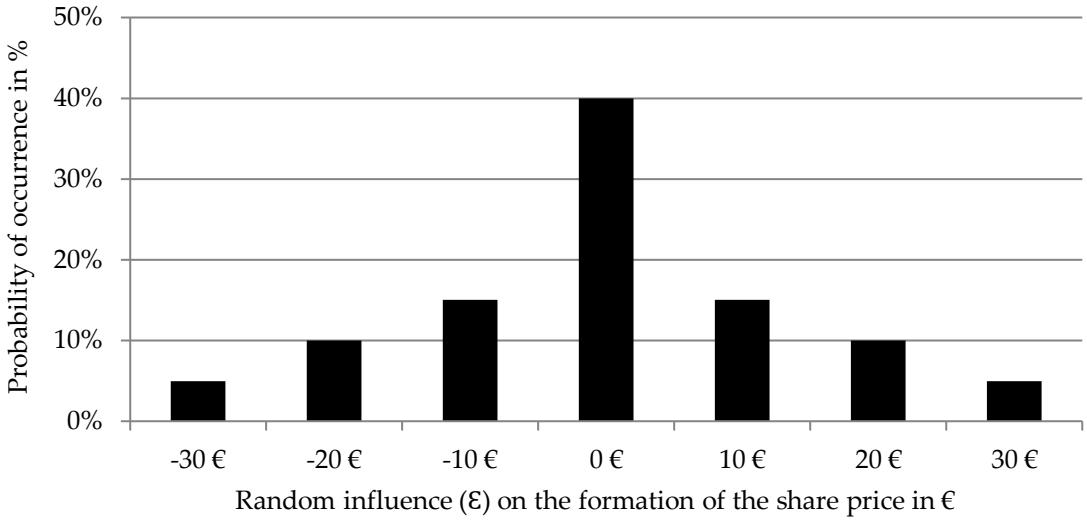
The fundamental influencing factors (A_t, B_t, C_t and D_t) would, without the random influence ε_t, lead to a price change of between –€10 and €0 or between €0 and +€10 in every period.

If the fundamental influencing factors generally develop favorably, without the random influence ε_t there would always be a rise of the share price between €0 and €10. This means: $€0 < (\Delta K_t - \varepsilon_t) < +€10$. If, however, the fundamental influencing factors generally develop unfavorably, without the random influence ε_t there would always be a fall in the share price of between €0 and €10. This means: $€0 > (\Delta K_t - \varepsilon_t) > -€10$.

The random influence ε_t has an expected value of 0 and is distributed as follows: with a probability of 40%, the random influence ε_t is equal to zero ($\varepsilon_t = 0$). With a probability of 15% each, the random influence ε_t obtains a value of –€10 or +€10. With a probability of 10% each, the random influence ε_t obtains a value of –€20 or +€20. With a probability of 5% each, the random influence ε_t obtains a value of –€30 or +€30 (Figure 1).

In each round of forecasting you have the opportunity to make your own assessment (the price rises or the price falls), or to delegate the decision to a forecasting computer. In 70% of cases, the forecasting computer estimates the future price of stock Z correctly.

Figure 1: Distribution probability of the random influence ε_t



Procedure

After reading the instructions and answering the test questions, you see the history of stock Z over the last 40 periods as well as a detailed chart of the price of Z during the last 10 periods. In addition, you will receive the figures of the fundamental data for the next period. You will be asked to forecast the trend of the share price in the next period. After making your forecast you will see what actually happens to the price of the share Z in the next period and receive the results of your prediction. A total of 40 rounds are played. Before every round you see the course of Z from period 1 to the current period as well as a detailed chart of the price of Z during the last 10 periods. In addition, you will receive the figures of the fundamental data for the next period.

Payment

For every successful share price forecast you receive €0.50. A forecast is considered successful and is rewarded accordingly when it correctly predicts the actual direction of the share price. In total you can earn up to €20. Payment is made at the end of the experiment.

Information

- Please remain quiet during the experiment
- Please do not look at your neighbor's screen
- Apart from a pen and a pocket calculator, no aids are permitted (smartphones, smart watches etc.)

Appendix B: Test questions

Test question 1: Which alternatives do you have when making your forecasts?

a) I can only use my own forecast.
b) I can either follow the algorithm or make my own forecast. *(correct*)
c) I can either follow the algorithm, make my own forecast, or ask other people in the room.

Test question 2: What is the success rate of the algorithm?

a) 40%
b) 50%
c) 70% *(correct)*

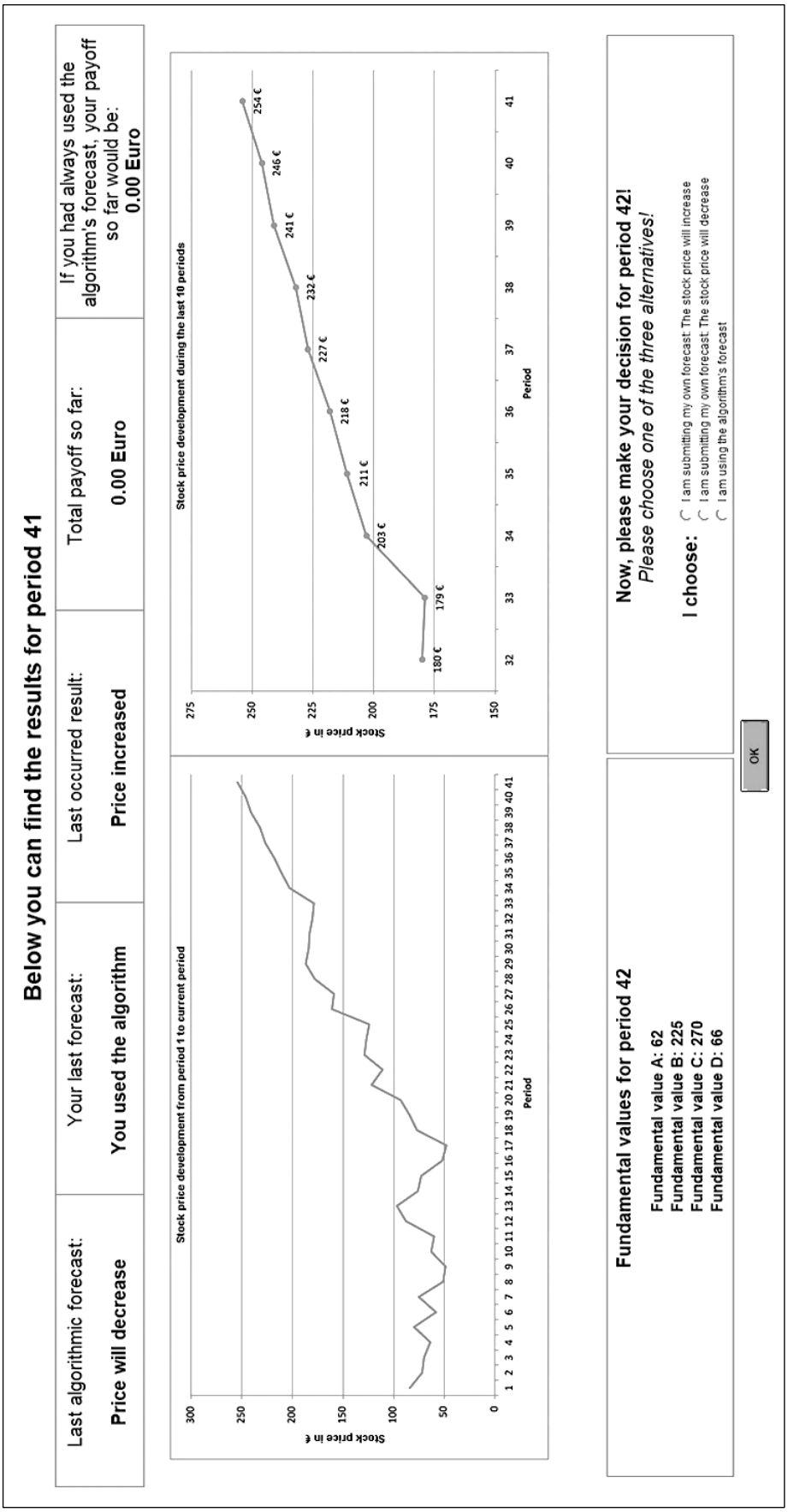
Test question 3: How much is the payment for a successful forecast?

a) €0.00
b) €0.50 *(correct)*
c) €1.00

Test question 4: How much is the payment for an unsuccessful forecast?

a) €0.50
b) €1.00
c) €0.00 *(correct)*

Appendix C: Subject's Screen



Appendix D: Variations in the price movements

In order to prevent distortion of the results due to conferring between subjects who take part in the experiment at different points in time, four different variants of price movements were used in the experiment.

Forecasting round	Variant A	Variant B	Variant C	Variant D
1	267	264	254	255
2	275	273	273	284
3	284	282	277	292
4	272	280	265	289
5	299	278	252	307
6	296	316	261	302
7	294	315	256	310
8	313	333	289	299
9	311	330	287	328
10	349	356	283	326
11	357	364	252	335
12	376	381	280	304
13	364	388	288	283
14	353	377	307	321
15	351	376	306	316
16	389	410	304	305
17	398	408	281	302
18	396	406	278	300
19	405	395	285	308
20	413	414	314	337
21	381	432	323	346
22	410	421	322	335
23	428	408	330	344
24	436	417	308	332
25	425	426	336	310
26	423	434	335	308
27	422	423	334	327
28	421	421	332	325
29	430	418	339	333
30	438	427	358	341
31	446	444	323	349
32	454	431	331	325
33	463	398	310	304
34	461	392	313	312
35	450	400	322	311
36	448	408	311	320
37	427	384	308	317
38	408	392	326	314
39	387	370	324	343
40	366	355	333	349

Chapter III

The Extent of Algorithm Aversion in Decision-making Situations with Varying Gravity

Co-authored by Ibrahim Filiz, Marco Lorenz, and Markus Spiwoks
Contribution Jan René Judek: 45%

Published:
PLoS ONE, Vol. 18, Issue 2, e0278751, 1-21. (Feb 2023)
https://doi.org/10.1371/journal.pone.0278751

Also published as:

Discussion Paper in Sonderforschungsgruppe Institutionenanalyse (sofia), No. 21-2, Darmstadt, January 2021.

Working Paper in Wolfsburg Working Papers (WWP), Ostfalia University of Applied Sciences, No. 21-02, Wolfsburg, February 2021.

Abstract

Algorithms already carry out many tasks more reliably than human experts. Nevertheless, some subjects have an aversion towards algorithms. In some decision-making situations an error can have serious consequences, in others not. In the context of a framing experiment, we examine the connection between the consequences of a decision-making situation and the frequency of algorithm aversion. This shows that the more serious the consequences of a decision are, the more frequently algorithm aversion occurs. Particularly in the case of very important decisions, algorithm aversion thus leads to a reduction of the probability of success. This can be described as the tragedy of algorithm aversion.

Keywords

Algorithm aversion; technology adoption; framing; behavioral economics; experiments.

JEL Classification

D81; D91; G41; O33.

1 Introduction

Automated decision-making or decision aids, so-called algorithms, are becoming increasingly significant for many people's working and private lives. The progress of digitalization and the growing significance of artificial intelligence in particular mean that efficient algorithms have now already been available for decades (see, for example, Dawes, Faust & Meehl, 1989). These algorithms already carry out many tasks more reliably than human experts (Grove et al., 2000). However, only a few algorithms are completely free of errors. Some areas of application of algorithms can have serious consequences in the case of a mistake – such as autonomous driving (cf. Shariff, Bonnefon & Rahwan, 2017), making medical diagnoses (cf. Majumdar & Ward, 2011), or support in criminal proceedings (cf. Simpson, 2016). On the other hand, algorithms are also used for tasks which might not have serious consequences in the case of an error, such as dating service (cf. Brozovsky & Petříček, 2007), weather forecasts (cf. Sawaitul, Wagh & Chatur, 2012), and the recommendation of cooking recipes (cf. Ueda, Takahata & Nakajima, 2011).

Some economic agents have a negative attitude towards algorithms. This is usually referred to as algorithm aversion (for an overview of algorithm aversion see Mahmud et al., 2022; Burton, Stein & Jensen, 2020). Many decision-makers thus tend to delegate tasks to human experts or carry them out themselves. This is also frequently the case when it is clearly recognizable that using algorithms would lead to an increase in the quality of the results (Castelo, Bos & Lehmann, 2020; Dietvorst, Simmons & Massey, 2015; Youyou, Kosinski & Stillwell, 2015).

In decision-making situations which lead to consequences which are not so serious in the case of an error, a behavioral anomaly of this kind does not have particularly significant effects. In the case of a dating service, the worst that can happen is meeting with an unsuitable candidate. In the case of an erroneous weather forecast, unless it is one for seafarers, the worst that can happen is that unsuitable clothing is worn, and if the subject is the recommendation of cooking recipes, the worst-case scenario is a bland meal. However, particularly in the case of decisions which can have serious consequences in the case of a mistake, diverging from the rational strategy would be highly risky. For example, a car crash or a wrong medical diagnosis can, in the worst case, result in someone's death. Being convicted in a criminal case can lead to many years of imprisonment. In these serious cases, it can be expected that people tend to think more thoroughly about what to do in order to make a reasonable decision. Can algorithm aversion be overcome in serious situations in order to make a decision which maximizes utility and which, at best, can save a life?

Tversky and Kahneman (1981) show that decisions can be significantly influenced by the context of the decision-making situation. The story chosen to illustrate the problem influences the salience of the information, which can also lead to an irrational neglect of the underlying mathematical facts. This phenomenon is also referred to as the framing effect (for an overview see Cornelissen & Werner, 2014). Irrespective of the actual probability of success, subjects do allow themselves to be influenced. This study therefore

uses six mathematically identical decision situations with different contexts to examine whether the extent of algorithm aversion can be influenced by a framing effect.

Moreover, it is analyzed precisely which aspects of a decision affect the choice between algorithms and human experts the most. In particular, it is examined whether subjects are prepared to desist from their algorithm aversion in decision-making situations which can have severe consequences (three of the six scenarios). Expectancy theory (Vroom, 1964) states that the importance of a task positively influences subjects' motivation in performing the task. Consistent with this, Mento, Cartledge and Locke (1980) show in five experiments that increasing valence of a goal leads to higher goal acceptance and determination to achieve it. Gollwitzer (1993) argues that the importance of a task determines the extent to which individuals develop and maintain commitment to the task. Similarly, Gendolla (1997) asserts that "outcome valence and importance have effects on expectancy formation," where importance refers to the "personal relevance of events."

If algorithm aversion is due to decisions being made on gut instinct rather than analytically thought through, it should decrease with more meticulous expectancy formation, and increasing motivation and commitment, all of which result from task importance. We thus consider whether there are significantly different frequencies of algorithm aversion depending on whether the decision-making situations can have serious consequences or not.

2 Literature Review

Previous publications have defined the term algorithm aversion in quite different ways (Table 1). These different understandings of the term are reflected in the arguments put forward as well as in the design of the experiments carried out. From the perspective of some scholars, it is only possible to speak of algorithm aversion when an algorithm recognizably provides the option with the highest quality result or probability of success (cf. Köbis & Mossink, 2021; Burton, Stein & Jensen, 2020; Ku, 2020; Castelo, Bos & Lehmann, 2020; Dietvorst, Simmons & Massey, 2015). However, other scholars consider algorithm aversion to be present as soon as subjects exhibit a fundamental disapproval of an algorithm in spite of its possible superiority (cf. Efendić, Van de Calseyde & Evans, 2020; Niszczota & Kaszás, 2020; Horne et al., 2019; Logg, Minson & Moore, 2019; Rühr et al., 2019; Yeomans et al., 2019; Prahl & Van Swol, 2017).

Another important aspect of how the term algorithm aversion is understood is the question of whether and possibly also how the subjects learn about the superiority of an algorithm. Differing approaches were chosen in previous studies. Dietvorst, Simmons and Massey (2015) focus on the gathering of experience in dealing with an algorithm in order to be able to assess its probability of success in comparison to one's own performance. In a later study, Dietvorst, Simmons and Massey (2018) specify the average error of an algorithm. Alexander, Blinder and Zak (2018) provide exact details on the probability of success of an algorithm, or they refer to the rate at which other subjects used an algorithm in the past.

Table 1: Definitions of algorithm aversion in the literature

Authors	Definition of algorithm aversion
Dietvorst, Simmons & Massey, 2015	"Research shows that evidence-based algorithms more accurately predict the future than do human forecasters. Yet when forecasters are deciding whether to use a human forecaster or a statistical algorithm, they often choose the human forecaster. This phenomenon, which we call *algorithm aversion (…)*"
Prahl & Van Swol, 2017	"The irrational discounting of automation advice has long been known and a source of the spirited "clinical versus actuarial" debate in clinical psychology research (Dawes, 1979; Meehl, 1954). Recently, this effect has been noted in forecasting research (Önkal et al., 2009) and has been called algorithm aversion (Dietvorst, Simmons, & Massey, 2015)."
Dietvorst, Simmons & Massey, 2018	"Although evidence-based algorithms consistently outperform human forecasters, people often fail to use them after learning that they are imperfect, a phenomenon known as *algorithm aversion*."
Commerford, Dennis, Joe & Wang, 2019	"(…) *algorithm aversion* – the tendency for individuals to discount computer-based advice more heavily than human advice, although the advice is identical otherwise."
Horne, Nevo, O'Donovan, Cho & Adali, 2019	"For example, Dietvorst et al. (Dietvorst, Simmons, and Massey, 2015) studied when humans choose the human forecaster over a statistical algorithm. The authors found that aversion of the automated tool increased as humans saw the algorithm perform, even if that algorithm had been shown to perform significantly better than the human. Dietvorst et al. explained that aversion occurs due to a quicker decrease in confidence in algorithmic forecasters over human forecasters when seeing the same mistake occur (Dietvorst, Simmons, and Massey, 2015)."
Ku, 2019	"(…) "algorithm aversion", a term refers by Dietvorst et al. (Dietvorst et al. 2015) means that humans distrust algorithm even though algorithm consistently outperform humans."
Leyer & Schneider, 2019	"In the particular context of the delegation of decisions to AI-enabled systems, recent findings have revealed a general algorithmic aversion, an irrational discounting of such systems as suitable decision-makers despite objective evidence (Dietvorst, Simmons and Massey, 2018)"
Logg, Minson & Moore, 2019	"(…) human distrust of algorithmic output, sometimes referred to as "algorithm aversion" (Dietvorst, Simmons, & Massey, 2015).[1] "; Footnote 1: "while this influential paper [of Dietvorst et al.] is about the effect that seeing an algorithm err has on people's likelihood of choosing it, it has been cited as being about how often people use algorithms in general."
Önkal, Gönül & De Baets, 2019	"(…) people are averse to using advice from algorithms and are unforgiving toward any errors made by the algorithm (Dietvorst et al., 2015; Prahl & Van Swol, 2017)."
Rühr, Streich, Berger & Hess, 2019	"Users have been shown to display an aversion to algorithmic decision systems [Dietvorst, Simmons, Massey, 2015] as well as to the perceived loss of control associated with excessive delegation of decision authority [Dietvorst, Simmons, Massey, 2018]."
Yeomans, Shah, Mullainathan & Kleinberg, 2019	"(...) people would rather receive recommendations from a human than from a recommender system (...). This echoes decades of research showing that people are averse to relying on algorithms, in which the primary driver of aversion is algorithmic errors (for a review, see Dietvorst, Simmons, & Massey, 2015)."
Berger, Adam, Rühr & Benlian, 2020	"Yet, previous research indicates that people often prefer human support to support by an IT system, even if the latter provides superior performance – a phenomenon called algorithm aversion." (…) "These differences result in two varying understandings of what algorithm aversion is: unwillingness to rely on an algorithm that a user has experienced to err versus general resistance to algorithmic judgment."
Burton, Stein & Jensen, 2020	"(…) algorithm aversion—the reluctance of human forecasters to use superior but imperfect algorithms— (…)"

Castelo, Bos & Lehmann, 2020	"The rise of algorithms means that consumers are increasingly presented with a novel choice: should they rely more on humans or on algorithms? Research suggests that the default option in this choice is to rely on humans, even when doing so results in objectively worse outcomes."
De-Arteaga, Fogliato & Chouldechova, 2020	"*Algorithm aversion*–the tendency to ignore tool recommendations after seeing that they can be erroneous (…)"
Efendić, Van de Calseyde & Evans, 2020	"Algorithms consistently perform well on various prediction tasks, but people often mistrust their advice."; "However, repeated observations show that people profoundly mistrust algorithm-generated advice, especially after seeing the algorithm fail (Bigman & Gray, 2018; Diab, Pui, Yankelevich, & Highhouse, 2011; Dietvorst, Simmons, & Massey, 2015; Önkal, Goodwin, Thomson, Gönül, & Pollock, 2009)."
Erlei, Nekdem, Meub, Anand & Gadiraju, 2020	"Recently, the concept of algorithm aversion has raised a lot of interest (see (Burton, Stein, and Jensen 2020) for a review). In their seminal paper, (Dietvorst, Simmons, and Massey 2015) illustrate that human actors learn differently from observing mistakes by an algorithm in comparison to mistakes by humans. In particular, even participants who directly observed an algorithm outperform a human were less likely to use the model after observing its imperfections."
Germann & Merkle, 2020	"The tendency of humans to shy away from using algorithms even when algorithms observably outperform their human counterpart has been referred to as algorithm aversion."
Ireland, 2020	"(…) some researchers find that, when compared to humans, people are averse to algorithms after recording equivalent errors."
Jussupow, Benbasat & Heinzl, 2020	"(…) literature suggests that although algorithms are often superior in performance, users are reluctant to interact with algorithms instead of human agents – a phenomenon known as algorithm aversion"
Niszczota & Kaszás, 2020	"When given the possibility to choose between advice provided by a human or an algorithm, people show a preference for the former and thus exhibit algorithm aversion (Castelo et al., 2019; Dietvorst et al., 2015, 2016; Longoni et al., 2019)."
Wang, Harper & Zhu, 2020	"(…) people tend to trust humans more than algorithms even when the algorithm makes more accurate predictions."
Kawaguchi, 2021	"The phenomenon in which people often obey inferior human decisions, even if they understand that algorithmic decisions outperform them, is widely observed. This is known as algorithm aversion (Dietvorst et al. 2015)."
Köbis & Mossink, 2021	"When people are informed about algorithmic presence, extensive research reveals that people are generally averse towards algorithmic decision makers. This reluctance of "human decision makers to use superior but imperfect algorithms" (Burton, Stein, & Jensen, 2019; p.1) has been referred to as algorithm aversion (Dietvorst, Simmons, & Massey, 2015). In part driven by the belief that human errors are random, while algorithmic errors are systematic (Highhouse, 2008), people have shown resistance towards algorithms in various domains (see for a systematic literature review, Burton et al., 2019)."

In addition, when dealing with algorithms, the way in which people receive feedback is of significance. Can subjects (by using their previous decisions) draw conclusions about the quality and/or success of an algorithm? Dietvorst, Simmons and Massey (2015) merely use feedback in order to facilitate experience in dealing with an algorithm. Prahl and Van Swol (2017) provide feedback after every individual decision, enabling an assessment of the success of the algorithm. Filiz et al. (2021) follow this approach and use feedback after every single decision in order to examine the decrease in algorithm aversion over time.

Other aspects which emerge from the previous definitions of algorithm aversion in the literature are the reliability of an algorithm (perfect or imperfect), the observation of its reliability (the visible occurrence of errors), access to historical data on how the algorithmic forecast was drawn up; the setting (algorithm vs. expert; algorithm vs. amateur; algorithm vs. subject) as well as extent of the algorithm's intervention (does the algorithm act as an aid to decision-making or does it carry out tasks automatically?).

In our view, the superiority of an algorithm (higher probability of success) and the knowledge of this superiority are the decisive aspects. Algorithm aversion can only be present when subjects are clearly aware that not using an algorithm reduces the expected value of their utility and they do not deploy it, nevertheless. A decision against the use of an algorithm which is known to be superior reduces the expected value of the subject's pecuniary utility and thus has to be viewed as a behavioral anomaly (cf. Frey, 1992; Kahneman & Tversky, 1979; Tversky & Kahneman, 1974).

3 Methods and Experimental Design

We carry out an economic experiment in the laboratory of the Ostfalia University of Applied Sciences, in which the subjects assume the perspective of a businessperson who offers a service to his/her customers. A decision has to be made on whether this service should be carried out by specialized algorithms or by human experts.

The involvement of students as subjects was approved by the dean's office of the business faculty and the research commission of the Ostfalia University of Applied Sciences. The economic experiment took place as part of a regular laboratory class. All participants were at least 18 years of age at the time of the experiment and are therefore considered to be of legal age in Germany. The participants had confirmed their consent by registration for the economic experiment in the online portal of the Ostfalia University, which is sufficient according to the dean's office and the research commission. Before the start of the economic experiment, they were informed again that their participation was completely voluntary and that they could leave at any time.

In this framing approach, six decision-making scenarios are contrasted that entail different degrees of gravity of their potential consequences if they are executed not successfully. We base our experimental approach on the factual contexts in which algorithms can be used, described in the introduction, and assume that subjects perceive gravity differently in these contexts. The following services are considered: (1) Driving service with the aid of autonomous vehicles (algorithm) or with the aid of drivers, (2) The evaluation of MRI scans with the help of a specialized computer program (algorithm) or with the aid of doctors, (3) The evaluation of files on criminal cases with the aid of a specialized computer program (algorithm) or with the help of legal specialists, (4) A dating site providing matchmaking with the aid of a specialized computer program (algorithm) or with the support of staff trained in psychology, (5) The selection of recipes for cooking subscription boxes with the aid of a specialized computer program or the help of staff trained as professional chefs, and (6) The drawing up of weather forecasts with

the help of a specialized computer program (algorithm) or using experienced meteorologists (Table 2).

Table 2: Decision-making scenarios

Decision-making scenarios
(1) Driving service
(2) Evaluation of MRI scans
(3) The assessment of criminal case files
(4) Dating service
(5) Selection of cooking recipes
(6) Drawing up weather forecasts

The six scenarios that are part of this study were identified through a pre-test, in which additional scenarios were also presented from a literature analysis and brainstorming process. The final selection was made based on three criteria: comprehensibility (do the subjects understand what this application area for algorithms is about?), familiarity (do the subjects know the application area from personal experience or from the media?), and scope (are the high and low scope scenarios actually evaluated as such?). The scenarios are selected in such a way that they are relevant in the literature and that the subjects should be familiar with them from public debates or from their own experience. In this way, it is easier for the subjects to immerse themselves in the respective context. Detailed descriptions of the scenarios can be viewed in Appendix C.

The study has a between-subjects design. Each subject is only confronted with one of a total of six scenarios. All six scenarios have the same probability of success: the algorithm carries out the service with a probability of success of 70%. The human expert carries out the service with a probability of success of 60%. The participants receive a show-up fee of €2, and an additional payment of €4 if the service is carried out successfully. Since we apply the same mathematical conditions of a successful execution of a service to each scenario, only the contextual framework of the six scenarios varies. A perfectly rational economic subject (homo oeconomicus) decides to use the algorithm in all six scenarios because this leads to the maximization of the expected value of the compensation. The context of the respective scenario does not play any role for a homo oeconomicus, because he exclusively strives to maximize his pecuniary benefit.

Before the experiment begins, all participants have to answer test questions (Appendix B). They have a maximum of two attempts at this. Participants who answer the test questions incorrectly twice are disregarded in the analysis, as the data should not be distorted by subjects who have misunderstood the task. The experiment starts with the participants being asked to assess the gravity of the shown decision-making scenario on a scale from 0 (not serious) to 10 (very serious). This allows us to evaluate the different scenarios based on the perceived gravity of the subjects. In this way, it is possible to assess how subjects perceive the potential effects in the context of one scenario compared to the

context of another scenario. In the case of the driving service and the evaluation of MRI scans, it could be a matter of life and death. In the evaluation of documents in the context of criminal cases, it could lead to serious limitations of personal freedom. These scenarios could thus have serious consequences for third parties if they end unfavorably. The situation is different in the case of matchmaking, selecting cooking recipes and drawing up weather forecasts. Even when these tasks cannot be accomplished in a satisfactory way sometimes, the consequences should usually not be very serious. A date might turn out to be dull, or one is disappointed by the taste of a lunch, or you are out without a jacket in the rain. None of those things would be pleasant, but the potential consequences would be trivial.

A *homo oeconomicus* (a person who acts rationally in economic terms) must – regardless of the context – prefer the algorithm to human experts, because it maximizes his or her financial utility. Every decision in favor of the human experts has to be considered algorithm aversion.

Algorithm aversion is a phenomenon which can occur in a wide range of decision-making situations (cf. Burton, Stein & Jensen, 2020). We thus presume that the phenomenon can also be observed in this study. Although the scenarios offer no rational reasons for choosing the human experts, some of the participants will do precisely this. Hypothesis 1 is: Not every subject will select the algorithm. Null hypothesis 1 is therefore: Every subject will select the algorithm.

There is some evidence that the extent of algorithm aversion is influenced by the framing of the conditions under which an algorithm operates. Hou and Jung (2021) have subjects complete estimation tasks using algorithms. They vary the description of the algorithm using a framing approach and find that this has a significant impact on the willingness to follow the algorithm's advice. Utz, Wolfers and Göritz (2021) investigate the perspective on a decision. In three scenarios, they use a framing approach to vary whether a subject is the decision maker or the one affected by the decision. The influence of perspective on the choice behavior between human and algorithm is significant only in one of the three scenarios, namely in the distribution of ventilators for Covid-19 patients.

Regarding the importance and consequences of a task, the findings to date are mixed. Castelo, Bos and Lehmann (2020) use a vignette study to show that framing is suited to influencing algorithm aversion. A self-reported dislike for or distrust in algorithms appears to various degrees in different contexts of a decision. Likewise, Renier, Schmid Mast and Bekbergenova (2021) study, among other things, the relationship between algorithm aversion and the magnitude of a decision for the human who must bear the consequences of the decision. In a vignette study, they vary the magnitude of the consequences that result from an algorithm error. According to their description of the task, the people affected may, for example, be wrongly denied a job contract or a loan. In contrast to Castelo, Bos and Lehmann (2020), they conclude that the scope has no influence on the extent of algorithm aversion.

The difference in the results of the mentioned studies already shows that there still seems to be a large knowledge gap here. Sometimes a framing approach seems suitable to change decision behavior in the context of algorithm use, and sometimes not. Nonetheless, in all four studies mentioned above, the algorithm was not recognizably the most reliable alternative, and there is also no performance-related payment for the subjects. Algorithm aversion is therefore not modeled as a behavioral anomaly.

To extend our understanding, we analyze the extent of algorithm aversion in six differently framed decision situations. We believe that a clear financial incentive that models algorithm aversion as a behavioral anomaly will enhance the framing effect. We expect that the frame will have an influence on algorithm aversion analogous to Castelo, Bos and Lehmann (2020) if the financial advantage of the algorithm is clearly recognizable. Hypothesis 2 is: The proportion of decisions made in favor of the algorithm will vary significantly between the decision situations perceived as serious and trivial. Null hypothesis 2 is therefore: The proportion of decisions made in favor of the algorithm will not vary significantly between the decision situations perceived as serious and trivial.

In the literature there are numerous indications that framing can significantly influence the decision-making behavior of subjects (cf. Tversky & Kahneman, 1981). If subjects acted rationally and maximized their utility, neither algorithm aversion nor the framing effect would arise. Nonetheless, real human subjects – as the research in behavioral economics frequently shows – do not act like *homo oeconomicus.* Their behavior usually tends to correspond more to the model of bounded rationality put forward by Herbert A. Simon (1959). Human beings suffer from cognitive limitations – they fall back on rules of thumb and heuristics. But they do try to make meaningful decisions – as long as this does not involve too much effort. This kind of 'being sensible' – which is often praised as common sense – suggests that great efforts have to be made when decisions can have particularly serious consequences (for an overview of bounded rationality see Grüne-Yanoff, 2007; Hoffrage & Reimer, 2004; Lipman, 1995).

Jack W. Brehm's motivational intensity theory (see, e.g., Brehm & Self, 1989) identifies three main determinants of effort to make successful decisions: (1) The importance of the outcome of a successful decision, (2) the degree of difficulty of the task, and (3) the subjective assessment that the task can be successfully accomplished. The more important the outcome of a successful decision, the more pronounced the effort to make a successful decision. The more difficult the task is in relation to the desired outcome and the lower the prospect of successfully accomplishing the task, the weaker the effort to make a successful decision is pronounced.

The last two aspects are unlikely to vary much across the six decision situations in this study. The degree of difficulty of the task is consistent in all six cases. All that is required is to weigh the algorithm's probability of success (70%) against the human expert's probability of success (60%). This is a simple task - in all six decision situations. It can be assumed that this level of difficulty is perceived as manageable by the subjects - in all six decision situations. However, the importance of the outcome of a decision differs in the

six decision situations. Three decision situations have potentially serious consequences, and the other three decision situations have potentially trivial consequences. Thus, it is to be expected that subjects will try harder to make a successful decision in the decisions that involve potentially serious consequences. This is in line with other research that shows that the valence of a goal influences expectancy formation (Gendolla, 1997) and leads to increasing motivation (Vroom, 1964) and commitment to a task (Gollwitzer, 1993; Mento, Cartledge & Locke, 1980).

This is also consistent with what would generally be recognized as common sense. This everyday common sense, which demands different levels of effort for decision-making situations with different degrees of gravity, could contribute towards the behavioral anomaly of algorithm aversion appearing more seldom in decisions with possible serious consequences than in decisions with relatively insignificant effects. The founding of a company is certainly given much more thought than choosing which television program to watch on a rainy Sunday afternoon. And much more care will usually be invested in the selection of a heart surgeon than in the choice of a pizza delivery service.

The assumption that higher valence of a situation leads to more effort in decision making has already been supported by experimental economics in other contexts. For example, Muraven and Slessareva (2003) tell a subset of their subjects that their responses in an effort task will be used for important research projects to combat Alzheimer's disease. The mere belief that their effort may possibly reduce the suffering of Alzheimer's patients leads subjects to perform significantly better than in a control group. Since higher task importance may contribute to exerting more effort, we hypothesize that it also leads subjects to focus on the relevant aspects of a decision (here: the different probabilities of success), thus eventually decreasing algorithm aversion. Hypothesis 3 is thus: The greater the gravity of a decision, the more seldom the behavioral anomaly of algorithm aversion arises. Null hypothesis 3 is therefore: Even when the gravity of a decision-making situation increases, there is no reduction in algorithm aversion.

4 Results

This economic experiment is carried out between 2-14 November 2020 in the Ostfalia Laboratory of Experimental Economic Research (OLEW) of Ostfalia University of Applied Sciences in Wolfsburg. A total of 143 students of the Ostfalia University of Applied Sciences take part in the experiment. Of these, 91 subjects are male (63.6%), 50 subjects are female (35%) and 2 subjects (1.4%) describe themselves as non-binary. Of the 143 participants, 65 subjects (45.5%) study at the Faculty of Business, 60 subjects (42.0%) at the Faculty of Vehicle Technology, and 18 subjects (12.6%) at the Faculty of Health Care. Their average age is 23.5 years.

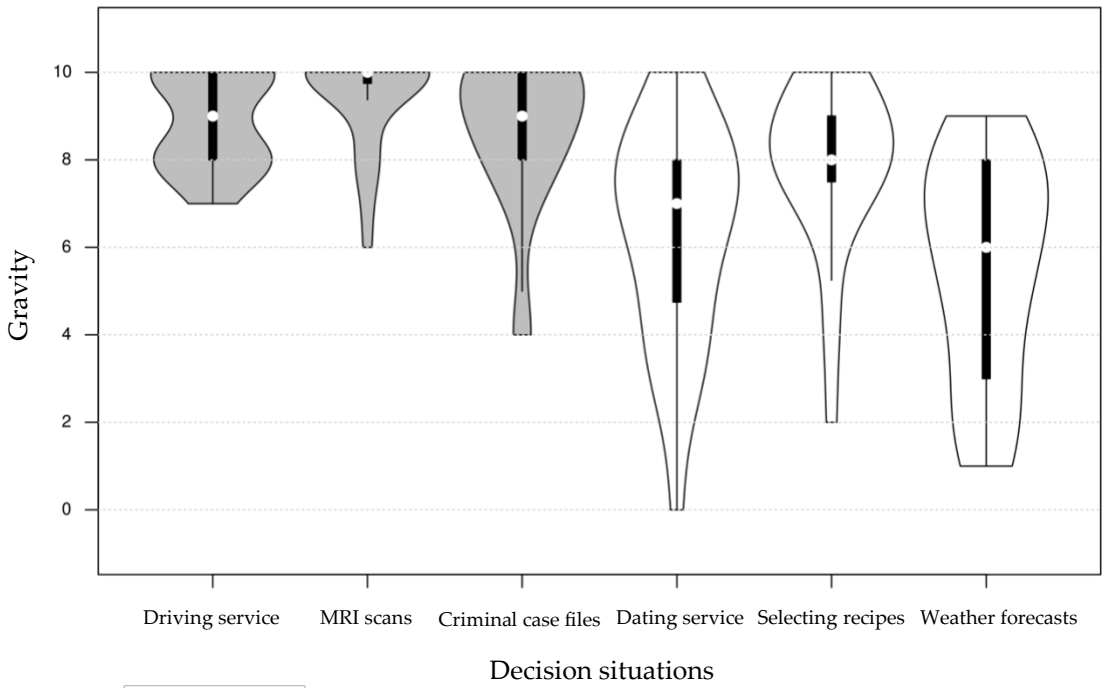
The experiment is programmed in z-Tree (cf. Fischbacher, 2007). Only the lottery used to determine the level of success when providing the service is carried out by taking a card from a pack of cards. In this way we want to counteract any possible suspicion that

the random event could be manipulated. The subjects see the playing cards and can be sure that when they choose the algorithm there is a probability of 70% that they will be successful (the pack of cards consists of seven +€4 cards and three ±€0 cards). In addition, they can be sure that if they choose a human expert their probability of success is 60% (the pack of cards consists of six +€4 cards and four €±0 cards) (see Figure A-1 and Figure A-2 in Appendix D).

The time needed for reading the instructions of the game (Appendix A), answering the test questions (Appendix B), and carrying out the task is 10 minutes on average. A show-up fee of €2 and the possibility of a performance-related payment of €4 seem appropriate for the time spent - it is intended to be sufficient incentive for meaningful economic decisions, and the subjects do give the impression of being concentrated and motivated.

We provide the subjects with only one contextual framework of a decision situation at a time. Here, the subjects are presented with decision situations in the context of driving service (25 subjects), evaluating MRI scans (24 subjects), assessing criminal case files (22 subjects), dating service (24 subjects), selecting cooking recipes (23 subjects), and drawing of weather forecasts (25 subjects). Subjects were distributed similarly evenly across the contextual decision situations in terms of gender and faculty.

Figure 1: Violin plots for the assessment of the gravity of the decision-making situations



Overall, only 87 out of 143 subjects (60.84%) decide to delegate the service to the (superior) algorithm. A total of 56 subjects (39.16%) prefer to rely on human experts in spite of the lower probability of success. Null hypothesis 1 thus has to be rejected. The result of the chi-square goodness of fit test is highly significant (χ^2 (N = 143) = 21.93,

$p < 0.001$). On average, around two out of five subjects thus tend towards algorithm aversion. All subjects should be aware that preferring human experts and rejecting the algorithm reduces the expected value of the performance-related payment. However, the need to decide against the algorithm is obviously strong in a part of the subjects. To investigate the effects of our framing approach on algorithm aversion (hypothesis 2), we must first examine how subjects evaluate the six differently framed scenarios. The subjects perceive the gravity of the decision situations differently (Figure 1). While in the contextual decision situations driving service (mean gravity of 8.88), evaluation of MRI scans (mean gravity of 9.42) and assessment of criminal case files (mean gravity of 8.68), the potential consequences of not successfully performing the service are perceived as comparatively serious, the contextual decision situations dating service (mean gravity of 6.33), selection of cooking recipes (mean gravity of 7.87) and drawing up weather forecasts (mean gravity of 5.52) show less pronounced perceived gravity of potential consequences (Table 3).

Table 3: Evaluation of gravity in a contextual decision situation

Scenario	#	Mean value of gravity	Median	Standard deviation
(1) Driving service	25	8.88	9	1.09
(2) Evaluation of MRI scans	24	9.42	10	1.14
(3) Criminal case files	22	8.68	9	1.78
(4) Dating service	24	6.33	7	2.50
(5) Selection of recipes	23	7.87	8	1.96
(6) Weather forecasts	25	5.52	6	2.57

In the six scenarios, each with identical chances of success for execution by a human expert or an algorithm, subjects decide by whom the service should be performed depending on the context in which the situation is presented. By considering the context, subjects arrive at an assessment of the gravity of the potential consequences of not successfully performing the service (Figure 1). Even though the six scenarios differ considerably from each other in their context, they are also similar in the assessment of their gravity when viewed individually. The perceived gravity of the scenarios as reported by the subjects suggests that the decision situations can be considered in two clusters, decisions with possibly serious consequences (for the highest mean gravity scores) and decisions with possibly trivial consequences (for the lowest mean gravity scores).

The comparison of the contextual decision situations with possibly serious consequences and those with possibly trivial consequences, as indicated by the mean values of the gravity levels per decision situation, show that the perceived gravity of the six scenarios is (highly) significantly different when using the Wilcoxon rank-sum test

(Table 4). For example, perceived gravity of driving service differs from dating service with a $p < 0.001$. Only the scenarios driving service and assessment of criminal case files differ from the scenario recipe selection only at a significance level of 0.1, as already suggested by their mean gravity. On average, the possible consequences of recipe selection are perceived as slightly more serious, but not as serious as driving service, evaluating MRI scans or criminal case files. Cohen's d shows how much the means of two samples differ. An effect size of 0.2 corresponds to small effects, 0.5 to medium effects, and 0.8 to large effects (Cohen, 1992).

Table 4: Comparison of perceived gravity using Wilcoxon rank-sum test and Cohen's d

	(4) Dating service		(5) Selection of recipes		(6) Weather forecasts	
	p-value*	Cohen's d	p-value*	Cohen's d	p-value*	Cohen's d
(1) Driving service	0.000	1.33	0.064	0.64	0.000	1.70
(2) Evaluation of MRI scans	0.000	1.59	0.000	0.97	0.000	1.95
(3) Criminal case files	0.000	1.07	0.061	0.43	0.000	1.41

*p-values using Wilcoxon rank-sum test.

This confirms that subjects perceive the consequences of decision contexts to vary in gravity and leads to the classification of decision contexts into cluster A1 (possibly serious consequences: driving service, evaluation of MRI scans, and criminal case files) and cluster A2 (possibly trivial consequences: dating service, selecting cooking recipes, and weather forecasts) that we propose in this framework. The violin plot of the summarized decision situations shows that the subjects rate the gravity higher in contexts with critical consequences for physical integrity than in contexts where it does not matter (Figure 2). However, a direct comparison of the violin plots also shows that the range of decision situations rated as having trivial consequences is wider than that of the others, since some subjects also rate the gravity as very high here.

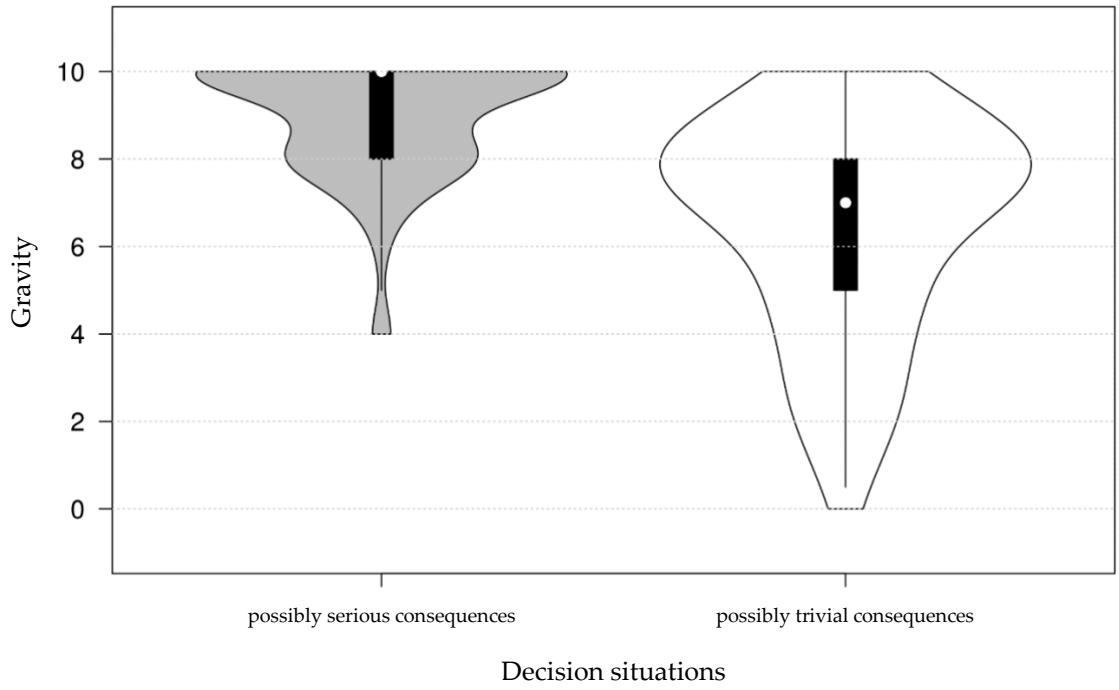
Table 5: Evaluation of gravity in clusters A1 and A2

	Cluster A1 (serious)	Cluster A2 (trivial)
Median	10	7
Average	9.00	6.54
Standard deviation	1.37	2.53

Still, the possible consequences of each decision situation from cluster A1 are rated by the subjects as more serious than those from cluster A2. According to this classification, the mean of the perceived gravity for the decision situations with possibly serious

consequences (A1) is 9.0 with a standard deviation of 1.37. In contrast, when the gravity of the decision situations with possibly trivial consequences (A2) is evaluated, the mean is 6.54 with a standard deviation of 2.53 (Table 5). The Wilcoxon rank-sum test shows that the gravity of the decision situations in cluster A1 is assessed as significantly higher than that of the decision situations in cluster A2 ($z = 6.689$; $p < 0.001$).

Figure 2: Violin plots for scenarios with possibly serious and trivial consequences



Furthermore, a difference in the number of decisions in favor of the algorithm between the two clusters can be observed. While 50.7% of the subjects in cluster A1 choose the algorithm, 70.83% in cluster A2 rely on it (for the individual decisions in the contextual decision situations, see Table 6). The chi-square test reveals that null hypothesis 2 has to be rejected (χ^2 (N = 143) = 6.08, p = 0.014). The frequency with which algorithm aversion occurs is influenced by the implications involved in the decision-making situation. The framing effect has an impact.

There may be situations in which people like to act irrationally at times. However, common sense suggests that one should allow oneself such lapses in situations where serious consequences must not be feared. For example, there is a nice barbecue going on and the host opens a third barrel of beer although he suspects that this will lead to hangovers the next day among some of his guests. In the case of important decisions, however, one should be wide awake and try to distance oneself from reckless tendencies. For example, if the same man visits a friend in hospital whose life would be acutely threatened by drinking alcohol after undergoing a complicated stomach operation, he would be wise to avoid bringing him a bottle of his favorite whisky. This comparison of

two examples illustrates what could be described as common sense and would be approved of by most neutral observers.

Table 6: Decisions for and against the algorithm

		Decisions for the algorithm		Decisions against the algorithm	
	Total	n	Percent	n	Percent
Cluster A1 (serious)	**71**	**36**	**50.70%**	**35**	**49.30%**
(1) Driving car	25	13	52.00%	12	48.00%
(2) Evaluation of MRI scans	24	13	54.17%	11	45.83%
(3) Criminal case files	22	10	45.45%	12	54.55%
Cluster A2 (trivial)	**72**	**51**	**70.83%**	**21**	**29.17%**
(4) Dating service	24	18	75.00%	6	25.00%
(5) Selection of recipes	23	14	60.87%	9	39.13%
(6) Weather forecasts	25	19	76.00%	6	24.00%
Σ (total)	**143**	**87**	**60.84%**	**56**	**39.16%**

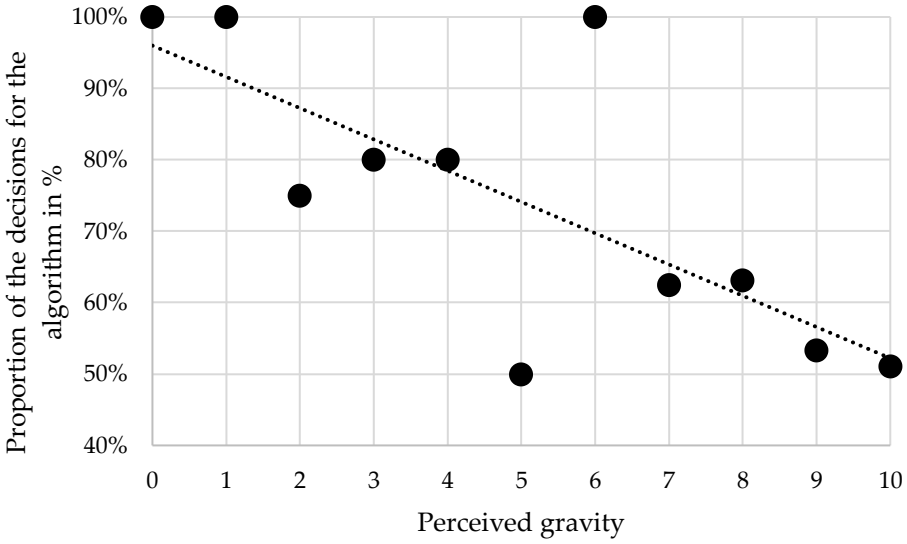
Nevertheless, the results of the experiment point in the opposite direction. A framing effect sets in, but not in the way one might expect. Whereas in cluster A1 (possibly serious consequences) 49.3% of the subjects do exhibit the behavioral anomaly of algorithm aversion, this is only the case in 29.17% of the subjects in cluster A2 (possibly trivial consequences) (Table 6). It seems that algorithm aversion is all the more pronounced in important tasks.

This is confirmed by a regression analysis which demonstrates the relationship between algorithm aversion and the perceived gravity of a scenario. To perform the regression analysis, we detach from the pairwise consideration of the two clusters and relate how serious an economic agent perceived the potential consequences of his or her decision and how it was decided. This is independent of the decision context, only the perceived gravity and the associated decision are considered. For the possible assessments of the consequences (from 0 = not serious to 10 = very serious), the respective average percentage of the decisions in favor of the algorithm is determined. The decisions of all 143 subjects are included in the regression analysis (Figure 3).

If the common sense described above would have an effect, the percentage of decisions for the algorithm from left to right (in other words with increasing perceived gravity of the decision-making situation) would tend to rise. Instead, the opposite can be observed. Whereas in the case of only a low level of gravity (zero and one) 100% of decisions are still made in favor of the algorithm, the proportion of decisions for the algorithm decreases with increasing gravity. In the case of very serious implications (nine and ten),

only somewhat more than half of the subjects decide to have the service carried out by an algorithm (Figure 3). If the perceived gravity of a decision increases by a unit, the probability of a decision in favor of the algorithm falls by 3.9% (t = -2.29; p = 0.023). The 95% confidence interval ranges from -7.27% to -0.54%. Null hypothesis 3 can therefore not be rejected. In situations which might have serious consequences in the case of an error, algorithm aversion is actually especially pronounced.

Figure 3: Decisions in favor of the algorithm depending on the perceived gravity of the decision-making situation



Further analysis shows that the choices between algorithms and human experts are also not statistically significantly influenced by gender (χ^2 (N = 143) = 2.22, p = 0.136), age (t (N = 143) = -0.44, = 0.661), mother tongue (χ^2 (N = 143) = 2.68, p = 0.102), faculty at which a subject is studying (χ^2 (N = 143) = 1.06, p = 0.589), semester (t (N = 143) = 0.63, p = 0.528), or previous participations in economic experiments (χ^2 (N = 143) = 0.21, p = 0.644).

The six scenarios differ in numerous aspects. In order to identify the main factors influencing the decisions of the subjects, clusters are formed based on different criteria and examined with regard to differences in the subjects' selection behavior. There are a total of ten ways to divide six scenarios into two clusters. All ten clustering opportunities are shown in Table 7.

The criterion in focus in this study is the scope of a decision (clustering opportunity). We can group the six scenarios into tasks that have potentially serious consequences if performed incorrectly, e.g., death or unjust imprisonment. These are mainly driving service, evaluation of MRI scans, and assessment of criminal case files (cluster A1). On the other hand, there are tasks where the consequences are trivial if performed poorly. These are dating service, selection of cooking recipes, and weather forecasts (cluster A2). The chi-square test shows that the willingness to use an algorithm is significantly higher

in the latter cluster (χ^2 (N = 143) = 6.08, p = 0.014). The more serious the consequences of a decision, the less likely subjects are to delegate the decision to an algorithm.

Another aspect is the familiarity with a task (Cluster J). A connection between algorithm aversion and familiarity has been suspected for some time. Luo et al. (2021) argue that the more familiar and confident sales agents in dealing with a task, the more pronounced their algorithm aversion is. Gaube et al. (2021) explicitly examine the influence of familiarity with a task on physicians' algorithm aversion in the context of evaluating human chest radiographs. They contrast experienced radiologists, who have a great deal of routine with this task, with inexperienced emergency physicians. Their results also suggest that algorithm aversion may increase with increasing experience in handling a task. We can group the six scenarios into tasks that are performed frequently, perhaps even daily, by an average person. These are driving a car, selection of cooking recipes, and weather forecasts. Almost every day, each of us commutes to work or other places, decides what to eat, and wonders what the weather will be like during the day. On the other hand, evaluation of MRI scans and assessment of criminal case files are activities that most of us may never have encountered, and dating service is something that those who are single may use from time to time, and those who are in a relationship (hopefully) not that much. The chi-square test shows no significant difference between the clusters J1 and J2 (χ^2 (N = 143) = 0.30, p = 0.586). Thus, the willingness to use an algorithm does not seem to be considerably affected by how often we engage in a particular activity.

Further interesting aspects are whether an algorithm requires an expert to operate it adequately or whether it can also be used by a layperson (Cluster H), whether a task requires human skills, such as empathy, or not (Cluster D), and the maturity of the technology, i.e., whether the use of algorithms is already widespread today or not (Cluster F). Algorithm aversion has been observed both in extremely simple algorithms that a layperson can easily operate by him- or herself and in extremely complex algorithms (numerous examples can be found in Castelo, Bos & Lehmann, 2020). Regarding human skills, Fuchs et al. (2016) find that algorithm aversion is particularly high for tasks that are driven more by human skills than by mathematical data analysis. Kaufmann (2021) shows that algorithm aversion can occur to a large extent in student performance evaluation, a task that is characterized as requiring a lot of empathy. On the maturity of technology, already 17 years ago Nadler and Shestowsky (2006) raise the question of whether subjects may become accustomed to using an algorithm the longer it is established in the market.

It turns out that the willingness to choose an algorithm does not depend on the amount of expertise required to operate it (χ^2 (N = 143) = 0.17, p = 0.682), nor on the extent to which human skills are involved in the task it is supposed to perform (χ^2 (N = 143) = 0.00, p = 0.994). Regarding the maturity of technology, we see that activities that are already automated frequently in practice today, such as making weather forecasts, are also delegated to the algorithm much more often in the experiment (χ^2 (N = 143) = 3.67, p = 0.056). However, the difference between the clusters F1 and F2 is not as large as between the clusters A1 and A2. Moreover, there is also no significant difference at a significance

level of 0.05 in the frequency with which an algorithm is selected in the remaining five clustering opportunities. It therefore seems that of all the differences between the frames, the gravity of consequences of a decision are the most important aspect.

Table 7: Overview of all possible clusters obtained by grouping the frames evenly

Clustering Opportunities	Cluster	Frames	n	Algorithm Use	χ^2	p-value
A	A1	(1) (2) (3)	71	50.70%	6.080	**0.014**
	A2	(4) (5) (6)	72	70.83%		
B	B1	(1) (2) (4)	73	60.27%	0.020	0.888
	B2	(3) (5) (6)	70	61.43%		
C	C1	(1) (2) (5)	72	55.56%	1.699	0.192
	C2	(3) (4) (6)	71	66.20%		
D	D1	(1) (2) (6)	74	60.81%	0.000	0.994
	D2	(3) (4) (5)	69	60.87%		
E	E1	(1) (3) (4)	71	57.75%	0.566	0.452
	E2	(2) (5) (6)	72	63.89%		
F	F1	(1) (3) (5)	70	52.86%	3.667	0.056
	F2	(2) (4) (6)	73	68.49%		
G	G1	(1) (3) (6)	72	58.33%	0.382	0.536
	G2	(2) (4) (5)	71	63.38%		
H	H1	(1) (4) (5)	72	62.50%	0.168	0.682
	H2	(2) (3) (6)	71	59.16%		
I	I1	(1) (4) (6)	74	67.57%	2.914	0.088
	I2	(2) (3) (5)	69	53.62%		
J	J1	(1) (5) (6)	73	63.01%	0.296	0.586
	J2	(2) (3) (4)	70	58.57%		

(1) = Driving service, (2) = Evaluation of MRI scans, (3) = Assessment of criminal case files, (4) = Dating service, (5) = Selection of cooking recipes, (6) = Weather forecasts.

5 Discussion

General

The results are surprising, given that common sense would deem – particularly in the case of decisions which might have serious consequences – that the option with the greatest probability of success should be chosen. Also, with regard to Brehm's motivational intensity theory, it can be argued that the importance of the successful execution of the action is not adequately reflected in the subjects' decisions. In line with Hou and Jung (2021) and Castelo, Bos and Lehmann (2020), our results also show that a framing approach is suitable to influence decisions to engage an algorithm. The study by Utz, Wolfers and Göritz (2021) shows that the preference to use an algorithm in moral scenarios (distribution of ventilators for Covid-19 treatment) is low. In our study, in scenarios that were perceived as scenarios with potentially serious consequences (driving service, evaluation of MRI scans and criminal case files) and also raise moral issues, a lower utilization rate of the algorithm is also shown. A survey by Grzymek and

Puntschuh (2019) clearly shows that people are less likely to use an algorithm in decision-making situations with potentially serious consequences, such as diagnosing diseases, evaluating creditworthiness, trading stocks, or pre-selecting job applicants, but more likely to use an algorithm in scenarios with less serious consequences, such as spell-checking, personalizing advertisements, or selecting the best travel route. In contrast, Renier, Schmid Mast and Bekbergenova (2021) found no effect of gravity on the extent to which participants demand an improvement to an algorithm.

If subjects allow themselves to be influenced by algorithm aversion to make decisions to their own detriment, they should only do so when they can take responsibility for the consequences with a clear conscience. In cases where the consequences are particularly serious, maximization of the success rate should take priority. But the exact opposite is the case. Algorithm aversion appears most frequently in cases where it can cause the most damage. To this extent it seems necessary to speak of the tragedy of algorithm aversion.

Implications

Our results suggest that algorithm aversion is particularly prevalent where potential errors have dire consequences. This means that algorithm aversion should be especially addressed by those developers, salespeople, and other staff whose supervised areas of operation are related to human health and safety. This can be done, for example, through staff training and intensive field testing with potential users. In addition, the results suggest that clever framing of the activity that an algorithm undertakes can make users more likely to use the algorithm. For example, neutral words should be chosen when advertising medical or investment algorithms, rather than unnecessarily pointing out the general risks of such activity.

Limitations and directions for future research

Despite their advantages in establishing causal relationships, framing studies always carry the risk that subjects may have many other associations with the decision-making situations that are not the focus of the study, and yet have an unintended influence on the results. In our study, these are in particular the complexity and subjectivity of the tasks, but also moral aspects, that may be more relevant in the decisions with potentially serious consequences, in which the physical well-being or the freedom of humans are at stake. In addition, we do not focus on the variation in perceived gravity within one scenario (e.g., MRI scans for live threatening diagnosis vs. MRI scans for less severe diagnosis), but rather on the variation in gravity between different scenarios, which could be a risk in regard to causality. It remains for further studies to vary the gravity within one scenario. Moreover, these aspects may also include the familiarity from the everyday experiences of the subjects, which should be higher, for example, for weather forecasts than for MRI scans. However, these associations do not affect our core result. Our regression analysis only considers the correlation between algorithm aversion and the subjectively perceived gravity of consequences, regardless of the scenario, and finds that higher perceived consequences in general lead to an increase in algorithm aversion.

Second, it should be noted that prior experiences of the subjects and the maturity of the technologies may lead to different expectations regarding the success rates. For example, the use of algorithms for weather forecasting is already advanced and it is to be expected that an algorithm would perform better here than a human. In autonomous driving, on the other hand, the technology is not yet as advanced. However, to ensure the comparability of the scenarios, in our framing approach the probabilities must be identical in all scenarios, which may not always fit the subjects' expectations. In addition, the success rates are directly given in the instructions of our experiment. In real life, we would first gather our own experience in all these areas to get an idea of when to rely on algorithms and when not to. Moreover, the sample size of our experiment is rather small with 143 participants. We therefore encourage future research efforts to further explore our results in a research design with more practice-oriented conditions and with a larger sample.

Finally, it is needless to say that the consequences of the decisions in our experiments might have to be borne by third parties. It would be possible to continue this line of research by giving up the framing approach and modeling a situation where the subjects are directly affected. In this case, different incentives would have to be introduced into the decision situations. Success in scenarios with possible serious consequences would then have to be rewarded with a higher amount than in scenarios with trivial consequences. However, we presume that our results would also be confirmed by an experiment based on this approach, given that it is a between-subjects design in which every subject is only presented with one scenario. Whether one receives €4 or €8 for a successful choice will probably not have a notable influence on the results. Nonetheless, the empirical examination of this assessment is something which will have to wait for future research efforts.

6 Summary

Many people decide against the use of an algorithm even when it is clear that the algorithm promises a higher probability of success than a human mind. This behavioral anomaly is referred to as algorithm aversion.

The subjects are placed in the position of a businessperson who has to choose whether to have a service carried out by an algorithm or by a human expert. If the service is carried out successfully, the subject receives a performance-related payment. The subjects are informed that using the respective algorithm leads to success in 70% of all cases, while the human expert is only successful in 60% of all cases. In view of the recognizably higher success rate, there is every reason to trust in the algorithm. Nevertheless, just under 40% of the subjects decide to use the human expert and not the algorithm. In this way they reduce the expected value of their performance-related payment and thus manifest the behavioral anomaly of algorithm aversion.

The most important objective of the study is to find out whether decision-making situations of varying gravity can lead to differing frequencies of the occurrence of algorithm aversion. To do this, we choose a framing approach. Six decision-making situations (with potentially serious / trivial consequences) have an identical payment structure. Against this background there is no incentive or reason to act differently in each of the six scenarios. It is a between-subjects approach – each subject is only presented with one of the six decision-making situations.

In the three scenarios with potentially serious consequences for third parties, just under 50% of the subjects exhibit algorithm aversion. In the three scenarios with potentially trivial consequences for third parties, however, less than 30% of the subjects exhibit algorithm aversion.

This is a surprising result. If a framing effect were to occur, it would have been expected to be in the opposite direction. In cases with implications for freedom or even danger to life, one should tend to select the algorithm as the option with a better success rate. Instead, algorithm aversion shows itself particularly strongly here. If it is only a matter of arranging a date, creating a weather forecast or offering cooking recipes, the possible consequences are quite clear. In a situation of this kind, one can still afford to have irrational reservations about an algorithm. Surprisingly, however, algorithm aversion occurs relatively infrequently in these situations.

One can call it the tragedy of algorithm aversion because it arises above all in situations in which it can cause particularly serious damage.

Author Contributions

Conceptualization, I.F., J.R.J., M.L., M.S.; Data curation, J.R.J., M.L.; Formal analysis, I.F., J.R.J., M.L., M.S.; Software, J.R.J., M.L.; Validation, I.F., J.R.J., M.L., M.S.; Visualization, J.R.J., M.L.; Writing - original draft, J.R.J., M.L., M.S.; Writing - review & editing, I.F., J.R.J., M.L., M.S.

Acknowledgements

The authors would like to thank the editor, the anonymous reviewers, the participants in the German Association for Experimental Economic Research e.V. (GfeW) Meeting 2021, the participants in the Economic Science Association (ESA) Meeting 2021, and the participants in the PhD seminar at the Georg August University of Göttingen, for their constructive comments and useful suggestions, which were very helpful in improving the manuscript.

References

Alexander, V., Blinder, C., & Zak, P. J. (2018), Why trust an algorithm? Performance, cognition, and neurophysiology, *Computers in Human Behavior*, 89(2018), 279-288.

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2020), Watch Me Improve—Algorithm Aversion and Demonstrating the Ability to Learn, *Business & Information Systems Engineering*, 1-14.

Brehm, J. W., & Self, E. A. (1989), The intensity of motivation, *Annual Review of Psychology*, 40, 109-131.

Brozovsky, L., & Petříček, V. (2007), Recommender System for Online Dating Service, *ArXiv*, abs/cs/0703042.

Burton, J., Stein, M., & Jensen, T. (2020), A Systematic Review of Algorithm Aversion in Augmented Decision Making, *Journal of Behavioral Decision Making*, 33(2), 220-239.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2020), Task-dependent algorithm aversion, *Journal of Marketing Research*, 56(5), 809-825.

Cohen, J. (1992), A power primer, *Psychological bulletin*, 112(1), 155-159.

Commerford, B. P., Dennis, S. A., Joe, J. R., & Wang, J. (2019), Complex estimates and auditor reliance on artificial intelligence, *SSRN*, 3422591.

Cornelissen, J., & Werner, M. D. (2014), Putting Framing in Perspective: A Review of Framing and Frame Analysis across the Management and Organizational Literature, *The Academy of Management Annals*, 8(1), 181-235.

Dawes, R., Faust, D., & Meehl, P. (1989), Clinical versus actuarial judgment, *Science*, 243(4899), 1668-1674.

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020), A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores, *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Paper 509, 1-12.

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.

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), 114-126.

Efendić, E., Van de Calseyde, P. P., & Evans, A. M. (2020), Slow response times undermine trust in algorithmic (but not human) predictions, *Organizational Behavior and Human Decision Processes*, 157(C), 103-114.

Erlei, A., Nekdem, F., Meub, L., Anand, A., & Gadiraju, U. (2020), Impact of Algorithmic Decision Making on Human Behavior: Evidence from Ultimatum Bargaining, *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1), 43-52.

Filiz, I., Judek, J. R., Lorenz, M., & Spiwoks, M. (2021), Reducing Algorithm Aversion through Experience, *Journal of Behavioral and Experimental Finance*, 31, 100524.

Fischbacher, U. (2007), z-Tree: Zurich Toolbox for Ready-made Economic Experiments, *Experimental Economics*, 10(2), 171-178.

Frey, B. S. (1992), Behavioural Anomalies and Economics, in: Economics As a Science of Human Behaviour, 171-195.

Fuchs, C., Matt, C., Hess, T., & Hoerndlein, C. (2016), Human vs. Algorithmic recommendations in big data and the role of ambiguity, *Twenty-second Americas Conference on Information Systems*, San Diego, 2016.

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E., & Ghassemi, M. (2021), Do as AI say: susceptibility in deployment of clinical decision-aids, *NPJ Digital Medicine*, 4(1), 1-8.

Gendolla, G. H. (1997), Surprise in the context of achievement: The role of outcome valence and importance, *Motivation and Emotion*, 21(2), 165-193.

Germann, M., & Merkle, C. (2019), Algorithm Aversion in Financial Investing, *SSRN*, 3364850.

Gollwitzer, P. M. (1993), Goal achievement: The role of intentions, *European Review of Social Psychology*, 4(1), 141-185.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000), Clinical versus mechanical prediction: a meta-analysis, *Psychological Assessment,* 12(1), 19-30.

Grüne-Yanoff, T. (2007), Bounded Rationality, *Philosophy Compass*, 2(3), 534-563.

Grzymek, V., & Puntschuh, M. (2019), What Europe Knows and Thinks About Algorithms Results of a Representative Survey. Bertelsmann Stiftung, *eupinions*, February 2019.

Hoffrage, U., & Reimer, T. (2004), Models of bounded rationality: The approach of fast and frugal heuristics, *Management Revue*, 15(4), 437-459.

Horne, B. D., Nevo, D., O'Donovan, J., Cho, J., & Adali, S. (2019), Rating Reliability and Bias in News Articles: Does AI Assistance Help Everyone?, *ArXiv*, abs/1904.01531.

Hou, Y. T. Y., & Jung, M. F. (2021), Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making, *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 1-25.

Ireland, L. (2020), Who errs? Algorithm aversion, the source of judicial error, and public support for self-help behaviors, *Journal of Crime and Justice*, 43(2), 174-192.

Jussupow, E., Benbasat, I., & Heinzl, A. (2020), Why are we averse towards Algorithms? A comprehensive literature Review on Algorithm aversion, *Proceedings of the 28th European Conference on Information Systems (ECIS)*, aisel.aisnet.org/ecis2020_rp/168.

Kahneman, D., & Tversky, A. (1979), Prospect theory: An analysis of decision under risk, *Econometrica*, 47(2), 263-291.

Kaufmann, E. (2021), Algorithm appreciation or aversion? Comparing in-service and pre-service teachers' acceptance of computerized expert models, *Computers and Education: Artificial Intelligence*, 2, 100028.

Kawaguchi, K. (2021), When Will Workers Follow an Algorithm? A Field Experiment with a Retail Business, *Management Science*, 67(3), 1670-1695.

Köbis, N., & Mossink, L. D. (2021), Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry, *Computers in Human Behavior*, 114(2021), 1-13.

Ku, C. Y. (2019), When AIs Say Yes and I Say No: On the Tension between AI's Decision and Human's Decision from the Epistemological Perspectives, *Információs Társadalom*, 19(4), 61-76.

Leyer, M., & Schneider, S. (2019), Me, You or Ai? How Do We Feel About Delegation, *Proceedings of the 27th European Conference on Information Systems (ECIS)*, 1-17.

Lipman, B. L. (1995), Information Processing and Bounded Rationality: A Survey, *Canadian Journal of Economics*, 28(1), 42-67.

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

Luo, X., Qin, M. S., Fang, Z., & Qu, Z. (2021), Artificial intelligence coaches for sales agents: Caveats and solutions, *Journal of Marketing*, 85(2), 14-32.

Mahmud, H., Islam, A. N., Ahmed, S. I., & Smolander, K. (2022), What influences algorithmic decision-making? A systematic literature review on algorithm aversion, *Technological Forecasting and Social Change*, 175, 121390.

Majumdar, A., & Ward, R. (2011), An algorithm for sparse MRI reconstruction by Schatten p-norm minimization, *Magnetic resonance imaging*, 29(3), 408-417.

Mento, A. J., Cartledge, N. D., & Locke, E. A. (1980), Maryland vs Michigan vs Minnesota: Another look at the relationship of expectancy and goal difficulty to task performance, *Organizational Behavior and Human Performance*, 25(3), 419-440.

Muraven, M., & Slessareva, E. (2003), Mechanisms of self-control failure: Motivation and limited resources, *Personality and Social Psychology Bulletin*, 29(7), 894-906.

Nadler, J., & Shestowsky, D. (2006), Negotiation, information technology, and the problem of the faceless other, *Negotiation Theory and Research*, 145-172, New York.

Niszczota, P., & Kaszás, D. (2020), Robo-investment aversion, *PLoS ONE*, 15(9), 1-19.

Önkal, D., Gönül, M. S., & De Baets, S. (2019), Trusting forecasts, *Futures & Foresight Science*, 1(3-4), 1-10.

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

Renier, L. A., Schmid Mast, M., & Bekbergenova, A. (2021), To err is human, not algorithmic–Robust reactions to erring algorithms, *Computers in Human Behavior*, 124, 106879.

Rühr, A., Streich, D., Berger, B., & Hess, T. (2019), A Classification of Decision Automation and Delegation in Digital Investment Systems, *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 1435-1444.

Sawaitul, S. D., Wagh, K., & Chatur, P.N. (2012), Classification and Prediction of Future Weather by using Back Propagation Algorithm-An Approach, *International Journal of Emerging Technology and Advanced Engineering*, 2(1), 110-113.

Shariff, A., Bonnefon, J. F., & Rahwan, I. (2017), Psychological roadblocks to the adoption of self-driving vehicles, *Nature Human* Behaviour, 1(10), 694-696.

Simon, H. A. (1959), Theories of Decision-Making in Economics and Behavioral Science, *The American Economics Review*, 49(3), 253-283.

Simpson, B. (2016), Algorithms or advocacy: does the legal profession have a future in a digital world?, *Information & Communications Technology Law*, 25(1), 50-61.

Tversky, A., & Kahneman, D. (1981), The framing of decisions and the psychology of choice, *Science*, 211(4481), 453-458.

Tversky, A., & Kahneman, D. (1974), Judgment under Uncertainty: Heuristics and Biases, *Science*, 185(4157), 1124-1131.

Ueda, M., Takahata, M., & Nakajima, S. (2011), User's food preference extraction for personalized cooking recipe recommendation, *Proceedings of the Second International Conference on Semantic Personalized Information Management: Retrieval and Recommendation*, 781, 98-105.

Utz, S., Wolfers, L. N., & Göritz, A. S. (2021), The effects of situational and individual factors on algorithm acceptance in covid-19-related decision-making: A preregistered online experiment, *Human-Machine Communication*, 3, 27-45.

Vroom, V. (1964), Work and Motivation, New York: John Wiley & Sons.

Wang, R., Harper, F. M., & Zhu, H. (2020, April), Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences, *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Paper 684, 1-14.

Yeomans, M., Shah, A. K., Mullainathan, S., & Kleinberg, J. (2019), Making Sense of Recommendations, *Journal of Behavioral Decision Making*, 32(4), 403-414.

Youyou, W., Kosinski, M., & Stillwell, D. (2015), Computer-based personality judgments are more accurate than those made by humans, *Proceedings of the National Academy of Sciences*, 112(4), 1036-1040.

Appendix A: Instructions for the game

The Game

You are a businessperson and have to decide whether you want a service you are offering for the first time carried out solely by an algorithm or solely by human experts. You are aware that the human experts carry out the task with a probability of success of 60%. You are also aware that the algorithm carries out the task with a probability of success of 70%.

Procedure

After reading the instructions and answering the test questions the decision-making situation is presented to you. This specifies the service which your company offers. First of all, you are asked to assess the gravity of the decision-making situation from the perspective of your customers. Then you decide whether the service should be carried out by human experts or by an algorithm.

Payment

You receive a show-up fee of €2 for taking part in the experiment. Apart from this, an additional payment of €4 is made if the service is carried out successfully.

Information

- Please remain quiet during the experiment
- Please do not look at your neighbor's screen
- Apart from a pen/pencil and a pocket calculator, no aids are permitted

Appendix B: Test questions

Test question 1: Which alternatives are available to you to carry out the service?

a) I can provide the service myself or have it done by an algorithm.

b) I can provide the service myself or have it done by human experts.

c) I can have the service carried out via human experts or by an algorithm. *(correct)*

Test question 2: For how many newly-offered services do you need to make a choice?

a) None.

b) One. *(correct)*

c) Two.

Test question 3: How much is the bonus payment for carrying out the task successfully?

a) €1
b) €2.50
c) €4 *(correct)*

Test question 4: How much is the bonus payment if you carry out the task wrongly?

a) -€2.50
b) €0 *(correct)*
c) €2.50

Appendix C: Decision-making situations

Decision-making situation 1: Driving service

You are the manager of a public transport company and have to decide whether you want to transport your 100,000 passengers solely with autonomous vehicles (algorithm) or solely with vehicles with drivers (human experts). The task will be considered to have been successfully completed when all of your customers have reached their destination safely. In an extreme case, a wrong decision could mean the death of a passenger.

I choose: O Autonomous vehicles (algorithm)
O Drivers (human experts)

Decision-making situation 2: Evaluation of MRI scans

You are the manager of a large hospital and have to decide whether the MRI scans of your 100,000 patients with brain conditions should be assessed solely by a specialized computer program (algorithm) or solely by doctors (human experts). The task will be considered to have been successfully completed when all life-threatening symptoms are recognized immediately. In an extreme case, a wrong decision could mean the death of a patient.

I choose: O Specialized computer program (algorithm)
O Doctors (human experts)

Decision-making situation 3: Criminal case files

You are the head of a large law firm and have to decide whether the analysis of the case documents of your 100,000 clients should be carried out exclusively by a specialized computer program (algorithm) or solely by defense lawyers (human experts). The task will be considered to have been successfully completed when the penalties issued to your

clients are below the national average. In an extreme case, a wrong decision could mean an unjustified long prison sentence for a client.

I choose: O Specialized computer program (algorithm)

O Defense lawyers (human experts)

Decision-making situation 4: Dating service

You are the manager of an online dating site and have to decide whether potential partners are suggested to your 100,000 customers solely by a specialized computer program (algorithm) or exclusively by trained staff (human experts). The task will be considered to have been successfully completed when you can improve the rating of your app in the App Store. For your customers, a wrong decision could lead to a date with a sub-optimal candidate.

I choose: O Specialized computer program (algorithm)

O Trained staff (human experts)

Decision-making situation 5: Selection of cooking recipes

You are the manager of an online food retailer and have to decide whether your 100,000 cooking boxes – with ingredients and recipes which are individually tailored to the customers – are put together solely by a specialized computer program (algorithm) or solely by trained staff (human experts). The task will be considered to have been successfully completed when you can increase the reorder rate as a key indicator of customer satisfaction. A wrong decision could mean that the customers don't like their meal.

I choose: O Specialized computer program (algorithm)

O Trained staff (human experts)

Decision-making situation 6: Weather forecasts

You are the manager of a news site and have to decide whether your 100,000 daily weather forecasts for various cities are carried out solely by a specialized computer program (algorithm) or exclusively by experienced meteorologists (human experts). The task will be considered to have been successfully completed when the temperatures forecast the previous day do not diverge by more than 1 degree Celsius from the actual temperature. A wrong decision could mean that the readers of the forecasts do not dress suitably for the weather.

I choose: O Specialized computer program (algorithm)

O Experienced meteorologists (human experts)

Appendix D: Determination of the random event with the aid of a lottery

Figure A-1: Pack of cards in the selection of the algorithm

Pack of cards in the selection of the algorithm: seven cards with the event +€4 and three cards with the event €±0.

Figure A-2: Pack of cards in the selection of the human expert

Pack of cards in the selection of the human expert: six cards with the event +€4 and four cards with the event €±0.

Chapter IV

Comparing different kinds of influence on an algorithm in its forecasting process and their impact on Algorithm Aversion

Co-authored by Zulia Gubaydullina, Marco Lorenz, and Markus Spiwoks
Contribution Jan René Judek: 45%

Published:
Businesses, Vol. 2, Issue 4, 448-470. (Oct 2022)
https://doi.org/10.3390/businesses2040029

Also published as:

Discussion Paper in Sonderforschungsgruppe Institutionenanalyse (sofia), No. 21-6, Darmstadt, June 2021.

Working Paper in Wolfsburg Working Papers (WWP), Ostfalia University of Applied Sciences, No. 21-04, Wolfsburg, June 2021.

Abstract

Although algorithms make more accurate forecasts than humans in many applications, decision-makers often refuse to resort to their use. In an economic experiment, we examine whether the ex-tent of this phenomenon known as algorithm aversion can be reduced by granting decision-makers the possibility to exert an influence on the configuration of the algorithm (an influence on the algorithmic input). In addition, we replicate the study carried out by Dietvorst et al. (2018). This shows that algorithm aversion recedes significantly if the subjects can subsequently change the results of the algorithm – and even if this is only by a few percent (an influence on the algorithmic output). The present study confirms that algorithm aversion is reduced significantly when there is such a possibility to influence the algorithmic output. However, exerting

an influence on the algorithmic input seems to have only a limited ability to reduce algorithm aversion. A limited opportunity to modify the algorithmic output thus reduces algorithm aversion more effectively than having the ability to influence the algorithmic input.

Keywords

Algorithm aversion; Technology adoption; Human in the loop; Human-computer interaction; Experiment and behavioral economics.

JEL Classification

D81; D91; G17; G41; O33.

1 Introduction

In many domains, the adoption of algorithmic decision making (ADM) has helped complete tasks more accurately, safely, and profitably (Alexander, Blinder & Zak, 2018; Youyou, Kosinski & Stillwell, 2015; Beck et al., 2011, Dawes, 1979; Meehl, 1954). In contrast to the recent successes, algorithm aversion is a major barrier to the adoption of ADM systems (Burton, Stein & Jensen, 2020; Dietvorst, Simmons & Massey, 2015). If effective means can be found to overcome algorithm aversion and enable the implementation of powerful algorithms, quality of life and prosperity can be enhanced (Castelo, Bos & Lehmann, 2020; Dietvorst, Simmons & Massey, 2018; Logg, 2017). Allowing decision makers to influence an algorithm and its prediction process has been shown to influence algorithm aversion (Dietvorst, Simmons & Massey, 2018). However, it is still largely unclear which ways of influencing an algorithm are appropriate and in which step of the process decision makers should be involved to effectively reduce their aversion. This study aims to fill this research gap by investigating different ways of influencing an algorithm and their effects on algorithm aversion in the context of an economic experiment. We draw on the research design of Dietvorst, Simmons and Massey (2018), but also extend their work by introducing a novel method for influencing an algorithm and testing its effectiveness.

Businesses throughout the world are driving the digital transformation. Progress in the field of ADM has wide-ranging effects on our everyday lives and is bringing about fundamental changes in all fields of human life (Mahmud et al., 2022; Nagtegaal, 2021; Fayyaz et al., 2020). ADM systems make a considerable contribution towards tasks being completed faster and above all more cheaply (Upadhyay & Khandelwal, 2018). In addition, algorithms can better the performance of humans (from lay persons to experts) in a multitude of areas and make more accurate predictions, including the following examples: forecasts on the performance of employees (Highhouse, 2008), the likelihood of ex-prisoners re-offending (Wormith & Goldstone, 1984), or in making medical diagnoses (Beck et al., 2011; Gladwell, 2007; Grove et al., 2000; Dawes, Faust & Meehl, 1989; Adams et al., 1986).

Nevertheless, in certain fields there is a lack of acceptance for the actual use of algorithms because subjects have reservations about them. This phenomenon, which is known as algorithm aversion, refers to the lack of trust in ADM systems which arises in subjects as soon as they recognize that the algorithms sometimes make inaccurate predictions (Jussupow, Benbasat & Heinzl, 2020; Prahl & Van Swol, 2017; Dietvorst, Simmons & Massey, 2015). We therefore focus on the issue of how algorithm aversion can be reduced and how the level of acceptance of algorithms can be increased.

In recent years, scholars have explored many ideas for reducing algorithm aversion. Some have proven effective, but others have not. For example, decision making on behalf of others (Filiz et al., 2022) has been shown to have no significant effect on algorithm aversion. Moreover, naming an algorithm has actually been shown to decrease willingness to use it (Hodge, Mendoza & Sinha, 2021).

On the other hand, considering predictions of experts using an algorithm as an input variable for the algorithm has been shown to increase willingness to use it (Kawaguchi, 2021). Moreover, it has been observed that a more precise representation of the algorithmic output (Kim, Giroux & Lee, 2021) and additional information about the process of an algorithm (Ben David, Resheff & Tron, 2021) decrease algorithm aversion. In particular, the latter implies that subjects like to exert some kind of influence on an algorithm. However, many of these tested means of reducing aversion are costly and difficult to implement in real-world scenarios, which is why it remains an important task to continue the research.

Most notably, Dietvorst, Simmons, and Massey (2018) demonstrate a way to significantly reduce algorithm aversion. In their experiment, the subjects can either choose an algorithm or make their own forecasts. Some of the subjects are – if they choose to use an algorithm – allowed to subsequently change the preliminary forecast of the algorithm by a few percent up or down (we describe this in our study as an opportunity to influence the 'algorithmic output'). When they have this opportunity to make retrospective changes to the forecasts, significantly more subjects are prepared to consult the algorithm for their forecasts than otherwise. However, the impact of a slight influence on the configuration of the algorithm (an influence on the algorithmic input) has not been the focus of research, a gap that the present study aims to fill.

As long as the subjects are able to change the results of the algorithm (i.e., they have an influence on the algorithmic output), algorithm aversion can be significantly reduced. Decisions in favor of an algorithm are made more frequently if the users retain an element of control over it, whereby the extent to which they are able to modify the algorithm is irrelevant. Furthermore, users who can make slight modifications report that they are no less content with the forecasting process than users who can make unlimited changes. To sum up, users will deploy algorithms more often when they have the final say in how they deal with them (Dietvorst, Simmons & Massey, 2018). So, is it crucial for lowering algorithm aversion that users are given an opportunity to influence the algorithmic output, or can algorithm aversion be generally reduced by providing a way of influencing the forecasting process?

Human decision-makers want to influence algorithms instead of being at the mercy of their calculations (Honeycutt, Nourani & Ragan, 2020; Stumpf et al., 2008). In other words, decision-makers need partial control over an algorithm in order to make a decision in favor of its use. Having real or at least perceived control over the decisions to be made satisfies the psychological needs and personal interests of users (Colarelli & Thompson, 2008). This feeling of control can arise either via a real understanding of the efficiency of an algorithm, or via adaptations to the algorithmic decision-making process which have little or no influence on the functioning or level of performance of an algorithm (Burton, Stein & Jensen, 2020). In other words, if a user is granted control over decisions, this leads to a higher level of acceptance: if a recommendation algorithm for hotel rooms is used which only recommends hotel rooms based on the person's previous search and purchasing behavior, the offers made are less readily accepted. However, if less than ideal

offers are included, levels of acceptance of the algorithm improve (Taylor, 2017). Participation in the decision-making process, or a belief that one can influence the decision-making process, can contribute towards the user exhibiting greater trust in a decision (Landsbergen et al., 1997).

Nolan and Highhouse (2014) argue that allowing subjects to modify mechanical prediction practices may enhance their perception of autonomy and thus their intentions to use them. In order to expand our understanding of algorithm aversion, we grant the subjects the opportunity to interact with an algorithm not only by modifying its predictions afterwards, but also, adding to the existing literature, by giving them an influence on the weighting of the algorithm's input variables. Analogous to the influence on the algorithmic output (both in the present study and in Dietvorst, Simmons & Massey, 2018), we keep the extent of the subjects' intervention in the algorithmic input small. In this way the algorithm can almost reach its maximum level of performance; however, this minor intervention could be of great significance in overcoming algorithm aversion (cf. Burton, Stein & Jensen, 2020). The present study is the first one to examine whether the opportunity to adjust the weighting of an algorithm's input factors has an effect on its acceptance. In this study, it is observed whether influencing the algorithmic input can contribute towards a reduction of algorithm aversion in the same way as influencing the algorithmic output does.

The economic experiment extends our understanding of algorithm aversion. As in previous studies, subjects do not behave at all like homo economicus. However, their algorithm aversion can be mitigated. The ability to adjust algorithm output significantly increases willingness to use it. The ability to adjust algorithmic input does not work to the same extent. We therefore advise managers dealing with algorithms to create means of influencing algorithmic output to get their customers to use algorithms more often.

2 Materials and Methods

Previous research indicates that economic agents interacting with ADM systems exhibit algorithm aversion and are reluctant to use them (for a synoptic literature review, see Jussupow, Benbasat & Heinzl, 2020; Burton, Stein & Jensen, 2020). This behavior of not relying on an algorithm persists even when an algorithm would be more competent in fulfilling a task than other available alternatives (Berger et al., 2021; Kawaguchi, 2021; Burton, Stein & Jensen, 2020; Efendić, Van de Calseyde & Evans, 2020; Dietvorst, Simmons & Massey, 2018; Dietvorst, Simmons & Massey, 2015). For instance, economic agents are less likely to rely on share price forecasts when they have been drawn up by an algorithm instead of a human expert, which shows the phenomenon of algorithm aversion in the field of share price forecasts (Önkal et al., 2009). Other economic experiments examine the perceived task objectivity and the human-likeness of an algorithm in the context of stock index forecasting and show that the task objectivity affects the willingness to use algorithms with different human-likeness (Castelo, Bos & Lehmann, 2020). The interaction of humans and algorithms is not only a subject in the field of share price

forecasts, but also linked to robo advisors in the financial market research (Filiz et al., 2022).

The fact that algorithms can make more accurate predictions than human forecasters has already been shown on numerous occasions (Grove et al., 2000; Dawes, 1979; Meehl, 1954). Thus, it is key to find ways to mitigate algorithm aversion so that economic agents can arrive at more successful and accurate forecasts. Algorithm aversion can be reduced by providing the opportunity to modify the algorithmic output, even when the possibilities for modification are severely limited (Dietvorst, Simmons & Massey, 2018).

In their literature review, Burton, Stein and Jensen (2020) pose the question of whether the reduction of algorithm aversion by the modification of the algorithmic output can also be achieved by a modification of the algorithmic input. Even the illusion of having the freedom to act and make decisions could be a possible solution to overcome algorithm aversion (Burton, Stein & Jensen, 2020). Users who interact with algorithms often receive their advice from a black box whose workings are a mystery to them. They thus develop theories about which kinds of information an algorithm uses as input and how this information is exactly processed (Logg, Minson & Moore, 2019). According to Colarelli and Thompson (2008), users need to at least have the feeling that they can exercise a degree of control in order to increase the acceptance of algorithms. This feeling of control can either come from a genuine understanding of how an algorithm works or by making modifications to the algorithmic decision-making process. Whether a genuine influence is exerted on the way the algorithm actually functions is not important here. It is only necessary to allow the users to have real or perceived control over decision-making in order to satisfy their need for a feeling of control (Colarelli & Thompson, 2008).

Kawaguchi (2021) has taken a look at how adding an input variable - in this case the predictions made by the subjects - to an algorithm's forecasting process influences algorithm aversion. We draw on this approach and examine how an opportunity to influence the algorithmic input affects the willingness to use an algorithm. We give our subjects the possibility to influence the weighting of an input factor the algorithm uses for its predictions. In this way we are testing an alternative approach to the reduction of algorithm aversion without influencing the algorithmic output. Since modification of the algorithmic output can also have a negative overall effect on forecasting performance, it is examined whether influencing the weighting of an input factor reduces algorithm aversion without allowing human modification of the algorithmic output. We do not want to deceive the subjects and thus give them – in the form of this input factor – the opportunity to exert an actual influence on the configuration of the forecasting computer. In this way, the subjects are given freedom to act in a limited way, which actually leads to slight differences in how the algorithm works. Thus, we address the issue of whether a general possibility to influence the algorithmic process is sufficient to reduce algorithm aversion, or whether an opportunity to influence the results themselves is necessary. We thus examine whether an opportunity to influence the weighting of the input variables of the algorithm (algorithmic input) can contribute towards a similar decrease in algorithm aversion as the opportunity to influence the algorithmic output.

To validate our results in light of previous research and to strengthen our findings, we first replicate the possibility of severely limited influence on algorithmic output (Dietvorst, Simmons & Massey, 2018). We determine whether this measure can also contribute to a reduction of algorithm aversion in the domain of share price forecasts when a choice is made between an algorithm and a subject's own forecasts. Hypothesis 1 is therefore: The proportion of decisions in favor of the algorithm will be higher when there is a limited possibility to influence the algorithmic output than when no influence is possible. Hence, null hypothesis 1 is: The proportion of decisions in favor of the algorithm will not be higher when there is a limited possibility to influence the algorithmic output than when no influence is possible.

Other studies suggest that an influence on the input of an algorithm may also reduce the extent of algorithm aversion (Kawaguchi, 2021; Jung & Seiter, 2021; Burton, Stein & Jensen, 2020; Nolan & Highhouse, 2014). In order to examine whether the possibility to influence the weighting of an algorithm's input variables can have an effect on the willingness to use the algorithm, and thus on algorithm aversion, without the negative effects on performance of influencing algorithmic output, we formulate hypothesis 2 as follows: The proportion of decisions in favor of the algorithm will be higher when there is a limited possibility to influence the algorithmic input than when no influence is possible. Null hypothesis 2 is therefore: The proportion of decisions in favor of the algorithm will not be higher when there is a limited possibility to influence the algorithmic input than when no influence is possible.

In order to answer our research question, an economic experiment is carried out between 17-27 March 2021 in the Ostfalia Laboratory of Experimental Economic Research (OLEW) with students of the Ostfalia University of Applied Sciences in Wolfsburg. In 51 sessions, a total of 157 subjects take part in the experiment. 118 subjects (75.16%) are male and 39 subjects (24.84%) are female. The subjects are distributed across the faculties as follows: 66 subjects (42.04%) study at the Faculty of Vehicle Technology, 56 subjects (35.67%) at the Faculty of Business, 9 subjects (5.73%) at the Faculty of Health Care and a further 26 subjects (16.56%) at other faculties based at other locations of the Ostfalia University of Applied Sciences. Their average age is 23.6 years.

The experiment is programmed with z-Tree (cf. Fischbacher, 2007). In the OLEW there are twelve computer workplaces. However, only a maximum of four are used per session. This ensures that in line with the measures to contain the Covid-19 pandemic a considerable distance can be maintained between the subjects. The workplaces in the laboratory are also equipped with divider panels, which makes it possible to completely separate the subjects from each other. The experiments are constantly monitored by the experimenter so that communication between the subjects and the use of prohibited aids (such as smartphones) can be ruled out. Overall a total of 51 sessions with a maximum of four subjects per session are carried out. A session lasts an average of 30 minutes.

In our study, the subjects are asked to forecast the exact price of a share in ten consecutive periods (Appendix A). Here, the price of the share is always the result of four

influencing factors (A, B, C and D) which are supplemented by a random influence (Ɛ) (see Filiz et al., 2021; Filiz, Nahmer & Spiwoks, 2019; Meub et al., 2015; Becker, Leitner & Leopold-Wildburger, 2009). First of all, the subjects are familiarized with the scenario and are informed that the influencing factors A, C and D have a positive effect on the share price. This means that - other things being equal - when these influencing factors rise the share price will also rise. The influencing factor B, on the other hand, has a negative effect on the share price. This means that - other things being equal - when the influencing factor B rises, the share price will fall (Table 1). In addition, the subjects are informed that the random influence (Ɛ) has an expected value of zero. However, the random influence can lead to larger or smaller deviations from the share price level which the four influencing factors would suggest.

Table 1: Influencing factors in the formation of the share price

Influencing factor	Influence	Strength of the influence
A	Positive	Strong
B	Negative	Strong
C	Positive	Strong
D	Positive	Medium

The subjects are informed of the four influencing factors before each of the ten rounds of forecasting. In addition, they always receive a graphic insight into the historical development of the share price, the influencing factors and the random influence in the last ten periods. In this way, the subjects can recognize in a direct comparison how the levels of the four influencing factors have an effect on the share price during the individual rounds of forecasting. Through test questions we ensure that all subjects have understood this (Appendix B).

Table 2: Performance-related payment for the forecasts

Deviation of the forecast from the actual share price	Payment for the forecast
€0 ≤ $\lvert K_t - P_t \rvert$ ≤ €5	€1.20
€5 < $\lvert K_t - P_t \rvert$ ≤ €10	€0.90
€10 < $\lvert K_t - P_t \rvert$ ≤ €15	€0.60
€15 < $\lvert K_t - P_t \rvert$ ≤ €20	€0.30
$\lvert K_t - P_t \rvert$ > €20	€0.00

Whereby K_t = share price at the point of time t, P_t = forecast at the point of time t.

The payment structure provides for a fixed show-up fee of €4 and a performance-related element. The level of the performance-related payment is dependent on the precision of the individual share price forecasts, whereby the greater the precision of the forecasts, the higher the payment (Table 2). The subjects can thus obtain a maximum payment of €16 (€4 show-up fee plus €12 performance-related payment from ten rounds of forecasting).

In order to help them make the share price forecasts, a forecasting computer (algorithm) is made available to the subjects. The subjects are informed that in the past the share price forecasts of the forecasting computer have achieved a payment of at least €0.60 per forecast in 7 out of 10 cases. The subjects are thus aware of the fact that the algorithm they are using does not function perfectly. In order to make its forecasts, the algorithm uses the information which it has been given on the fundamental influencing factors, the direction and strength of the influence and the random influence (ε) in a way that maximizes the accuracy and thus the expected payoff. In this way, however, it by no means achieves 'perfect' forecasts (for a detailed description of how the algorithm works, see Appendix D). Based on the same information and the historical share prices, the subjects can make their own assessments. They would, however, be wrong to assume that they can outperform the algorithm in this way. Following the suggestions of the algorithm would thus seem to be the more sensible option. Before making their first share price forecast, the subjects make a one-off decision on whether they wish to base their payment for the subsequent ten rounds of forecasting on their own forecasts or on those made by the forecasting computer. Our set-up is oriented towards that used in the study carried out by Dietvorst, Simmons and Massey (2018). Algorithm aversion is thus modeled as the behavior of not choosing an ADM system that would increase subjects' payoff.

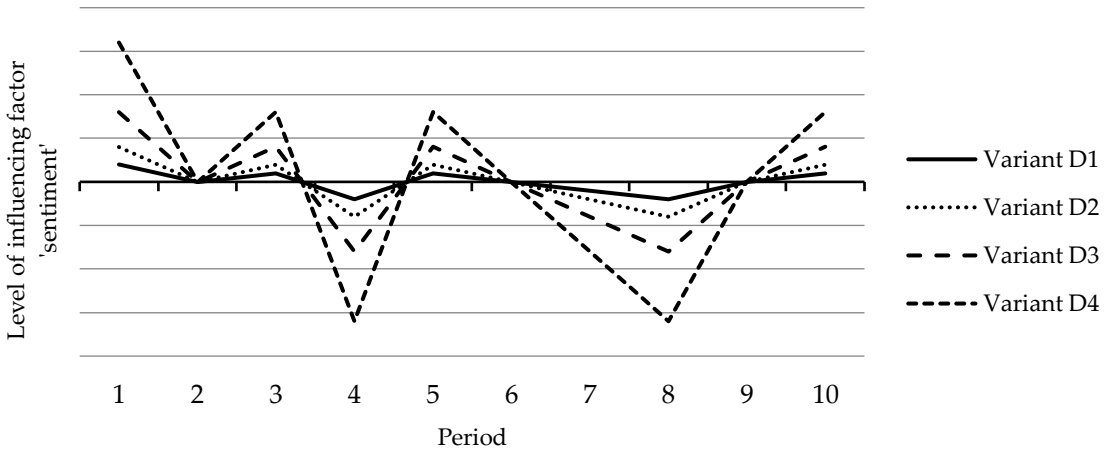
The experiment is carried out in three treatments. The 157 participants are divided up evenly over the three treatments, so that 52 subjects each carry out Treatments 1 and 2, and 53 subjects carry out Treatment 3. The distribution of the subjects among the three treatments has similarities to their distribution among the faculties as well as to their gender. The study uses a between-subjects design: each subject is assigned to only one treatment and encounters the respective decision-making situation. In Treatment 1 (no opportunity to influence the algorithm), the subjects make the decision (once only) whether they want to use their own share price forecasts as the basis for their payment or whether they want to use the share price forecasts made by the forecasting computer. Even if the subjects choose the algorithm for determining their bonus, they have to make their own forecasts. In this case, their payoff only depends on the algorithm's forecasts, not on the forecasts made by the subjects themselves. The obligation to submit one's own forecasts even when choosing the algorithm is based on the study by Dietvorst, Simmons and Massey (2018). Regardless of this decision, the subjects make their own forecasts without having access to the forecast of the algorithm (Figure C-1 in Appendix C).

With Treatment 2 (opportunity to influence the algorithmic output), we intend to replicate the results of Dietvorst, Simmons and Massey (2018). To this end, the subjects make the decision (once only) whether they solely want to use their own share price

forecasts as the basis for their payment or whether they solely wish to use the share price forecasts made by the forecasting computer (which, however, can be adjusted by up to +/- €5) as the basis for their performance-related payment. The algorithmic forecast is only made available to the subjects if they decide in favor of the forecasting computer (Figure C.2).

In Treatment 3, we introduce the opportunity to influence the configuration of the algorithm (algorithmic input). Before handing in their first share price forecast, the subjects again make the decision (once only) whether they want to solely use their own share price forecasts as the basis for their performance-related payment or whether they want to solely use the share price forecasts made by the forecasting computer. If they decide in favor of the share price forecasts of the forecasting computer, the subjects receive a one-off opportunity to influence the configuration of the algorithm (Figure C.3). To this end, they are given a more detailed explanation. The algorithm uses data on four different factors which influence the formation of the share price (A, B, C and D). The last of these four influencing factors is identified as the sentiment of capital market participants and can be considered to various extents by the forecasting computer. To do so, the subjects can choose from four different levels. Whereas variant D1 attaches relatively little importance to sentiment, the extent to which sentiment is considered in the other variants increases continuously and is relatively strong in variant D4 (Figure 1).

Figure 1: Level of the influencing factor 'Sentiment of capital market actors'



Subjects who decide to use the forecasting computer in Treatment 3 and thus receive the opportunity to influence the configuration of the algorithm have a one-off chance to change the weighting of the input variable D of the algorithm. This occurs solely by means of their choice of which degree of sentiment should be taken into account (variant D1, D2, D3 or D4).

3 Results

The results show that the various possibilities to influence the forecasting process led to different decisions on the part of the subjects. In Treatment 1 (no influence possible), 44.23% of the subjects opt for the use of the algorithm. The majority of the subjects here (55.77%) put their faith in their own forecasting abilities. In Treatment 2 (opportunity to influence the algorithmic output) on the other hand, 69.23% of the subjects decide to use the forecasting computer and 30.77% of the subjects choose to use their own forecasts. In Treatment 3 (opportunity to influence the algorithmic input), 58.49% of the subjects decide to use the forecasting computer and 41.51% of the subjects choose to use their own forecasts (Figure 2).

Figure 2: Comparison of the decisions in favor of the algorithm or the subjects' own forecasts per treatment

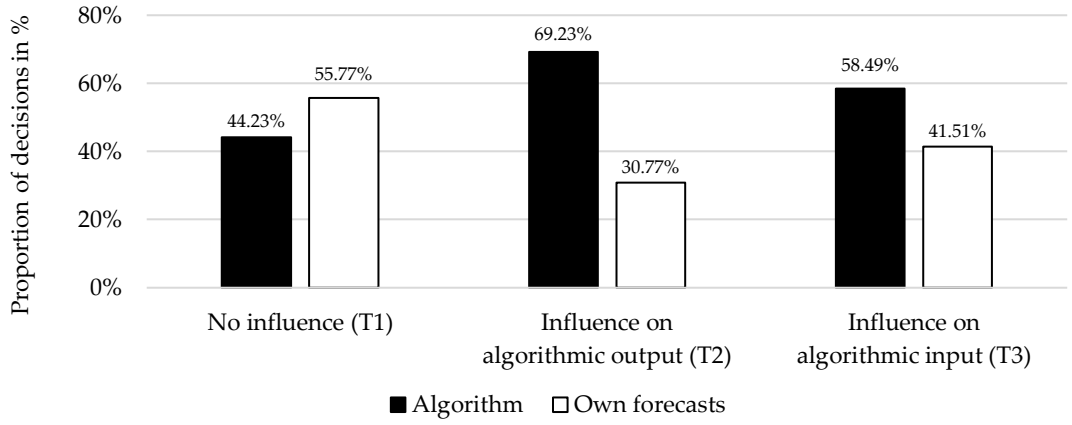
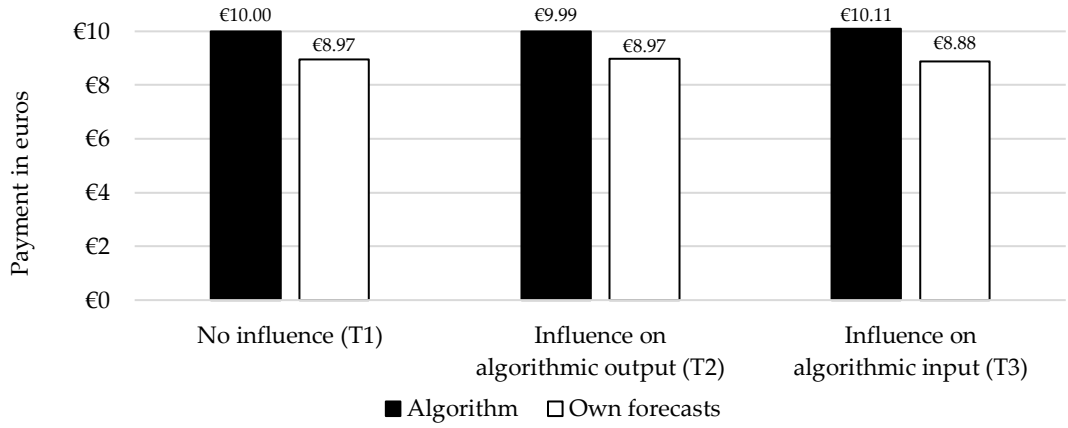


Figure 3: Average payment in the three treatments depending on the strategy chosen when making the forecasts (own forecast or delegation to the algorithm)



On average across all three treatments, the subjects obtain a payment of €9.57. However, there are large differences in the amounts of the payment depending on the strategy chosen. Subjects who choose their own forecasts achieve an average total

payment of €8.94. When the algorithm is chosen, the average payment in all three treatments is between €9.99 and €10.11 (Figure 3). The Wilcoxon rank-sum test shows that the average payment – regardless of the treatment – is significantly higher if the algorithm is used as the basis of the forecasts (T1: $z = 4.27$, $p \leq 0.001$; T2: $z = 3.25$, $p \leq 0.001$; T3: $z = 5.27$, $p \leq 0.001$). No matter which treatment is involved, it is thus clearly in the financial interests of the subjects to put their faith in the algorithm. The algorithm consistently outperforms human judgment, yet, across all treatments, 42.68% of the subjects refrain from using it. In our study too, the phenomenon of algorithm aversion is thus evident in the field of share price forecasts (Önkal et al., 2009; Castelo, Bos & Lehman, 2020).

We perform the Chi-square test on subject's decisions between the algorithm and their own forecasts among the individual treatments. Whereas in Treatment 1 a total of 44.23% of the decisions are in favor of the algorithm, 69.23% of the subjects who can make changes to the algorithmic output (Treatment 2) decide to use the forecasting computer (χ^2 ($N = 104$) = 6.62, $p = 0.010$). Null hypothesis 1 thus has to be rejected; the opportunity to modify the algorithmic output by up to +/- €5 leads to the subjects selecting the forecasting computer significantly more frequently to determine their payment.

When subjects are given the opportunity to influence the algorithmic input (Treatment 3), the majority of the subjects (58.49%) choose to use the forecasting computer (χ^2 ($N = 105$) = 2.14, $p = 0.144$). Nevertheless, null hypothesis 2 is not rejected. The possibility to influence the algorithmic input (via the extent to which the influencing factor D is taken into account) does not lead to the subjects selecting the forecasting computer significantly more often as the basis for their performance-related payment.

In each treatment, there are between 52 and 53 participants, leading to a total of 157 participants. The 67 subjects who, regardless of which treatment they are in, use their own forecasts as the basis of their payment, diverge by an average of €18.28 from the actual share price and thus achieve an average bonus of €0.49 per round of forecasting. The 90 subjects who decide to use the forecasts of the forecasting computer exhibit a lower average forecasting error independently of which treatment they are in. The average bonus and the average payment of the subjects who use the forecasting computer are also higher than that of subjects who rely on their own forecasting abilities. Because of the different ways in which the algorithm can be influenced, the average forecast error, average bonus per round, and average total payment also vary between treatments for those subjects who rely on the ADM (Table 3).

In Treatment 2, the subjects are given the opportunity to adapt the algorithmic output in each round of forecasting by up to +/- €5. The subjects do not fully exploit the scope granted to them to exert an influence on the algorithm and make an average change to the algorithmic forecast of €2.11. In Treatment 3 the subjects are given a one-off opportunity via the influencing factor D (sentiment) to exert an influence on the configuration of the algorithm (input). Eight subjects select variant D1, which takes sentiment into account to a minor extent. Eleven subjects choose to take sentiment into account to a moderate extent, seven to a considerable extent, and five to a great extent.

Table 3: Performance of the subjects in relation to their chosen strategy when making their forecasts (own forecasts or delegation to the algorithm)

	n	Ø Forecast error [in €]*	Ø Bonus per round [in €]	Ø Total payment [in €]
Own forecasts	67	18.2776	0.4939	8.94
Forecasts by the algorithm without the opportunity to influence it *(Treatment 1)*	23	13.4000	0.6000	10.00
Forecasts by the algorithm with an opportunity to influence the output *(Treatment 2)*	36	13.5167	0.5992	9.99
Forecasts by the algorithm with an opportunity to influence the input *(Treatment 3)*	31	13.2968	0.6106	10.11
Total	157	15.4879	0.5566	9.57

* Ø Deviation between the forecasted share price and the actually occurring share price

If the results are viewed in isolation, a similar picture is revealed. Regardless of whether subjects used their own forecasts or the forecasts of the forecasting computer to determine their payment, the average forecast error in Treatment 1 (no influence possible) is higher than in the other two treatments, which offer the subjects the opportunity to influence the algorithm. Whereas the forecasts in Treatment 1 deviate by an average of €16.18 from the resulting share price, the average forecast error in Treatment 2 is €15.14 and €15.15 in Treatment 3. That those subjects who are given the opportunity to influence the algorithm are more successful is shown by their average bonus and higher average overall payment (Table 4).

Table 4: Comparison of the performance of the subjects across all three treatments

	n	Ø Forecast error [in €]*	Ø Bonus per round [in €]	Ø Total payment [in €]
No influence possible *(Treatment 1)*	52	16.1788	0.5423	9.42
Influence on the algorithmic output *(Treatment 2)*	52	15.1442	0.5677	9.68
Influence on the algorithmic input *(Treatment 3)*	53	15.1472	0.5598	9.60
Total	157	15.4879	0.5566	9.57

* Ø Deviation between the forecasted share price and the actually occurring share price

4 Discussion

Algorithm aversion is characterized by the fact that it mostly occurs when algorithms recognizably do not function perfectly and prediction errors occur (Dietvorst, Simmons & Massey, 2015). Even when it is recognizable that the algorithm provides significantly more reliable results than humans (lay persons as well as experts), many subjects are still reluctant to trust the algorithm (Dietvorst, Simmons & Massey, 2018). Due to the advancing technological transformation and the increasing availability of algorithms, it is inevitable to enhance the understanding of algorithm aversion and to study ways to mitigate it.

Previous research had shown that giving users the ability to influence algorithmic output in terms of minimal adjustments to the forecasts contributes to a significant reduction of algorithm aversion (Dietvorst, Simmons & Massey, 2018). This groundbreaking finding is confirmed in the context of share price forecasts in the present paper. As shown in our introduction section, as a reaction to this interesting concept, a rich literature that focuses on further ways to mitigate algorithm has emerged (Filiz et al., 2022; Hodge, Mendoza & Sinha, 2021; Kim, Giroux & Lee, 2021; Ben David, Resheff & Tron, 2021).

Most noteworthy in the context of our research, Kawaguchi (2021) examined the effect of having an algorithm select its users' individual forecasts as an additional input variable, and Jung and Seiter (2021) examined the effect of having subjects self-select the variables an algorithm should consider. Both studies report significant changes in the extent of algorithm aversion due to the manipulation on the input they investigate. The results from the present study are in line with previous findings regarding influence on an algorithm's output (Dietvorst, Simmons & Massey, 2018), but point in a different direction regarding influence on algorithm's input (Kawaguchi, 2021; Jung & Seiter, 2021).

The algorithm used in our study does not give perfect forecasts, and if there are no opportunities to influence the algorithm's decision-making process, the majority of users choose not to use the forecasting computer. But the ability to influence algorithmic output (replicated from Dietvorst, Simmons & Massey, 2018) leads subjects to use the algorithm significantly more often compared to the control treatment, even when the amount of adjustment allowed in the process is relatively small (T1 vs. T2). By using the algorithm more frequently, the subjects also enhance their financial performance.

Our study essentially contributes to the scientific discourse by testing the possibility of influence on the weighting of the variables an algorithm uses in its forecasting process (the algorithmic input). Even though the financial performance is slightly enhanced, there is no significantly higher willingness to use the algorithm compared to the control treatment when there is a possibility to influence the input (T1 vs. T3). The assumption of Nolan and Highhouse (2014) that intention to use a forecasting aid can be improved by the possibility to influence its configuration is not confirmed. We also cannot confirm Burton, Stein and Jensen's (2020) conjecture that changing an ADM's input mitigates algorithm aversion, at least for the consideration of a limited influence on the weighting

of the algorithmic input. The differences in our results compared to Kawaguchi (2021) and Jung and Seiter (2021) are likely due to the fact that the extent of the subjects' influence on the input of the algorithm is much smaller in the present study. Another crucial difference is that the input factors of the algorithm in our study are predetermined and only the weighting can be changed.

We examine whether major reductions in algorithm aversion are due to the fact that the subjects can exercise an influence on the process of algorithmic decision-making in general, or only because they can influence its forecasts. We expand the research about algorithm aversion by showing that a general opportunity to influence an algorithm is obviously not sufficient to significantly reduce algorithm aversion. Subjects want to retain control over the results and to have the final say in the decision-making process, even if this intervention is limited by considerable restrictions. Since no significantly higher willingness to use can be achieved by adjusting the input, we recommend focusing on the output of algorithms in order to identify further possibilities for mitigating algorithm aversion. It seems to be of considerable relevance at which point in the process of algorithmic decision making an intervention is allowed.

Implications

Nevertheless, our study has interesting implications for real-life situations. The overall financial benefit can be maximized by influencing the algorithmic output. Decision-makers tend to trust an algorithm more if they can keep the upper hand in the decision-making process. This even applies when the possibilities to exert an influence are limited. The average quality of the forecasts is slightly reduced due to the changes made by the decision-maker (Table 3), but this is over-compensated for by a significantly higher utilization rate of the – still clearly superior – algorithm, and in a comparison between the treatments this leads to a higher average total payment (Table 4). The opportunity to influence the algorithmic input has a similar effect with regard to the overall pecuniary benefit. The forecasts made after the subjects have made changes to the algorithm actually exhibit a slightly lower forecast error and a somewhat higher bonus. To a similar degree to which the subjects do not fully take advantage of the opportunity to influence the algorithmic output, they also fail to put their faith in the algorithm. Their average payment is nevertheless significantly higher than that of the subjects who cannot influence the algorithm. From this we conclude for real-world settings that customers should not be involved in formulas or configuration options of algorithms, but rather be given the opportunity to influence the output, for example through override functions, veto rights, emergency stop buttons, etc.

Limitations

Our study also has some limitations which should be noted. We give the subjects a genuine opportunity to influence the algorithmic input. However, we also make it clear in the instructions that the influencing factor D, which can be taken into account to different degrees, only has a moderate influence on the formation of the share price. The influencing factors A, B and C, on the other hand, have a considerable influence. This

circumstance could contribute towards the subjects not developing enough trust in their opportunity to influence the input and thus tending to rely on their own forecasts. In addition, our results were obtained in the context of share price forecasts. The validity of our results for the many other areas of ADM systems has yet to be verified.

Future Research

Future research work may wish to investigate further possibilities to reduce algorithm aversion. This study has again shown that granting subjects the opportunity to influence the algorithmic output can effectively reduce algorithm aversion. However, there is a risk that the forecasting performance of the algorithm can deteriorate as a result of the modifications. For this reason, it is important to examine alternative forms of reducing algorithm aversion. Our study has shown that modifying the algorithmic input to a small extent is only of limited use here. It would be interesting to see what happens when the possible adjustments to the algorithmic input, and thus the perceived control over the algorithm, are greater. In our study, opportunities to influence the algorithmic input cannot reduce algorithm aversion to the same extent as giving subjects the chance to influence the algorithmic output. We therefore recommend that further research be carried out to search for other alternatives to reduce algorithm aversion. One possible approach could be to merely give users the illusion of having control over the algorithmic process. In this way, algorithm aversion could be decreased without a simultaneous reduction of the forecasting quality.

Conclusion

In an economic experiment we examine whether providing a possibility to influence the algorithmic input contributes towards mitigating algorithm aversion. We ask subjects to make forecasts of share prices. In return, they receive a performance-related payment which increases in line with the precision of their share price forecasts. In three treatments the subjects have a forecasting computer (algorithm) available to them which provides different options for influencing the process: In Treatment 1 we do not grant the subjects any opportunity to influence the forecasting process. In Treatment 2 the subjects can influence the algorithmic output, and in Treatment 3 they can influence the algorithmic input. In line with the literature on algorithm aversion, we show that even a considerably limited opportunity to influence the algorithmic output is able to reduce algorithm aversion significantly. However, being able to influence the algorithmic input does not lead to a significant reduction in algorithm aversion. Granting subjects a general possibility to influence the algorithmic decision-making process is therefore not a crucial factor in reducing algorithm aversion. What does lead to a significantly higher rate of using the forecasting computer is the opportunity to influence the algorithmic output. For this reason, further efforts to mitigate algorithm aversion for real-world events should focus on the possibilities of adjusting the algorithmic output.

References

Adams, I., Chan, M., Clifford, P., Cooke, W. M., Dallos, V., Dombal, F. T., Edwards, M., Hancock, D., Hewett, D. J., & McIntyre, N. (1986), Computer aided diagnosis of acute abdominal pain: a multicentre study, *British Medical Journal*, 293(6550), 800-804.

Alexander, V., Blinder, C., & Zak, P. J. (2018), Why trust an algorithm? Performance, cognition, and neurophysiology, *Computers in Human Behavior*, 89(2018), 279-288.

Beck, A., Sangoi, A., Leung, S., Marinelli, R. J., Nielsen, T., Vijver, M. J., West, R., Rijn, M. V., & Koller, D. (2011), Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival, *Science Translational Medicine*, 3(108), 108-113.

Becker, O., Leitner, J., & Leopold-Wildburger, U. (2009), Expectation formation and regime switches, *Experimental Economics*, 12(3), 350-364.

Ben David, D., Resheff, Y. S., & Tron, T. (2021), Explainable AI and Adoption of Financial Algorithmic Advisors: An Experimental Study, *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 390-400.

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2021), Watch me improve—algorithm aversion and demonstrating the ability to learn, *Business & Information Systems Engineering*, 63(1), 55-68.

Burton, J., Stein, M., & Jensen, T. (2020), A Systematic Review of Algorithm Aversion in Augmented Decision Making, *Journal of Behavioral Decision Making*, 33(2), 220-239.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2020), Task-dependent algorithm aversion, *Journal of Marketing Research*, 56(5), 809-825.

Colarelli, S. M., & Thompson, M. B. (2008), Stubborn Reliance on Human Nature in Employee Selection: Statistical Decision Aids Are Evolutionarily Novel, *Industrial and Organizational Psychology*, 1(3), 347-351.

Dawes, R. (1979), The Robust Beauty of Improper Linear Models in Decision Making, *American Psychologist*, 34(7), 571-582.

Dawes, R., Faust, D., & Meehl, P. (1989), Clinical Versus Actuarial Judgment, *Science*, 243(4899), 1668-74.

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.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015), Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err, *Journal of Experimental Psychology*, 144(1), 114-126.

Efendić, E., Van de Calseyde, P. P., & Evans, A. M. (2020), Slow response times undermine trust in algorithmic (but not human) predictions, *Organizational Behavior and Human Decision Processes*, 157, 103-114.

Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A., & Kashef, R. (2020), Recommendation systems: Algorithms, challenges, metrics, and business opportunities, *Applied Sciences*, 10(21), 7748.

Filiz, I., Judek, J. R., Lorenz, M., & Spiwoks, M. (2022), Algorithm Aversion as an Obstacle in the Establishment of Robo Advisors, *Journal of Risk and Financial Management*, 15(8), 353.

Filiz, I., Judek, J. R., Lorenz, M., & Spiwoks, M. (2021), Reducing algorithm aversion through experience, *Journal of Behavioral and Experimental Finance*, 31(5), 100524.

Filiz, I., Nahmer, T., & Spiwoks, M. (2019), Herd behavior and mood: An experimental study on the forecasting of share prices, *Journal of Behavioral and Experimental Finance*, 24, 1-10.

Fischbacher, U. (2007), z-Tree: Zurich toolbox for ready-made economic experiments, *Experimental Economics*, 10(2), 171-178.

Gladwell, M. (2007), Blink: The Power of Thinking Without Thinking, Back Bay Books, New York City.

Grove, W., Zald, D., Lebow, B., Snitz, B., & Nelson, C. (2000), Clinical versus mechanical prediction: A meta-analysis, *Psychological Assessment*, 12(1), 19-30.

Highhouse, S. (2008), Stubborn Reliance on Intuition and Subjectivity in Employee Selection, *Organizational Psychology*, 1(3), 333-342.

Hodge, F. D., Mendoza, K. I., & Sinha, R. K. (2021), The effect of humanizing robo-advisors on investor judgments, *Contemporary Accounting Research*, 38(1), 770-792.

Honeycutt, D., Nourani, M., & Ragan, E. (2020), Soliciting Human-in-the-Loop User Feedback for Interactive Machine Learning Reduces User Trust and Impressions of Model Accuracy, *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1), 63-72.

Jung, M., & Seiter, M. (2021), Towards a better understanding on mitigating algorithm aversion in forecasting: an experimental study, *Journal of Management Control*, 32, 495-516.

Jussupow, E., Benbasat, I., & Heinzl, A. (2020), Why are we averse towards Algorithms? A comprehensive literature Review on Algorithm aversion, *Proceedings of the 28th European Conference on Information Systems (ECIS)*, 1-16.

Kawaguchi, K. (2021), When Will Workers Follow an Algorithm? A Field Experiment with a Retail Business, *Management Science*, 67(3), 1670-1695.

Kim, J., Giroux, M., & Lee, J. C. (2021), When do you trust AI? The effect of number presentation detail on consumer trust and acceptance of AI recommendations, *Psychology & Marketing*, 38(7), 1140-1155.

Landsbergen, D., Coursey, D. H., Loveless, S., & Shangraw, R. (1997), Decision Quality, Confidence, and Commitment with Expert Systems: An Experimental Study, *Journal of Public Administration Research and Theory*, 7(1), 131-158.

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

Logg, J. M. (2017), Theory of Machine: When Do People Rely on Algorithms?, *Harvard Business School Working Paper 17-086*, March 2017.

Mahmud, H., Islam, A. N., Ahmed, S. I., & Smolander, K. (2022), What influences algorithmic decision-making? A systematic literature review on algorithm aversion, *Technological Forecasting and Social Change*, 175, 121390.

Meehl, P. (1954), Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence, University of Minnesota Press, Minneapolis.

Meub, L., Proeger, T., Bizer, K., & Spiwoks, M. (2015), Strategic coordination in forecasting - An experimental study, *Finance Research Letters*, 13(1), 155-162.

Nagtegaal, R. (2021), The impact of using algorithms for managerial decisions on public employees' procedural justice, *Government Information Quarterly*, 38(1), 101536.

Nolan, K. P., & Highhouse, S. (2014), Need for autonomy and resistance to standardized employee selection practices, *Human Performance*, 27(4), 328-346.

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.

Stumpf, S., Sullivan, E., Fitzhenry, E., Oberst, I., Wong, W. K., & Burnett, M. (2008), Integrating rich user feedback into intelligent user interfaces, *Proceedings of the 13th international conference on Intelligent user interfaces*, 50-59.

Taylor, E. L. (2017), Making sense of "algorithm aversion", *Research World*, 2017(64), 57-57.

Upadhyay, A. K., & Khandelwal, K. (2018), Applying artificial intelligence: implications for recruitment, *Strategic HR Review*, 17(5), 255-258.

Wormith, J. S., & Goldstone, C. S. (1984), The Clinical and Statistical Prediction of Recidivism, *Criminal Justice and Behavior*, 11(1), 3-34.

Youyou, W., Kosinski, M., & Stillwell, D. (2015), Computer-based personality judgments are more accurate than those made by humans, *Proceedings of the National Academy of Sciences*, 112(4), 1036-1040.

Appendix A: Instructions for the Experiment

Appendix A.1: Instructions (Treatment 1: no opportunity to influence the algorithm)

The Game

In this game you are requested to make forecasts on the future trend of a share price. You will forecast the price movements of a share (share Z) in 10 periods.

The price of share Z is always the result of four influencing factors (**A**, **B**, **C** and **D**) and a random influence (**ε**). The influencing factors are announced before every round of forecasting. In addition, you receive an insight into the past development of the share price, the influencing factors and the random influence in the last ten periods.

The influencing factors **A**, **C** and **D** have a positive effect on the share price. This means that when these influencing factors rise, the share price will also tend to rise (Table 1).

The influencing factor **B** has a negative effect on the share price. This means that when the influencing factor **B** rises, the share price will tend to fall (Table 1).

Table 1: Influencing factors in the formation of the share price

Influencing factor	Influence	Strength of the influence
A	Positive	Strong
B	Negative	Strong
C	Positive	Strong
D	Positive	Medium

The random influence **ε** has an expected value of 0, but it can lead to smaller or larger deviations of the share price from the level which the influencing factors would suggest.

You can choose whether your own share price forecasts or the share price forecasts of a forecasting computer (algorithm) are used to determine your payment. Regardless of your choice, you will make your own share price forecasts.

You will receive a show-up fee of €4 for participating. In addition, you receive a performance-related payment: the more accurate your share price forecasts are, the higher your payment. For each forecast made, you receive…

- €1.20 in the case of a deviation of a maximum of €5 of the forecast from the actual share price;
- €0.90 in the case of a deviation of a maximum of €10 of the forecast from the actual share price;
- €0.60 in the case of a deviation of a maximum of €15 of the forecast from the actual share price;
- €0.30 in the case of a deviation of a maximum of €20 of the forecast from the actual share price.

In the past, the share price forecasts of the algorithm have achieved a payment of at least €0.60 per forecast in 7 out of 10 cases.

Procedure

After reading the instructions and answering the test questions, you initially choose whether your own share price forecasts or the forecasts of the forecasting computer (algorithm) are used to determine your payment.

Following this, you will see the price history of share Z, the trend of the influencing factors and the trend of the random influence ε in the last ten periods. In addition, you will receive the influencing factors for the next period. You will be asked to forecast the trend of the share price in the next period.

After making your share price forecast you will see the actual price of share Z. Following this, you will hand in your share price forecasts for the next period. A total of ten rounds are played.

You have a time limit of two minutes available for handing in each share price forecast.

Information

- Please remain quiet during the experiment!
- Please do not look at your neighbor's screen!
- Apart from a pen/pencil and a pocket calculator, **no** other aids are permitted (smartphones, smart watches etc.).
- Only use the sheet of white paper issued to you for your notes.

Appendix A.2: Instructions (Treatment 2: opportunity to influence the algorithmic output)

The Game

In this game you are requested to make forecasts on the future trend of a share price. You will forecast the price movements of a share (share Z) in 10 periods.

The price of share Z is always the result of four influencing factors (**A**, **B**, **C** and **D**) and a random influence (**ε**). The influencing factors are announced before every round of forecasting. In addition, you receive an insight into the past development of the share price, the influencing factors and the random influence in the last ten periods.

The influencing factors **A**, **C** and **D** have a positive effect on the share price. This means that when these influencing factors rise, the share price will also tend to rise (Table 1).

The influencing factor **B** has a negative effect on the share price. This means that when the influencing factor **B** rises, the share price will tend to fall (Table 1).

Table 1: Influencing factors in the formation of the share price

Influencing factor	Influence	Strength of the influence
A	Positive	Strong
B	Negative	Strong
C	Positive	Strong
D	Positive	Medium

The random influence **ε** has an expected value of 0, but it can lead to smaller or larger deviations of the share price from the level which the influencing factors would suggest.

- You can choose the basis which is used to determine your payment:
- Either you can forecast the future share price yourself and forego the use of a forecasting computer (algorithm)
- Or you can use the forecasts of the forecasting computer. If you decide to use the forecasting computer's forecasts (algorithm), you are not bound to the exact forecast provided by the computer. You can change the computer's proposal by up to +/- €5.

You will receive a show-up fee of €4 for participating. In addition, you receive a performance-related payment: the more accurate your share price forecasts are, the higher your payment. For each forecast made, you receive…

- €1.20 in the case of a deviation of a maximum of €5 of the forecast from the actual share price;
- €0.90 in the case of a deviation of a maximum of €10 of the forecast from the actual share price;
- €0.60 in the case of a deviation of a maximum of €15 of the forecast from the actual share price;
- €0.30 in the case of a deviation of a maximum of €20 of the forecast from the actual share price.

In the past, the share price forecasts of the algorithm have achieved a payment of at least €0.60 per forecast in 7 out of 10 cases.

Procedure

After reading the instructions and answering the test questions, you initially choose which basis is used to determine your payment. You can forecast the future share prices without the help of the forecasting computer (algorithm). Or you can use the forecasts of the forecasting computer and change them by up to +/- €5.

Following this, you will see the price history of share Z, the trend of the influencing factors and the trend of the random influence ε in the last ten periods. In addition, you will receive the influencing factors for the next period. You will be asked to forecast the trend of the share price in the next period.

After making your share price forecast you will see the actual price of share Z. Following this, you will hand in your share price forecasts for the next period. A total of ten rounds are played.

You have a time limit of two minutes available for handing in each share price forecast.

Information

- Please remain quiet during the experiment!
- Please do not look at your neighbor's screen!

- Apart from a pen/pencil and a pocket calculator, **no** other aids are permitted (smartphones, smart watches etc.).
- Only use the sheet of white paper issued to you for your notes.

Appendix A.3: Instructions (Treatment 3: opportunity to influence the algorithmic input)

The Game

In this game you are requested to make forecasts on the future trend of a share price. You will forecast the price movements of a share (share Z) in 10 periods.

The price of share Z is always the result of four influencing factors (**A**, **B**, **C** and **D**) and a random influence (**ε**). The influencing factors are announced before every round of forecasting. In addition, you receive an insight into the past development of the share price, the influencing factors and the random influence in the last ten periods.

The influencing factors **A**, **C** and **D** have a positive effect on the share price. This means that when these influencing factors rise, the share price will also tend to rise (Table 1).

The influencing factor **B** has a negative effect on the share price. This means that when the influencing factor **B** rises, the share price will tend to fall (Table 1).

Table 1: Influencing factors in the formation of the share price

Influencing factor	Influence	Strength of the influence
A	Positive	Strong
B	Negative	Strong
C	Positive	Strong
D	Positive	Medium

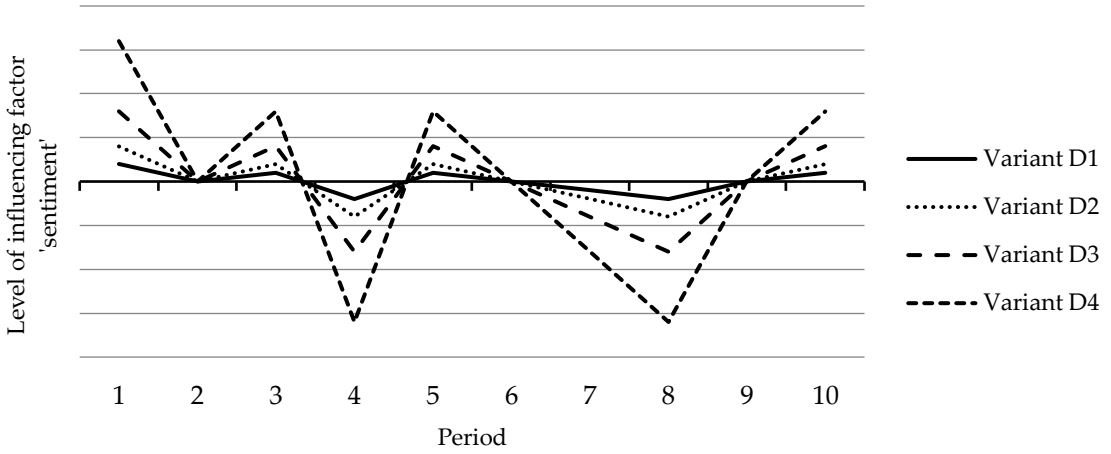
The random influence ε has an expected value of 0, but it can lead to smaller or larger deviations of the share price from the level which the influencing factors would suggest.

You can choose whether your own share price forecasts or the share price forecasts of a forecasting computer (algorithm) are used to determine your payment. Regardless of your choice, you will make your own share price forecasts.

If you decide to use the forecasting computer's forecasts (algorithm), you have the opportunity to influence the design of the algorithm.

As mentioned above, the influencing factor D also has an effect on the formation of the price alongside the influencing factors A, B and C. The influencing factor D is the sentiment of capital market participants. The influencing factor D can be taken into account to differing extents (D1, D2, D3 or D4) (Figure 1). You decide which of these four variants should be taken into account by the forecasting computer (algorithm).

Figure 1: Variants of the influencing factor D (sentiment)



You will receive a show-up fee of €4 for participating. In addition, you receive a performance-related payment: the more accurate your share price forecasts are, the higher your payment. For each forecast made, you receive…

- €1.20 in the case of a deviation of a maximum of €5 of the forecast from the actual share price;
- €0.90 in the case of a deviation of a maximum of €10 of the forecast from the actual share price;
- €0.60 in the case of a deviation of a maximum of €15 of the forecast from the actual share price;
- €0.30 in the case of a deviation of a maximum of €20 of the forecast from the actual share price.

In the past, the share price forecasts of the algorithm have achieved a payment of at least €0.60 per forecast in 7 out of 10 cases.

Procedure

After reading the instructions and answering the test questions, you initially choose whether your own share price forecasts or the forecasts of the forecasting computer (algorithm) are used to determine your payment.

Following this, you will see the price history of share Z, the trend of the influencing factors and the trend of the random influence ε in the last ten periods. In addition, you will receive the influencing factors for the next period. You will be asked to forecast the trend of the share price in the next period.

After making your share price forecast you will see the actual price of share Z. Following this, you will hand in your share price forecasts for the next period. A total of ten rounds are played.

You have a time limit of two minutes available for handing in each share price forecast.

Information

- Please remain quiet during the experiment!
- Please do not look at your neighbor's screen!
- Apart from a pen/pencil and a pocket calculator, **no** other aids are permitted (smartphones, smart watches etc.).
- Only use the sheet of white paper issued to you for your notes.

Appendix B: Test questions

Test question 1: For how many periods should a share price forecast be made?

a) 5.
b) 10. *(correct*)
c) 15.

Test question 2: On which influences is the share price dependent?

a) Influencing factors A and B as well as the random influence.
b) Influencing factors A, B and C as well as the random influence.
c) Influencing factors A, B, C and D as well as the random influence. *(correct)*

Test question 3: Which alternatives do you have when submitting your forecast?

a) I can only submit my own forecasts.
b) I can either submit my own forecasts or use a forecasting computer (algorithm). (*correct)*
c) I can either submit my own forecasts, use a forecasting computer or consult a financial expert.

Test question 4: How much is the payment for a forecast which deviates no more than €15 from the actual price?

a) €1.20.
b) €0.90.
c) €0.60. *(correct)*

Appendix C: Screens

Figure C-1: Screen when submitting one's own forecasts (Treatments 1, 2 and 3)

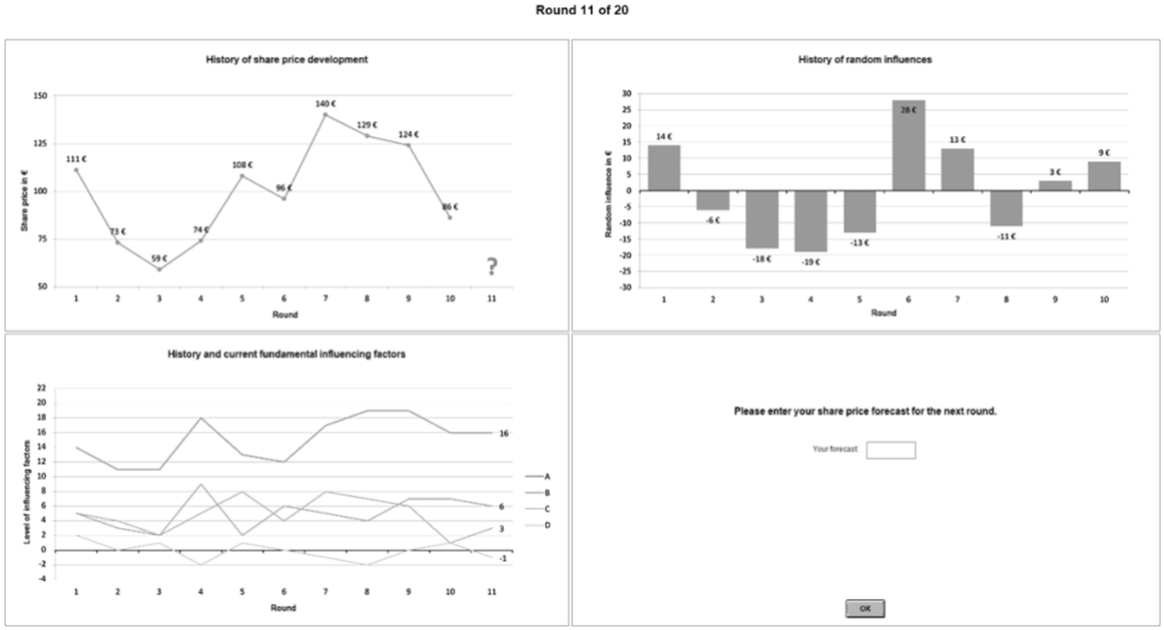


Figure C-2: Screen when influencing the algorithmic output (Treatment 2)

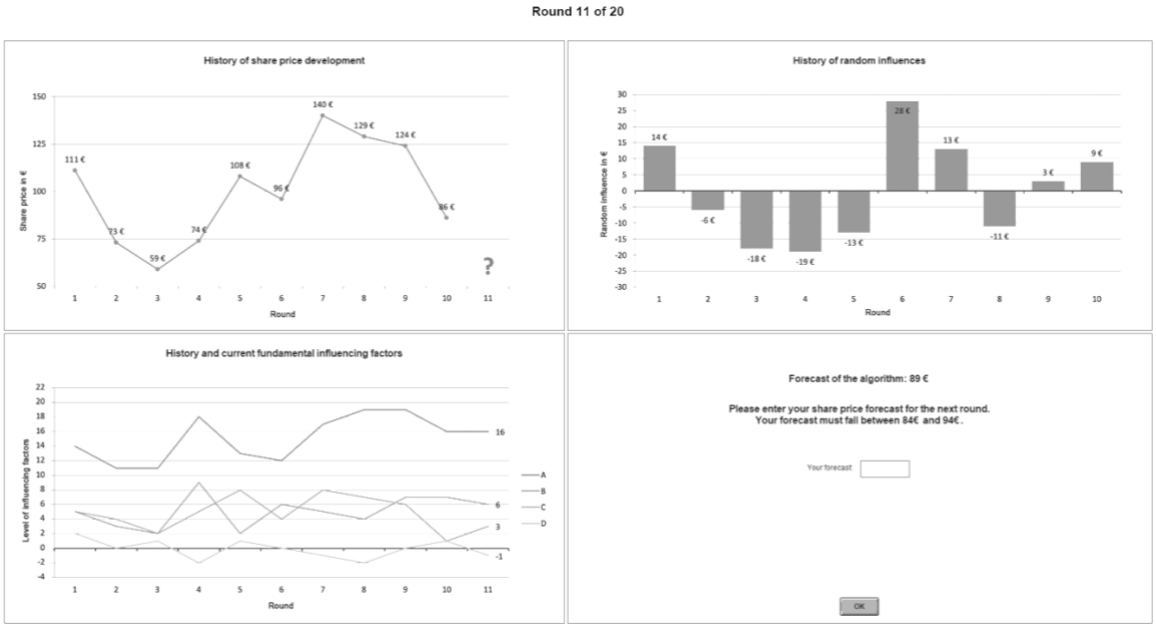
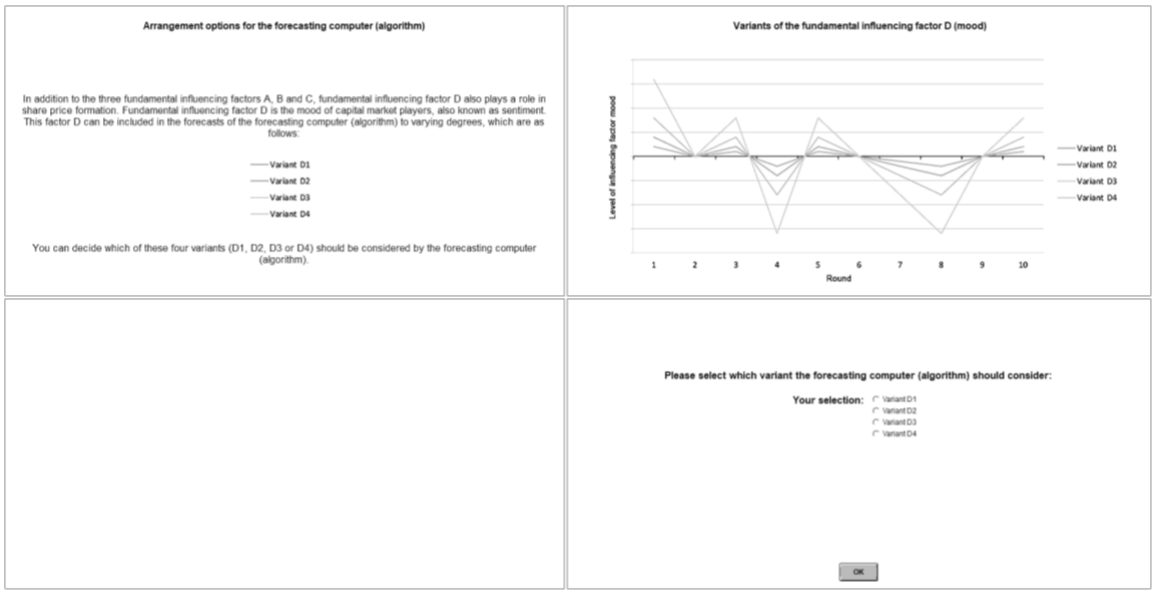


Figure C-3: Screen when influencing the algorithmic input (Treatment 3)



Appendix D: The functioning of the algorithm

The mechanism with which the share price is formed functions as follows:

$$K_t = 7A - 6B + 5C + 2D + \varepsilon \quad (1)$$

The level of the influencing factors A, B, C and D are announced before every round of forecasting. The level of the random influence is not announced. What is known, however, is that the random influence has an expected value of 0. The algorithm used in this experiment is a system that exploits the given information ideally through statistical processes. In every round, the algorithm inserts the values of the four influencing factors A, B, C and D into the formula for the formation of the price. Due to the fact that the subjects can influence the algorithmic input, the weighting of the influencing factor D can diverge somewhat in Treatment 3. For the random influence, the algorithm sets the expected value at €0. The result of this equation is the forecast of the algorithm P_t (see Table D-1). In period 1, the algorithm calculates as follows:

$$P_1 = 7 \cdot 14 - 6 \cdot 5 + 5 \cdot 5 + 2 \cdot 2 + 0 = 97 \quad (2)$$

For the calculation of the actual price, the random influence also has an effect. In period 1 it has a value of €+14. The actual price is thus calculated as follows:

$$K_1 = 7 \cdot 14 - 6 \cdot 5 + 5 \cdot 5 + 2 \cdot 2 + 14 = 111 \quad (3)$$

The difference between the actual share price K_t and the forecast of the algorithm P_t is the forecast error. This determines the amount of the bonus of the current forecasting round as described in accordance with the formula described in Table D-2. For a forecast whose forecast error lies within the interval $10 < |K_t - P_t| \leq 15$ for example, there is a bonus of €0.60.

Table D-1: Illustration of the modus operandi of the algorithm, how the share price is formed, and the calculation of the bonus

Period	Influencing factors				Forecast of the algorithm P_t	Random influence	Actual price K_t	Forecast error	Bonus
	A	B	C	D					
1	14	5	5	2	€97	+€14	€111	€14	€0.60

In practice one can see that perfect share price forecasts are not possible, even with knowledge of the most important influencing factors. On the contrary: share price trends have a number of similarities with random processes. This circumstance is taken into account by introducing the random influence. The random influence has the effect that the algorithm cannot make perfect forecasts. The forecast error of the algorithm thus corresponds to the random influence.

In this economic experiment, the random influence consistently lies within the interval -€30 ≤ Ɛ ≤ €30. It is always a whole number without decimal places. The exact distribution is described in Table D-2. The area -€15 ≤ Ɛ ≤ €15 (grey background) has a cumulative probability of 70%. For a forecast with a maximum forecasting error of €15 there is a payment of €0.60. In this way it can be ensured – as stated in the instructions – that the forecasts of the algorithm lead to a payment of at least €0.60 in 70% of cases.

Table D-2: Distribution of the random influence, which has an effect on the share price

Level of the random influence	Probability
-€30 ≤ Ɛ ≤ -€21 and €21 ≤ Ɛ ≤ €30	5% each (10%)
-€20 ≤ Ɛ ≤ -€16 and €16 ≤ Ɛ ≤ €20	10% each (20%)
-€15 ≤ Ɛ ≤ -€11 and €11 ≤ Ɛ ≤ €15	20% each (40%)
-€10 ≤ Ɛ ≤ -€6 and €6 ≤ Ɛ ≤ €10	10% each (20%)
-€5 ≤ Ɛ < €0 and €0 ≤ Ɛ ≤ €5	5% each (10%)

Lines highlighted in grey add up to the 70% success probability of achieving at least EUR 0.60 per forecast of the algorithm. Cells in white correspond to the remaining 30%.

As the level of the random influence is not known when handing in a forecast, the optimal strategy is to insert the values of the influencing factors A, B, C and D into the formula for the price formation mechanism and to assume an expected value of 0 for the random influence. This is precisely what the algorithm does. With the information available, it is thus not possible to make better forecasts than the algorithm.

When they make their own forecasts, the subjects also have the additional disadvantage that they do not know the exact formula for the price formation mechanism. They can only create an approximate picture of the price formation mechanism on the basis of examples of rounds of the game for which no payments were made (price history). For this purpose they are provided with the exact level of the share price, the influencing factors A, B, C and D as well as the random influence from ten previous rounds. From this information it is also already clear that making naïve forecasts – i.e., using the current price K_t without adaptation as a forecast for the following period P_{t+1} –

and continuously forecasting the average price of the last ten rounds are not promising approaches.

Given the advantage which the algorithm has in terms of information, there is thus no reason to presume that the subjects could succeed in making better forecasts. In effect they achieve an average total payment of €8.94 with their approach. They are thus clearly behind the payment of €10.03 obtained with the algorithm (p-value Wilcoxon rank-sum test ≤ 0.001). Decisions against using the algorithm can thus be considered to be algorithm aversion.

Chapter V

Algorithm Aversion as an Obstacle in the Establishment of Robo Advisors

Co-authored by Ibrahim Filiz, Marco Lorenz, and Markus Spiwoks
Contribution Jan René Judek: 45%

Published:
Journal of Risk and Financial Management, *Vol. 15, Issue 8, 353, 1-25. (Aug 2022)*
https://doi.org/10.3390/jrfm15080353

Also published as:

Discussion Paper in Sonderforschungsgruppe Institutionenanalyse (sofia), No. 22-2, Darmstadt, July 2022.

Working Paper in Wolfsburg Working Papers (WWP), Ostfalia University of Applied Sciences, No. 22-01, Wolfsburg, July 2022.

Abstract

Within the framework of a laboratory experiment, we examine to what extent algorithm aversion acts as an obstacle in the establishment of robo advisors. The subjects had to complete diversification tasks. They could either do this themselves or they could delegate them to a robo advisor. The robo advisor evaluated all the relevant data and always made the decision which led to the highest expected value for the subjects' payment. Although the high level of efficiency in the robo advisor was clear to see, the subjects only entrusted their decisions to the robo advisor in around 40% of cases. In this way, they reduced their success and their payment. Many subjects orientated themselves towards the 1/n-heuristic, which also contributed to their suboptimal decisions. As long as the subjects had to make decisions for others, they noticeably made a greater effort and were also more successful than when they made decisions for themselves. However, this did not have an effect on their acceptance of robo advisors. Even when they made

decisions on behalf of others, the robo advisor was only consulted in around 40% of cases. This tendency towards algorithm aversion among subjects is an obstacle to the broader establishment of robo advisors.

Keywords

Algorithm aversion; Robo advisors; Decisions for others; Portfolio choice; Diversification; Behavioral finance; Experiments.

JEL Classification

D81; D84; D91; G11; G21; G41; O31; O33.

1 Introduction

The traditional portfolio management business is demanding in terms of human resources and therefore comparatively expensive. Wealthy private customers have, however, become more price-sensitive since the establishment of low-cost investment opportunities such as exchange-traded funds (ETFs) in recent decades. Many banks are thus trying to find low-cost alternatives, particularly for the support of customers with smaller and medium-sized assets. The increased use of automated processes in portfolio management offers considerable scope for cost reduction. Many banks thus offer robo advisors (see, for example, Rühr et al., 2019; Jung et al., 2018; Singh & Kaur, 2017). Robo advisors are algorithms which are specialized in making investment decisions for customers and processing them. Using new technologies such as artificial neural networks, robo advisors are becoming increasingly more powerful and can potentially maximize clients' returns (Méndez-Suárez et al., 2019).

However, many customers have reservations about interacting with automated processes (robo advisors), although the latter are often remarkably effective (see, for example, Rossi & Utkus, 2020; Bhatia et al., 2020; D'Acunto et al., 2019; Beketov et al., 2018; Uhl & Rohner, 2018). So-called algorithm aversion is thus a significant problem for the banking sector.

Algorithm aversion particularly occurs when algorithms have to deal with stochastic processes. This is undoubtedly the case with robo advisors. Even when the algorithm makes very good investment decisions, it will—given the stochastic nature of financial market trends—never be able to always make perfect investment decisions. Dietvorst et al. (2015) showed that the tolerance of occasional errors by algorithms is much lower than the tolerance shown regarding occasional poor decisions which one has taken oneself or are made by an expert. We speak of algorithm aversion when subjects decline the use of an algorithm even though it is clearly recognizable that their own decisions or those of experts are by no means more successful (for the usual definitions, see, for example, Filiz et al., 2021a). There is a considerable amount of research results available on measures which can mitigate algorithm aversion (see, for example, Hinsen et al., 2022; Filiz et al., 2021b; Gubaydullina et al., 2021; Kim et al., 2021; Jung & Seiter, 2021; Castelo et al., 2019; Dietvorst et al., 2018; Taylor, 2017).

The efficiency of robo advisors is due—among other things—to the fact that they can make meaningful diversification decisions effortlessly. By contrast, investors often find it difficult to determine the expected earnings and the risk (variance) of alternative investments and to take into account the correlations of different investment opportunities in an appropriate way (see, for example, Ungeheuer & Weber, 2021; Cornil et al., 2019; Enke & Zimmermann, 2019; Gubaydullina & Spiwoks, 2015; Eyster & Weizsäcker, 2011; Kallir & Sonsino, 2009; Hedesstrom et al., 2006). This is why in practice many portfolios prove to be under-diversified or diversified in unsuitable ways (see, for example, Gomes et al., 2021; Chu et al., 2017; Dimmock et al., 2016; Anderson, 2013; Hibbert et al., 2012; Götzmann & Kumar, 2008; Meulbroek, 2005; Polkovnichenko, 2005;

Huberman & Sengmüller, 2004; Agnew et al., 2003; Guiso et al., 2002; Benartzi, 2001; Benartzi & Thaler, 2001; Barber & Odean, 2000; Bode et al., 1994; Blume & Friend, 1975; Lease et al., 1974).

We build on studies that have examined what influences the willingness to use a robo advisor. Alemanni et al. (2020) showed that the willingness to follow a robo advisor is lower when the robo advisor suggests a portfolio change. If the current portfolio is to be retained, the willingness to use is similar to advice from human advisors (Alemanni et al. 2020). In a questionnaire-based study, von Walter et al. (2022) found that consumers who believe artificial intelligence is better than human intelligence are more likely to accept advice from a robo advisor. Hodge et al. (2021) showed that subjects follow advice from a robo advisor without a name more closely than advice from a robo advisor that has been given a name. Robo advisors with names tend to be more popular for simple tasks than for complex ones (Hodge et al., 2021). The age of the decision-maker may also be a factor: Robillard (2018) argued that millennials may rely more heavily on robo advisors because this generation has lower trust in fellow humans than other generations.

Users' risk attitudes and attitudes toward automated processes also influence robo advisors and their investment decisions. Robo advisors can identify different risk profiles of their users, although there are large differences in risk preferences within the same investor type group (Boreiko & Massarotti, 2020). User preferences, however, have different effects on the perception of and intention to use robo advisors. For financial investments, a higher perceived level of automation leads to higher performance expectations and higher user control leads to lower perceived risk (Rühr et al., 2019). Since robo advisors should take user preferences into account to increase their usage intent, a performance-control dilemma arises that needs to be mitigated (Rühr, 2020).

Another important aspect seems to be the transfer of responsibility to the robo advisor. Niszczota and Kaszás (2020) discovered that moral investment decisions are rather delegated to humans than to robo advisors. On the other hand, Back et al. (2021) showed that subjects feel better in cases of loss if they have delegated some of the responsibility to the robo advisor. For tasks outside the world of finance, it has already been shown that punishment by third parties can be significantly lower if errors are committed by an algorithm rather than by a human (Feier et al., 2022). The idea that, under certain circumstances, subjects may be happy to hand over responsibility for possible future mistakes to a robo advisor is remarkable, and we explore it in more detail in the present study.

We carried out an economic experiment in which the subjects had to make four investment decisions. They could choose between different investment alternatives in each of the four cases. They were informed of the possible returns, the probability that these returns would materialize, and the correlations of the different investment opportunities. The subjects could either make their own diversification decisions or entrust the task to a robo advisor. The subjects knew that the robo advisor took all of the relevant data into account (the expected value of the returns, the probability that the

returns would materialize, and the correlation coefficients of the return development of the different investments), evaluated them optimally, and took them into account in its investment decisions. However, the subjects were also aware of the fact that the robo advisor could not know which random event will occur next. The subjects received the risk-adjusted return of their investment decisions as payment. This had the advantage that the subjects' risk preferences had no meaning for the assessment of the investment alternatives. We examine whether algorithm aversion occurs in this context and whether this can lead to a reduction in risk-adjusted returns. In this context, adding to previous research, we also consider whether algorithm aversion is less pronounced when a person has to make decisions for others.

Some empirical research findings indicate that when making decisions for others a change in the willingness of subjects to take risks can come into play (see, for example, Andersson et al., 2022; Eriksen et al., 2020; Vieider et al., 2016; Pahlke et al., 2015; Füllbrunn & Luhan, 2015; Bolton et al., 2015; Pahlke et al., 2012; Chakravarty et al., 2011; Charness & Jackson, 2009; Reynolds et al., 2009). This is particularly true when the person for whom a decision is being made is actually present (Polman, 2012). Later on, the persons for whom a decision is made may demand that the decision-maker justify their choices. If this is known in advance, it can lead to particular care on the part of the decision-maker. If the decision is delegated to an algorithm, however, the decision-makers do not have to justify their choices. This could possibly contribute towards a reduction in algorithm aversion.

This study examines the circumstances under which robo advisors can become an important complementary tool in wealth management. In addition to the performance of robo advisors in the meaningful diversification of investment alternatives, the reluctance of subjects to use automated processes (in this case: robo advisors) is the focus of attention. Measures to dampen algorithm aversion are of considerable interest. In this context, we raise the question of whether people are more likely to use a robo advisor if the consequences of the robo advisor's decision affect third parties. Exploring this question can help reduce hurdles to establish robo advisors. It also contributes to our understanding of the relatively new research field of algorithm aversion. The rest of this research paper is organized as follows: In Section 2, the experimental design is explained. Section 3 deals with the elaboration of the hypotheses. In Sections 4 and 5, the results of the economic experiments are presented and discussed in the context of the existing literature. To wrap up this study, Section 6 provides a summary of the key findings and a conclusion.

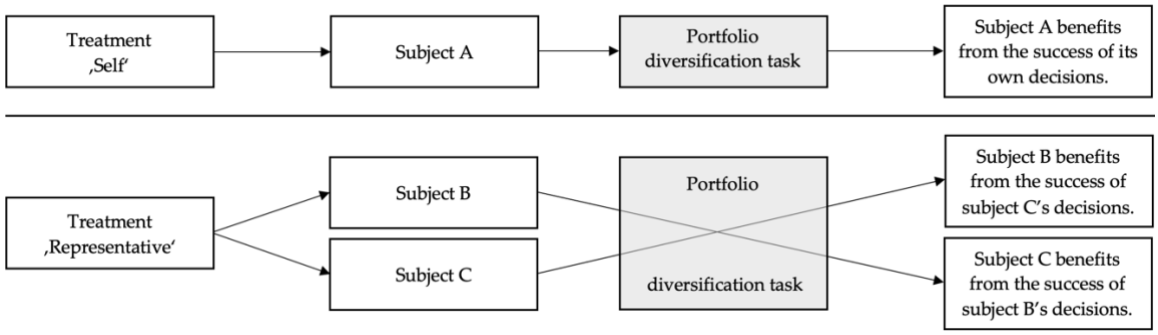
2 Experimental Design

In order to answer the research question, an economic experiment with two treatments was carried out between 20 and 28 April 2022 in the Ostfalia Laboratory of Experimental Economic Research (OLEW) at Ostfalia University of Applied Sciences in Wolfsburg. A total of 160 students of the Ostfalia University of Applied Sciences took part in the

experiment. Of these, 112 subjects (70%) were male, and 48 subjects (30%) were female. Of the 160 participants, 98 subjects (61.25%) studied at the Faculty of Economics and Business, 38 subjects (23.75%) at the Faculty of Vehicle Technology, and 24 subjects (15%) at other faculties. Their average age was 23.6 years.

In each treatment, subjects have to make four investment decisions (tasks 1-4) whose success directly affects them (or others). However, the subjects do not profit from gains in the share prices – they only profit (once) from the dividend payments of the shares in 2022. The subjects can either make their own diversification decisions or entrust the task to a robo advisor. In the treatment entitled 'Self' the subjects make a diversification decision for their own portfolio and receive the payment themselves. In the treatment entitled 'Representative' the subjects make a diversification decision for another participant's portfolio and the other participant in the session receives the payment which has been obtained. In the treatment 'Representative', after the payment has been made the subjects are informed about who is responsible for which payment.

Figure 1: The treatment 'Self' and the treatment 'Representative'



Let us assume, for example, that subject C receives the payment achieved by subject B and vice-versa (Figure 1). After the experiment, subject B could demand in a personal conversation that subject C justifies his or her decisions. And subject C could also demand that subject B justifies their decisions. All of the subjects who participate in the treatment 'Representative' are informed about this at the beginning of the experiment.

In the first task, there are two shares to choose from: share Y and share Z. The dividend payments of both companies are independent random processes with two possible configurations: 8 experimental currency units (ECU) and ECU 0. The probability of each of these occurring is 50%. The expected values of the dividend payments are thus ECU 4 each. The dividend payments of the two shares are wholly uncorrelated (correlation coefficient = 0). Table 1 shows the level of the dividend payments of the two shares in the past ten years. In this task, as well as in all other tasks in the experiment, subjects are given the dividend payments for the years 2012 to 2021. The dividend payment for 2022, which are relevant for their payoff, are still unknown, which is illustrated by the question mark.

Table 1: History of the random events of the dividend payments in task 1

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Share Y	ECU 8	ECU 0	ECU 8	ECU 8	ECU 8	ECU 0	ECU 8	ECU 0	ECU 0	ECU 0	?
Share Z	ECU 0	ECU 0	ECU 8	ECU 0	ECU 8	ECU 8	ECU 0	ECU 0	ECU 8	ECU 8	?

The subjects are allowed to compile a portfolio consisting of two shares. They can thus choose two Y shares, two Z shares, or one Y share and one Z share. As payment they receive the risk-adjusted dividends for 2022. A risk-adjusted dividend is equivalent to the dividend payment divided by the variance of the dividend payments of the chosen portfolio. The task thus consists of achieving the highest possible dividends with the lowest possible risk (low variance). The total of all risk-adjusted dividends (in ECU) which are obtained via portfolio decisions is multiplied by five at the end and then paid in euros.

As the subjects do not know the next random events for the dividend payments of share Y and share Z, it makes sense for them to orientate themselves towards the expected values and the variances of the three possible portfolios (see Table 2).

Table 2: Expected values and variances in task 1

Possible Portfolios	Expected Value of the Dividend	Variance	Expected Value of the Payment
2 Y shares	ECU 8	64	ECU 0.125 or EUR 0.625
2 Z shares	ECU 8	64	ECU 0.125 or EUR 0.625
1 Y share + 1 Z share	ECU 8	32	ECU 0.25 or EUR 1.25

Rational economic subjects orientate themselves towards the expected values of the payment, i.e., they select the mixed securities portfolio (1 Y share + 1 Z share). This is exactly how the robo advisor works.

All of the subjects have been familiarized with stochastic processes and the calculation of probabilities at school and also at the beginning of their degree programmes. They are aware of the fact that one cannot draw any conclusions about future random occurrences from an independent random event. Nevertheless, the temptation is great to make a forecast on which events will occur in the cases of the two shares in 2022 which is derived from the sequence of favorable and unfavorable dividend payments. People tend to see patterns even where there are definitely none (see, for example, Zielonka, 2004; Wärneryd, 2001; Gilovich et al., 1985; Roberts, 1959). Subjects who have succumbed to the hot hand fallacy (Burns, 2001; Gilovich et al., 1985) will tend to choose the portfolio of 2 Z shares. Subjects who believe in the gambler's fallacy (Rogers, 1998; Tversky & Kahneman,

1971) will prefer the 2 Y shares portfolio. Subjects who think they can predict the next random events will not make use of the robo advisor. Subjects who want to maximize the expected value of their payment can, however, sleep easily if they delegate the decision to the robo advisor because the robo advisor is specialized in making meaningful portfolio decisions and takes all of the relevant information into account in an optimal way in order to achieve risk-adjusted dividend payments which are as high as possible. The subjects are informed of this.

The second task is somewhat more complex. Once again, there are two shares to choose from (share X and share Q). Both of the shares can pay a dividend of either ECU 4 or ECU 0. The probability of each of these occurring is 50%. The expected values of the dividend payments are thus ECU 2 each. Once again, they are independent random events. The dividend payments of share X and share Q are completely uncorrelated (correlation coefficient = 0). Table 3 shows the level of the dividend payments of the two shares in the last 10 years.

Table 3: History of the random events of the dividend payments in task 2

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Share X	ECU 0	ECU 0	ECU 4	ECU 0	ECU 0	ECU 0	ECU 4	ECU 4	ECU 4	ECU 4	?
Share Q	ECU 0	ECU 4	ECU 4	ECU 4	ECU 0	ECU 4	ECU 0	ECU 0	ECU 4	ECU 0	?

The subjects can compile a portfolio consisting of four shares. They can thus choose four X shares, four Q shares, three X shares and one Q share, three Q shares and one X share, or two X shares and two Q shares. Neither the subjects nor the robo advisor know what the random events (dividend payments for share X and share Q) will be in 2022. A rational subject would orientate themselves towards the expected value of the payment and select the portfolio 2 X shares + 2 Q shares (see Table 4). This is exactly what the robo advisor does.

Table 4: Expected values and variances in task 2

Possible Portfolios	Expected Value of the Dividend	Variance	Expected Value of the Payment
4 X shares	ECU 8	64	ECU 0.125 or EUR 0.625
4 Q shares	ECU 8	64	ECU 0.125 or EUR 0.625
3 X shares + 1 Q share	ECU 8	40	ECU 0.20 or EUR 1
3 Q shares + 1 X share	ECU 8	40	ECU 0.20 or EUR 1
2 X shares + 2 Q shares	ECU 8	32	ECU 0.25 or EUR 1.25

The third task and the fourth task can no longer be accomplished with a crude diversification strategy such as the 1/n heuristic (see, for example, Fernandes, 2013; Baltussen & Post, 2011) because these are companies which belong to the same industry sector and whose dividend payments depend on the success of the sector. The dividend payments of the two shares are thus completely positively correlated (correlation coefficient = 1). Table 5 shows the amount of the dividend payments in the past ten years.

Table 5: History of the random events of the dividend payments in task 4

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Share M	ECU 4	ECU 0	ECU 4	ECU 0	ECU 0	ECU 0	ECU 4	ECU 4	ECU 0	ECU 4	?
Share P	ECU 3	ECU 1	ECU 3	ECU 1	ECU 1	ECU 1	ECU 3	ECU 3	ECU 1	ECU 3	?

A phase in which companies in this sector are either successful or are struggling occurs purely coincidentally with a probability of 50%. Previous events thus provide no indication of which random events might occur in the future. The expected value of the dividend payments is thus ECU 2 for both shares. The subjects can compile a portfolio consisting of four shares.

Given that the dividend payments for both shares are 100% positively correlated, a mixture of the two shares does not create any diversification effect. The optimal strategy is to select four P shares because that is the minimum variance portfolio (see Table 6). This is precisely the strategy pursued by the robo advisor.

Table 6: Expected values and variances in task 4

Possible Portfolios	Expected Value of the Dividend	Variance	Expected Value of the Payment
4 M shares	ECU 8	64	ECU 0.125 or EUR 0.625
4 P shares	ECU 8	16	ECU 0.50 or EUR 2.50
3 M shares + 1 P share	ECU 8	49	ECU 0.165 or EUR 0.825
3 P shares + 1 M share	ECU 8	25	ECU 0.32 or EUR 1.60
2 M shares + 2 P shares	ECU 8	36	ECU 0.225 or EUR 1.125

The experiment proceeds as follows: First, the subjects read the instructions and answer the control questions (see Appendices A and B). Afterwards, they make the four portfolio decisions of tasks 1 to 4 either with the help of the robo advisor or independently (see Appendix C). For each of the four tasks, the subjects can decide again whether they want to delegate the task to the robo advisor or whether they want to choose a portfolio

composition themselves. Only after the four tasks have been completed is it revealed which random events have occurred in this session and to which compensation the subjects have progressed. The payment is then made in cash.

3 Hypotheses

The most meaningful strategy is to delegate all four tasks to the robo advisor. The robo advisor always makes the most meaningful decisions. It always selects the portfolio composition which maximizes the expected value of the payment in euros. It would actually be possible to work out this optimal decision oneself. However, the amount of effort required to do so is considerable. The subjects can make mistakes when calculating the expected payment amount. The robo advisor, on the other hand, always evaluates all of the relevant data in an optimal way and always makes the decision which maximizes the expected value of the payment. Nevertheless, it has to be expected that some subjects will have reservations about using a robo advisor. The wide variety of previous findings on the occurrence of algorithm aversion make this highly likely (Mahmud et al., 2022; Kawaguchi, 2021; Burton et al., 2020; Castelo et al., 2019; Prahl & Van Swol, 2017).

Hypothesis 1. *Not all of the subjects will trust the robo advisor (algorithm), although it is not possible for them to make a better decision. This means that algorithm aversion will occur.*

Null Hypothesis 1. *All of the subjects will trust the robo advisor (algorithm). This means that algorithm aversion will not occur.*

If the subjects are wary of using the robo advisor (algorithm aversion), this may well lead—on average—to a reduction in the payment they obtain. Algorithm aversion will presumably cause a loss in potential earnings.

Hypothesis 2. *The more frequently the subjects delegate their decision to the robo advisor, the higher their payments will be.*

Null Hypothesis 2. *The frequency with which the subjects delegate their decisions to the robo advisor does not have a positive influence on their payment.*

Among the subjects, there will presumably be some who pursue a crude diversification strategy (1/n-heuristic; see, for example, Fernandes, 2013; Morrin et al., 2012; Baltussen & Post, 2011; Huberman & Jiang, 2006; Benartzi & Thaler, 2001). This strategy can lead to success in tasks 1 and 2. In tasks 3 and 4, on the other hand, it cannot lead to success. For an optimal solution of tasks 3 and 4, it is necessary to also take into account the correlation coefficients alongside the expected values of the dividends.

Hypothesis 3. *Subjects who do not deploy the algorithm partly neglect the correlations, and in the cases of tasks 3 and 4 they find the optimal solution significantly less often than in tasks 1 and 2.*

Null Hypothesis 3. *Subjects who do not deploy the algorithm do not neglect the correlations, and in the cases of tasks 3 and 4 they do not find the optimal solution significantly less often than in tasks 1 and 2.*

Based on the existing research on decision making for others (see, for example, Pahlke et al., 2015; Polman, 2012; Pahlke et al., 2012; Charness & Jackson, 2009; Reynolds et al., 2009) we presume that the subjects who make decisions for others (the treatment 'Representative') consider their decisions more carefully and try harder to make meaningful decisions. After all, the persons for whom the decisions are being made are actually present. At the end of the experiment, who decided for whom and what the results were is announced. All of the subjects in the treatment 'Representative' are aware of this. In other words, they have to expect that they will need to justify their decisions. The subjects in the treatment 'Self', on the other hand, are only responsible for themselves. They need not fear that someone will demand that they justify their decisions. We therefore presume that algorithm aversion will occur less frequently in the treatment 'Representative' than in the treatment 'Self'. In addition, we presume that those persons in the treatment 'Representative' who do not want to trust the robo advisor—for whatever reason—will make a greater effort to select meaningfully diversified portfolios.

Hypothesis 4. *The solution of the tasks is delegated to the robo advisor significantly more often in the treatment 'Representative' than in the treatment 'Self'.*

Null Hypothesis 4. *The solution of the tasks is not delegated to the robo advisor significantly more often in the treatment 'Representative' than in the treatment 'Self'.*

Hypothesis 5. *Those persons who do not want to trust the robo advisor will choose the optimal portfolio structure significantly more often in the treatment 'Representative' than in the treatment 'Self'.*

Null Hypothesis 5. *Those persons who do not want to trust the robo advisor will not choose the optimal portfolio structure significantly more often in the treatment 'Representative' than in the treatment 'Self'.*

The general research question of this study is: Can robo advisors become useful complementary tools in the modern wealth management business? In order to explore our research question, we assume that a robo advisor cannot forecast future capital market developments without errors. However, a robo advisor can effortlessly make meaningful diversification decisions. This leads to the question if economic agents need a robo advisor in order to achieve good diversification decisions with certainty. In four very clear decision situations where shares are assembled into a portfolio, optimal decisions can easily be made. However, the facts (expected value, the dispersion of events around expected value, and the correlation of the events of different shares) are neglected or misinterpreted by many economic agents. Therefore, a greater willingness to delegate the decision to the robo advisor presumably leads to greater investment success or higher compensation (Hypothesis 2). The fact that is most often neglected is probably the correlation of the returns of different shares (Hypothesis 3).

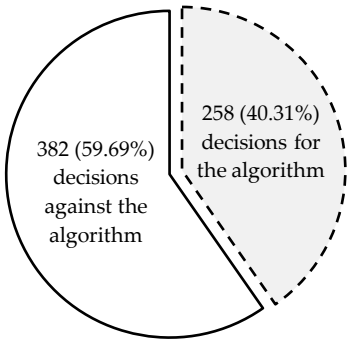
Although the subjects know that the robo advisor optimally evaluates all relevant information and makes the best possible diversification decision in each case, experience has shown that many economic subjects are reluctant to entrust themselves to an algorithm— in this case a robo advisor (Hypothesis 1). Thus, if robo advisors are to be successfully established, measures to mitigate algorithm aversion have to be considered. One possible measure would be to place the decision to use a robo advisor in the context of decision for others. After all, investment decisions are not only important for the wealthy person but also for his or her family, especially children and grandchildren. Thus, in the case of decision for others, the willingness to use the robo advisor might increase (Hypothesis 4) because economic agents might try harder to make a meaningful decision when making decisions that (also) affect others (Hypothesis 5).

4 Results

Of the 160 participants, 80 subjects played the treatment 'Self' and 80 played the treatment 'Representative'. The experiment was carried out using z-Tree (Fischbacher, 2007). The time needed for reading the instructions of the experiment (Appendix A), answering the test questions (Appendix B), and carrying out the four tasks took 15 min on average. An average payment of EUR 6.89 seemed very attractive for the amount of time required. It was intended to be sufficient incentive for meaningful economic decisions, and the subjects did actually give the impression of being concentrated and motivated.

In the first instance, it could be seen that algorithm aversion occurred to a considerable extent. Although it was clear to all of the participants that using the algorithm (robo advisor) definitely led to the best possible decisions, the robo advisor was deployed in less than half of the cases. A total of 160 subjects had to make four decisions each. This was a total of 640 decisions. The subjects decided to delegate the task to the robo advisor in only 258 cases (40.31%). In 382 cases (59.69%), the subjects refrained from using the algorithm (Figure 2). The reason why this is so remarkable is that all of the subjects knew that the robo advisor evaluated all of the relevant data in an optimal way and therefore always made the best possible decision.

Figure 2: Decisions for and against the algorithm (robo advisor)



An average subject relied on the algorithm in only 1.612 out of 4 rounds. The *t*-test shows in all clarity that Null Hypothesis 1 has to be rejected (p-value ≤ 0.001). The Z-test supports that only very few subjects (36 out of 160) consistently follow the rational strategy and rely on the algorithm in all rounds of the experiment (p-value ≤ 0.001). Algorithm aversion thus obviously occurs to a considerable extent (59.69% of all decisions).

It is of particular interest whether this tendency towards algorithm aversion really led to a smaller number of optimal diversification decisions and whether the payments were lower than would have been the case when the subjects had consistently trusted the robo advisor. After all, one cannot simply presume that the decisions of the subjects who did not always use the robo advisor were less successful.

A total of 53 subjects did not delegate their decision to the robo advisor a single time. In 89 out of 212 decisions (41.98%), these subjects selected optimal portfolios. On average, they achieved an expected payment value of EUR 6.36. How much the actual payment is also depended on the specific random events (dividend payments). Here, there was an average payment of EUR 6.67 (Table 7).

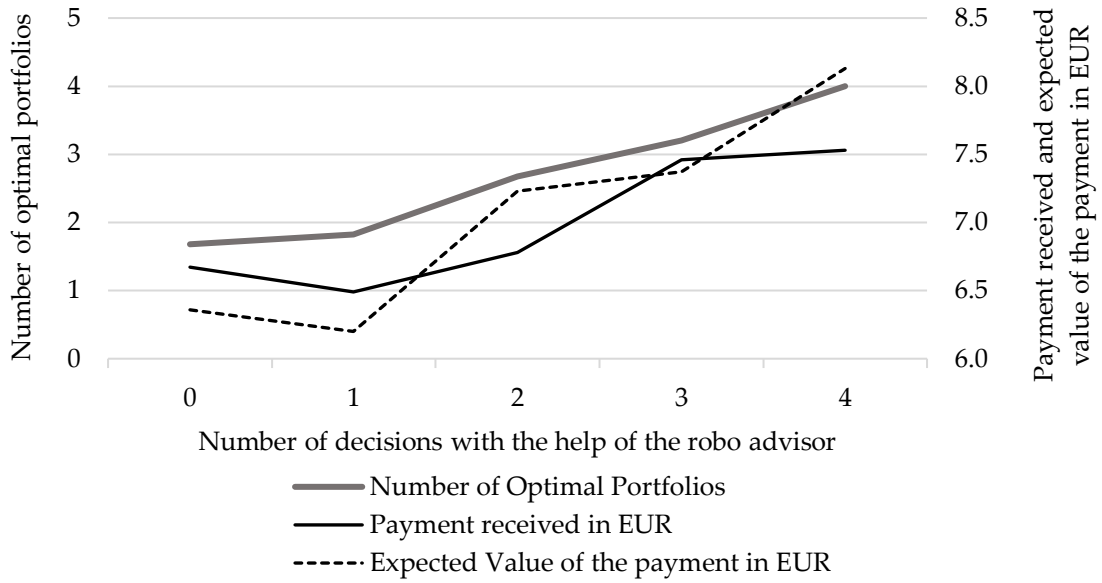
Table 7: Average success in relation to the extent of algorithm aversion

Number of times the algorithm was chosen	Number of subjects	Optimal portfolios	Expected value of the payment in Euros	Actual payment in Euros
0	53	89 (41.98%)	EUR 6.36	EUR 6.67
1	39	71 (45.51%)	EUR 6.20	EUR 6.49
2	19	51 (67.11%)	EUR 7.23	EUR 6.78
3	15	48 (80.00%)	EUR 7.37	EUR 7.46
4	34	136 (100%)	EUR 8.13	EUR 7.53

A total of 34 subjects delegated all four of their decisions to the robo advisor. As was to be expected, in 136 out of 136 decisions (100%), the optimal portfolios were chosen. The subjects achieved an expected payment value of EUR 8.13. The specific random events (dividend payments) led to an average payment of EUR 7.53 (Table 7).

Figure 3 shows clearly that the more frequently the subjects delegated their decision to the robo advisor, the more successful they were. The subjects who did not put their faith in the robo advisor a single time achieved an average of only 1.68 optimal portfolios. The subjects who used the robo advisor to solve all four tasks made 4.00 optimal decisions. The F-test confirms: the more frequently the robo advisor was used, the more optimal portfolios were compiled (thick grey line, left scale, p-value ≤ 0.001), the higher the expected value of the payment (dashed black line, right scale, p-value ≤ 0.001), and the higher the actual payment (continuous black line, right scale, p-value ≤ 0.001).

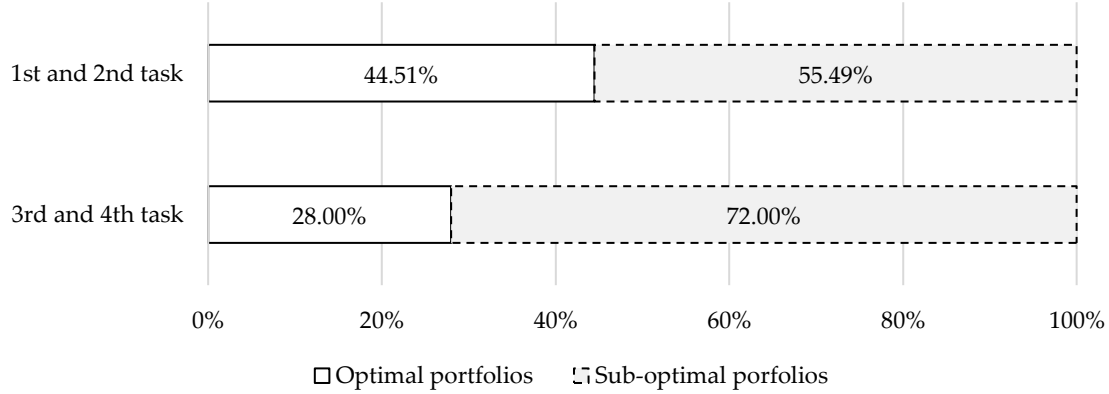
Figure 3: Average success in relation to the extent of algorithm aversion



The stronger the effect of algorithm aversion, the less successful the subjects were. Null Hypothesis 2 thus has to be discarded.

Now let us look at the success of the decisions which were not delegated to the robo advisor. Tasks 1 and 2 can be solved well with the simple understanding of diversification of the 1/n heuristic. In tasks 3 and 4, however, it is absolutely necessary to take the correlations between the dividend payments of the two shares into account and to understand the variances of the dividend payments of the two shares. Among the decisions which are not delegated to the robo advisor, a clear difference can indeed be seen between the success rate in tasks 1 and 2 on the one hand and the success rates in tasks 3 and 4 on the other. In tasks 1 and 2, 81 out of 182 decisions (44.51%) led to optimal portfolios. In tasks 3 and 4, on the other hand, only 56 out of 200 decisions (28%) led to optimal portfolios which maximized the expected value of the payment. In the chi square test, this difference proves to be significant (p-value ≤ 0.001.) Null Hypothesis 3 thus has to be rejected (Figure 4).

Figure 4: Percentage share of optimal portfolios according to tasks

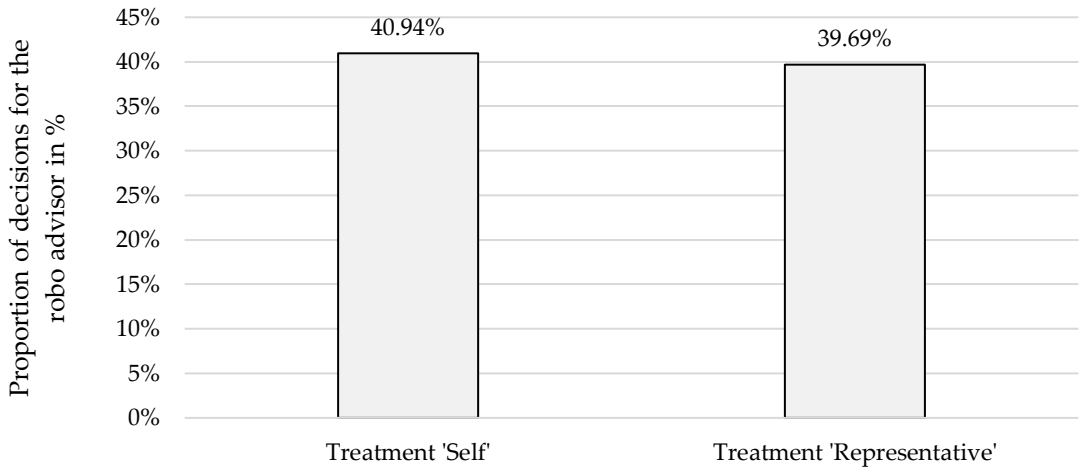


In a comparison of the two treatments 'Self' and 'Representative', no noteworthy differences with regard to use of the robo advisor can be seen. In the treatment 'Self', 131 of out 320 decisions (40.94%) were delegated to the robo advisor. In the treatment 'Representative', 127 out of 320 decisions (39.69%) were delegated to the robo advisor (Table 8 and Figure 5). This is only a very small difference. It proves to be insignificant both in the Wilcoxon rank sum test (p-value = 0.752) as well as in the chi square test (p-value = 0.747). Null Hypothesis 4 can therefore not be rejected.

Table 8: Influence of the treatments on algorithm aversion

Treatment	Decisions for the Robo advisor	Own decisions	Total
'Self'	131	189	320
'Representative'	127	193	320

Figure 5: Acceptance of the robo advisor according to treatments



This is a surprising result. The subjects in the treatment 'Representative' could have easily transferred their responsibility for the payment of another person to the robo advisor. Given that the robo advisor is known for the fact that it always makes optimal decisions, nobody needs to be afraid of being criticized. However, a large part of the subjects obviously had such far-reaching reservations regarding the deployment of a robo advisor that they did not want to take this route. We thus have to come to the conclusion that algorithm aversion occurs frequently and is by no means easy to overcome.

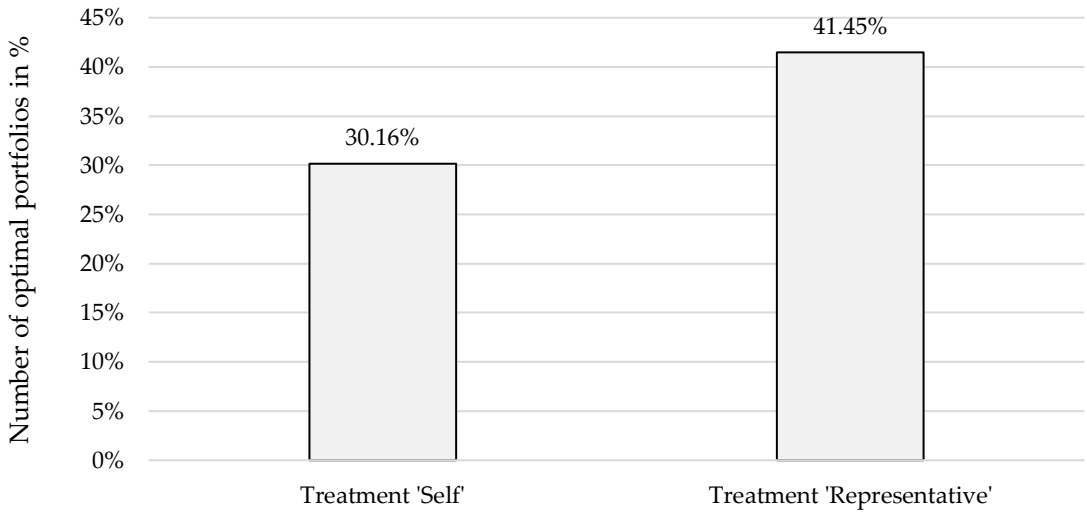
However, it is noticeable that it does make a difference whether one makes decisions for oneself or for others. The subjects in the treatment 'Representative' really did make a greater effort to make meaningful decisions. This can be seen in the decisions they made

without using the robo advisor. In 57 out of 189 decisions (30.16%) the subjects in the treatment 'Self' succeed in building optimal portfolios (portfolios with the highest expectation value for the payment in euros). In 80 out of 193 decisions (41.45%), the subjects in the treatment 'Representative' succeed in building optimal portfolios (portfolios with the highest expectation value for the payment in euros) (Table 9 and Figure 6). This difference turns out to be statistically significant in the chi square test (p-value = 0.021).

Table 9: Success of portfolio decisions without the robo advisor according to treatments

Treatment	Number of subjects	Number of optimal portfolios without the robo advisor	Number of sub-optimal portfolios without the robo advisor	Number of decisions made by the robo advisor	Total
'Self'	80	57	132	131	320
'Representative'	80	80	113	127	320

Figure 6: Success of the portfolio decisions without the robo advisor according to treatments



A clear difference between the two treatments can be seen. The subjects behaved differently depending on whether they were deciding for themselves or for others. They obviously acted less impulsively in the treatment 'Representative' and weighed up more precisely which portfolio composition would presumably lead to the largest payment. However, this effort to make meaningful decisions did not lead to a greater acceptance of

robo advisors. The subjects' reservations about using an algorithm were obviously stronger than their wish to make decisions for others with particular care.

5 Discussion

Our results contribute to the academic debate in three ways. First, it has been shown that many subjects have massive reservations about robo advisors despite their obvious advantages. In our study, robo advisors consistently outperformed subjects. Still, most subjects chose not to use them. Although robo advisors have enormous potential and perform significantly better on average, they seemed to be very unpopular among subjects. This is in line with previous studies, which also found that algorithm aversion in particular can be a hurdle in establishing robo advisors (Hodge et al., 2021; Alemanni et al., 2020; Niszczota & Kaszás, 2020).

Second, our research confirms that algorithm aversion is a serious barrier to the diffusion of innovative business fields in general. In this respect, we may also be facing a societal problem. Already today, the use of algorithms clearly provides humans with more powerful options for solving problems. Yet, decision-makers refuse to use them. Instead, they perform tasks themselves, leading to higher costs and poorer results. It therefore remains an important task of research, especially with regard to cognitive biases and heuristics, to further explore the background of algorithm aversion in order to contribute to the progress of society.

Third, it turns out that it makes little difference to the extent of algorithm aversion who has to bear the consequences (oneself or third parties). Research by Back et al. (2021) suggests that one reason to consult a robo advisor might be that it feels like relinquishing some of the responsibility for unpleasant tasks and potential mistakes. However, this assumption was not confirmed in our study. If subjects made decisions for others who may have demanded a justification for possible mistakes, the robo advisor was nevertheless just as unpopular.

To save taxes, many wealthy private clients transfer part of their assets to their children while they are still minors. These assets also need to be managed. The parents now have to decide on behalf of their children how this should be accomplished. If algorithm aversion were less prominent in decisions for others, this could be a starting point to resolve or at least mitigate the bias against robo advisors. However, no evidence for this has emerged. Algorithm aversion is reflected to the same extent in the decisions that economic agents make for themselves and in the decisions that they make for others.

Of course, there are also some limitations that may affect the validity of our results for practical applications. First, it should be mentioned that the results were obtained in the context of financial decisions with robo advisors. Financial decisions are influenced by a variety of factors, such as financial literacy or experience. Algorithm aversion is far from

being the only influencing factor. It may therefore be worthwhile to revisit our research question in relation to other areas of use for algorithms.

Moreover, robo advisors from reputable banks go through a detailed accreditation process. In this process, independent experts verify, for example, whether the robo advisors take appropriate measures to hedge risks and also make decisions that are justifiable from an ethical point of view. Accreditation is thus a tool that can increase user confidence. However, it cannot be replicated in the same way in an economic laboratory experiment.

Finally, when making decisions on behalf of others, it may always make a difference what one's relationship is to the person who has to bear the consequences. We conducted a laboratory experiment at our research institution. Usually, students go there together with fellow students whom they know from classes. Sometimes students also come alone. As such, the consequences of the decision in the treatment 'Representative' were largely borne either by complete strangers or loose acquaintances. It must be left to future research efforts to see if a different outcome emerges when we decide, for example, on behalf of loved ones.

6 Conclusion

Robo advisors are algorithms which can automatically make investment decisions for asset management customers. Given the increased price sensitivity of wealthy private clients, robo advisors are one way to offer solid portfolio management decisions at a low cost. However, customers have considerable reservations about algorithms, even when they are very efficient systems. This phenomenon, which is known as algorithm aversion, is considered in more detail in this study.

In a laboratory experiment, subjects made a total of four portfolio decisions. They could either try to determine the optimal portfolio composition in each case themselves or they could delegate the task to a robo advisor. The robo advisor took all the relevant information into account in an optimal way and always chose the portfolio composition which led to the highest expected value of the payment in euros. The subjects were familiar with the qualities of the robo advisor. Nevertheless, they only used it in around 40% of all cases. In around 60% of all decision-making situations, the subjects trusted in their own judgement, although it must have been clear to them that they were not able to make better decisions than the robo advisor. Algorithm aversion thus occurred to a great extent.

The actual success rate of the subjects who did not put their faith in the robo advisor was indeed lower than that of the robo advisor. This applied to the average number of optimal portfolio compositions, to the average expected values of their payment in euros, and also with regard to the actually obtained payment in euros. It is crystal clear that the more frequently the subjects delegated their decisions to the robo advisor, the greater

their success. With their aversion towards the algorithm, the subjects were recognizably damaging themselves.

The subjects had particular difficulties when trying to take into account the correlation between the different investments. Tasks which could be solved with the simple diversification strategy of the 1/n heuristic (tasks 1 and 2) were dealt with successfully significantly more often than tasks which could not be suitably dealt with using the 1/n heuristic (tasks 3 and 4).

Ultimately, it became clear that subjects who had to make decisions for others approached the task in a more careful and concentrated way. Among the decisions which were not made by the robo advisor, there were significantly more optimal portfolios within the subjects who made decisions for others than among those who decided for themselves. However, this did not have an effect on algorithm aversion. Regardless of whether the subjects decided for themselves or for others, a readiness to delegate the decision to the robo advisor could only be seen in around 40% of decisions.

To summarize, the following can be stated: The deployment of robo advisors can, under certain circumstances, be a low-cost and very efficient alternative to traditional asset management. However, algorithm aversion hinders the establishment of the business which could be had with robo advisors.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/jrfm15080353/s1.

Author Contributions: Conceptualization, M.S.; Software, I.F., J.R.J., and M.L.; Validation, I.F., J.R.J., M.L., and M.S.; Formal analysis, I.F., J.R.J., M.L., and M.S.; Data curation, I.F., J.R.J., M.L., and M.S.; Writing—original draft preparation, J.R.J., M.L., and M.S.; Writing—review and editing, I.F., J.R.J., M.L., and M.S.; Visualization, J.R.J., and M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Our data can be accessed at Supplementary Materials.

Acknowledgments: The authors thank the editor and the anonymous reviewers for their constructive comments and useful suggestions, which were very helpful to enhance the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

Agnew, J., Balduzzi, P., & Sundén, A. (2003), Portfolio Choice and Trading in a Large 401(k) Plan, *The American Economic Review*, 93(1), 193-215.

Alemanni, B., Angelovski, A., di Cagno, D. T., Galliera, A., Linciano, N., Marazzi, F., & Soccorso, P. (2020), Do Investors Rely on Robots? Evidence From an Experimental Study. *CONSOB Fintech Series*, (7).

Anderson, A. (2013), Trading and Under-Diversification, *Review of Finance*, 17(5), 1699-1741.

Andersson, O., Holm, H. J., Tyran, J.-R., & Wengström, E. R. (2022), Deciding for Others Reduces Loss Aversion, *Management Science*, 62(1), 29-36.

Back, C., Morana, S., & Spann, M. (2021), Do robo-advisors make us better investors?, Discussion Paper No. 276, University of Munich (LMU) and Humboldt University Berlin, Collaborative Research Center Transregio 190: Rationality and Competition, München & Berlin.

Baltussen, G., & Post, G. T. (2011), Irrational Diversification: An Examination of Individual Portfolio Choice, *Journal of Financial and Quantitative Analysis*, 46(5), 1463-1491.

Barber, B. M., & Odean, T. (2000), Trading is Hazardous to your Wealth: The Common Stock Investment Performance of Individual Investors, *Journal of Finance*, 55(2), 773-806.

Beketov, M., Lehmann, K., & Wittke, M. (2018), Robo Advisors: quantitative methods inside the robots, *Journal of Asset Management*, 19, 363-370.

Benartzi, S. (2001), Excessive Extrapolation and the Allocation of 401(k) Accounts to Company Stock, *The Journal of Finance*, 56(5), 1747-1764.

Benartzi, S., & Thaler, R. H. (2001), Naïve Diversification Strategies in Defined Contribution Saving Plans, *American Economic Review*, 91(1), 79-98.

Bhatia, A., Chandani, A., & Chhateja, J. (2020), Robo advisory and its potential in addressing the behavioral biases of investors — A qualitative study in Indian context, *Journal of Behavioral and Experimental Finance*, 25.

Blume, M. E., & Friend, I. (1975), The Asset Structure of Individual Portfolios and Some Implications for Utility Functions, *The Journal of Finance*, 30(2), 585-603.

Bode, M., van Echelpoel, A., & Sievi, C. R. (1994), Multinationale Diversifikation: Viel zitiert, kaum befolgt, *Die Bank*, 94(4), 202-206.

Bolton, G. E., Ockenfels, A., & Stauf, J. (2015), Social responsibility promotes conservative risk behavior, *European Economic Review*, 74(C), 109-127.

Boreiko, D., & Massarotti, F. (2020), How Risk Profiles of Investors Affect Robo-Advised Portfolios, *Frontiers in Artificial Intelligence*, 3(60), 1-9.

Burns, B. D. (2001), The hot hand in basketball: Fallacy or adaptive thinking, *Proceedings of the Annual Meeting of the Cognitive Science Society*, 23(23), 152-157.

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

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

Chakravarty, S., Harrison, G. W., Haruvy, E., & Rutstrom, E. (2011), Are You Risk Averse over Other People's Money?, *Southern Economic Journal*, 77(4), 901-913.

Charness, G., & Jackson, M. O. (2009), The role of responsibility in strategic risk-taking, *Journal of Economic Behavior & Organization*, 69(3), 241-247.

Chu, Z., Wang, Z., Xiao, J. J., & Zhang, W. (2017), Financial literacy, portfolio choice and financial well-being, *Social Indicators Research*, 132(2), 799-820.

Cornil, Y., Hardisty, D. J., & Bart, Y. (2019), Easy, breezy, risky: Lay investors fail to diversify because correlated assets feel more fluent and less risky, *Organizational Behavior and Human Decision Processes*, 153, 103-117.

D'Acunto, F., Prabhala, N., & Rossi, A. G. (2019), The Promises and Pitfalls of Robo-Advising, *The Review of Financial Studies*, 32(5), 1983-2020.

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.

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), 114-126.

Dimmock, S. G., Kouwenberg, R., Mitchell, O. S., & Peijnenburg, K. (2016), Ambiguity Aversion and Household Portfolio Choice Puzzles: Empirical Evidence, *Journal of Financial Economics*, 119, 559-577.

Enke, B., & Zimmermann, F. (2019), Correlation neglect in belief formation, *The Review of Economic Studies*, 86(1), 313-332.

Eriksen, K. W., Kvaløy, O., & Luzuriaga, M. (2020), Risk-taking on behalf of others, *Journal of Behavioral and Experimental Finance*, 26(C), 1-13.

Eyster, E., & Weizsäcker, G. (2011), Correlation Neglect in Financial Decision Making, *DIW Discussion Papers*, No. 1104, Berlin.

Feier, T., Gogoll, J., & Uhl, M. (2022). Hiding behind machines: artificial agents may help to evade punishment, *Science and Engineering Ethics*, 28(2), 1-19.

Fernandes, D. (2013), The 1/N Rule Revisited: Heterogeneity in the Naïve Diversification Bias, *International Journal of Research in Marketing*, 30(3), 310-313.

Filiz, I., Judek, J. R., Lorenz, M., & Spiwoks, M. (2021a), The Tragedy of Algorithm Aversion, *WWP – Wolfsburg Working Papers*, 21-02, Wolfsburg.

Filiz, I., Judek, J. R., Lorenz, M., & Spiwoks, M. (2021b), Reducing algorithm aversion through experience, *Journal of Behavioral and Experimental Finance*, 31(5), 1-8.

Fischbacher, U. (2007), z-Tree: Zurich Toolbox for Ready-made Economic Experiments, *Experimental Economics*, 10(2), 171-178.

Füllbrunn, S., & Luhan, W. J. (2015), Am I My Peer's Keeper? Social Responsibility in Financial Decision Making, *Ruhr Economic Paper*, No. 551.

Gilovich, T., Vallone, R., & Tversky, A. (1985), The hot hand in basketball: On the misperception of random sequences, *Cognitive psychology*, 17(3), 295-314.

Goetzmann, W. N., & Kumar, A. (2008), Equity Portfolio Diversification, *Review of Finance*, 12(3), 433-463.

Gomes, F., Haliassos, M., & Ramadorai, T. (2021), Household finance, *Journal of Economic Literature*, 59(3), 919-1000.

Gubaydullina, Z., Judek, J. R., Lorenz, M., & Spiwoks, M. (2021), Creative Drive and Algorithm Aversion – The Impact of Influence in the Process of Algorithmic Decision-making on Algorithm Aversion, *WWP – Wolfsburg Working Papers*, 21-04, Wolfsburg.

Gubaydullina, Z., & Spiwoks, M. (2015), Correlation Neglect, Naïve Diversification, and Irrelevant Information as Stumbling Blocks for Optimal Diversification, *Journal of Finance and Investment Analysis*, 4(2), 1-19.

Guiso, L., Haliassos, M., & Japelli, T. (2002), Household Portfolios, *MIT Press*, Cambridge, MA.

Hedesstrom, T. M., Svedsater, H., & Garling, T. (2006), Covariation Neglect among Novice Investors, *Journal of Experimental Psychology-Applied*, 12(3), 155-165.

Hibbert, A. M., Lawrence, E. R., & Prakash, A. J. (2012), Can Diversification Be Learned? *The Journal of Behavioral Finance*, 13(1), 38-50.

Hinsen, S., Hofmann, P., Jöhnk, J., & Urbach, N. (2022), How Can Organizations Design Purposeful Human-AI Interactions: A Practical Perspective From Existing Use Cases and Interviews, *Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS)*, Honolulu, HI, University of Hawai'i at Manoa, Hamilton Library.

Hodge, F. D., Mendoza, K. I., & Sinha, R. K. (2021), The effect of humanizing robo-advisors on investor judgments, *Contemporary Accounting Research*, 38(1), 770-792.

Huberman, G., & Sengmueller, P. (2004), Performance and Employer Stock in 401(k) Plans, *Review of Finance*, 8(3), 403-443.

Huberman, G., & Jiang, W. (2006), Offering versus choice in 401 (k) plans: Equity exposure and number of funds, *The Journal of Finance*, 61(2), 763-801.

Jung, D., Dorner, V., Glaser, F., & Morana, S. (2018), Robo-Advisory - Digitalization and Automation of Financial Advisory, *Business & Information Systems Engineering*, 60(1), 81-86.

Jung, M., & Seiter, M. (2021), Towards a better understanding on mitigating algorithm aversion in forecasting: an experimental study, *Journal of Management Control*, 32, 495-516.

Kallir, I., & Sonsino, D. (2009), The Neglect of Correlation in Allocation Decisions, *Southern Economic Journal*, 75(4), 1045-1066.

Kawaguchi, K. (2021), When will workers follow an algorithm? A field experiment with a retail business, *Management Science*, 67(3), 1670-1695.

Kim, J., Giroux, M., & Lee, J. C. (2021), When do you trust AI? The effect of number presentation detail on consumer trust and acceptance of AI recommendations, *Psychology & Marketing*, 38(7), 1140-1155.

Lease, R. C., Lewellen, W. G., & Schlarbaum, G. G. (1974), The Individual Investor: Attributes and Attitudes, *The Journal of Finance*, 29(2), 413-433.

Mahmud, H., Islam, A. N., Ahmed, S. I., & Smolander, K. (2022), What influences algorithmic decision-making? A systematic literature review on algorithm aversion, *Technological Forecasting and Social Change*, 175, 121390, 1-26.

Méndez-Suárez, M., García-Fernández, F., & Gallardo, F. (2019), Artificial intelligence modelling framework for financial automated advising in the copper market, *Journal of Open Innovation: Technology, Market, and Complexity*, 5(4), 81.

Meulbroek, L. (2005), Company Stock in Pension Plans: how costly is it?, *The Journal of Law and Economics*, 48(2), 443-474.

Morrin, M., Inman, J. J., Broniarczyk, S. M., Nenkov, G. Y., & Reuter, J. (2012), Investing for Retirement: The Moderating Effect of Fund Assortment Size on the 1/N Heuristic, *Journal of Marketing Research*, 49(4), 537-550.

Niszczota, P., & Kaszás, D. (2020). Robo-fund aversion: People prefer it when humans and not computers make investment decisions with moral undertones, *PsyArXiv*, March, 13.

Pahlke, J., Strasser, S., & Vieider, F. M. (2015), Responsibility effects in decision making under risk, *Journal of Risk and Uncertainty*, 51(2), 125-146.

Pahlke, J., Strasser, S., & Vieider, F. M. (2012), Risk-taking for others under accountability, *Economics Letters*, 114(1), 102-105.

Polkovnichenko, V. (2005), Household Portfolio Diversification: a Case for Rank-dependent Preferences, *Review of Financial Studies*, 18, 1467-1502.

Polman, E. (2012), Self–other decision making and loss aversion, *Organizational Behavior and Human Decision Processes*, 119(2), 141-150.

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

Reynolds, D. B., Joseph, J., & Sherwood, R. (2009), Risky Shift Versus Cautious Shift: Determining Differences In Risk Taking Between Private And Public Management Decision-Making, *International Journal of Economics and Business Research*, 7(1), 63-78.

Roberts, H. V. (1959), Stock market ''patterns'' and financial analysis: Methodological suggestions, *Journal of Finance*, 1(14), 1-10.

Robillard, J. (2018), Millennial Attitudes Towards Financial Advisors and Emerging Investment Technologies, *Wharton Research Scholars*, 171, https://repository.upenn.edu/wharton_research_scholars/171

Rogers, P. (1998), The cognitive psychology of lottery gambling: A theoretical review, *Journal of gambling studies*, 14(2), 111-134.

Rossi, A. G., & Utkus, S. P. (2020), Who Benefits from Robo-advising? Evidence from Machine Learning, *SSRN*, 3552671, dx.doi.org/10.2139/ ssrn.3552671.

Rühr, A. (2020), Robo-Advisor Configuration: An Investigation of User Preferences and the Performance-Control Dilemma, Proceedings of the 28th European Conference on Information Systems (ECIS), *An Online AIS Conference*, June 15-17.

Rühr, A., Berger, B., & Hess, T. (2019), Can I Control My Robo-Advisor? Trade-Offs in Automation and User Control in (Digital) Investment Management, *Proceedings of the 25th Americas Conference on Information Systems (AMCIS)*, 1-10.

Rühr, A., Streich, D., Berger, B., & Hess, T. (2019), A Classification of Decision Automation and Delegation in Digital Investment Systems, *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 1435-1444.

Singh, I., & Kaur, N. (2017), Wealth Management Through Robo Advisory, *International Journal of Research - Granthaalayah*, 5(6), 33-43.

Taylor, E. (2017), Making sense of "algorithm aversion", *Research World*, 64, 57-57.

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

Uhl, M. W., & Rohner, P. (2018), Robo-advisors versus traditional investment advisors: An unequal game, *The Journal of Wealth Management*, 21(1), 44-50.

Ungeheuer, M., & Weber, M. (2021), The perception of dependence, investment decisions, and stock prices, *The Journal of Finance*, 76(2), 797-844.

Vieider, F., Villegas-Palacio, C., Martinsson, P., & Mejía, M. (2016), Risk taking for oneself and others: A structural model approach, *Economic Inquiry*, 2016, 54(2), 879-894.

von Walter, B., Kremmel, D., & Jäger, B. (2022), The impact of lay beliefs about AI on adoption of algorithmic advice, *Marketing Letters*, 33(1), 143-155.

Wärneryd, K.-E. (2001), Stock-market psychology, Cheltenham: Edward Elgar.

Zielonka, P. (2004), Technical analysis as the representation of typical cognitive biases, *International Review of Financial Analysis*, 13, 217-225.

Appendix A: Instructions for the Experiment

Appendix A.1: Instructions (Treatment 'Self')

You have the task of creating portfolios of shares. A portfolio of shares is a compilation of several shares.

The development of the share prices is of no concern to you because you profit only once from the dividend payments of the shares in 2022. The dividend is the distribution of profits of a stock exchange-listed company to its shareholders.

You will receive information about how the dividend payments might turn out, and about the probabilities of different amounts of dividend. In addition, you will be shown how the dividends of the shares have developed over the last ten years.

You are paid the risk-adjusted dividend. A risk-adjusted dividend is the dividend payment divided by the variance of the dividend payments of the selected portfolio. Your task thus consists of achieving the highest possible dividends with the lowest possible risk (low variance).

The total of all risk-adjusted dividends (in ECU) which you achieve via your portfolio decisions is multiplied by five at the end and then paid in euros.

You can make the portfolio decisions yourself or delegate them to an algorithm (robo advisor). The robo advisor is specialized in making meaningful portfolio decisions and takes all of the relevant information into account in an optimal way in order to achieve risk-adjusted dividend payments which are as high as possible.

Appendix A.2: Instructions (Treatment 'Representative')

You have the task of creating portfolios of shares. A portfolio of shares is a compilation of several shares.

The development of the share prices is of no concern to you because you profit only once from the dividend payments of the shares in 2022. The dividend is the distribution of profits of a stock exchange-listed company to its shareholders.

You will receive information about how the dividend payments might turn out, and about the probabilities of different amounts of dividend. In addition, you will be shown how the dividends of the shares have developed over the last ten years.

You are paid the risk-adjusted dividend. A risk-adjusted dividend is the dividend payment divided by the variance of the dividend payments of the selected portfolio. Your task thus consists of achieving the highest possible dividends with the lowest possible risk (low variance).

The total of all risk-adjusted dividends (in ECU) which you achieve via your portfolio decisions is multiplied by five at the end and then paid in euros. However, this amount is not paid to you, but to another participant. If you make successful decisions, one of the other participants will have something to be pleased about. If you make unsuccessful decisions, one of the other participants will be annoyed.

At the same time, another participant is making the decisions which determine your payment. Who has made portfolio decisions for whom will be announced at the end of the session.

So please remember why you made which decisions. The other participant might want you to justify your decisions if the results are disappointing.

You can make the portfolio decisions yourself or delegate them to an algorithm (robo advisor). The robo advisor is specialized in making meaningful portfolio decisions and takes all of the relevant information into account in an optimal way in order to achieve risk-adjusted dividend payments which are as high as possible.

Appendix B: Test questions

Appendix B.1: Test questions (Treatment 'Self')

Test question 1: What is a share portfolio?

a) A compilation of shares, bonds and derivative instruments.
b) A compilation of shares. *(correct*)
c) A compilation of various securities without shares.

Test question 2: What is a dividend?

a) It is the opposite of a multiplication.
b) It is a major military unit.
c) It is the distribution of profits by a stock exchange-listed company to its shareholders. *(correct)*

Test question 3: What do you profit from?

a) From increases in the price of the shares that I choose.
b) From the risk-adjusted dividends of the shares that I choose. *(correct)*
c) From increases in the price of the shares that I choose, and from the dividends.

Test question 4: How can the algorithm (robo advisor) be deployed?

a) I have to use the robo advisor.
b) The robo advisor is not available to me.
c) I have a free choice between either making the portfolio decisions myself or delegating the task to a robo advisor which is specialised in this field. *(correct)*

Appendix B.2: Test questions (Treatment 'Representative')

Test question 1: What is a share portfolio?

d) A compilation of shares, bonds and derivative instruments.
e) A compilation of shares. *(correct*)
f) A compilation of various securities without shares.

Test question 2: From whose decisions do you profit?

d) From my own decisions.
e) From the decisions of all participants.
f) From the decisions of the participant who makes the decisions for me. *(correct)*

Test question 3: What determines the payment of the person for whom you make the decisions?

d) The changes in the prices of the shares that I choose.
e) The risk-adjusted dividends of the shares that I choose. (*correct)*
f) The increases in the price of the shares that I choose, and the dividends of the shares that I choose.

Test question 4: How can the algorithm (robo advisor) be deployed?

d) I have to use the robo advisor.
e) The robo advisor is not available to me.
f) I have a free choice between either making the portfolio decisions myself or delegating the task to a robo advisor which is specialised in this field. *(correct)*

Appendix C: The Tasks

Appendix C.1: Task 1 (Treatment 'Self')

There are two shares to choose from: share Y and share Z. The dividend payments of the two companies are independent random processes with two possible configurations: ECU 8 and ECU 0, and with an expected value of ECU 4. In the table you can see how high the dividend payments of the two shares were in the last 10 years.

Table C-1: Dividend payments of the shares in task 1 of treatment 'Self'

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Share Y	ECU 8	ECU 0	ECU 8	ECU 8	ECU 8	ECU 0	ECU 8	ECU 0	ECU 0	ECU 0	?
Share Z	ECU 0	ECU 0	ECU 8	ECU 0	ECU 8	ECU 8	ECU 0	ECU 0	ECU 8	ECU 8	?

You may choose two shares. As payment you receive the risk-adjusted dividends of the two selected shares. The risk-adjusted dividend corresponds to the dividend payment divided by the variance of the dividend payments of the selected portfolio. Depending on the portfolio selected, you thus receive the risk-adjusted dividends of 2 Y shares, of 2 Z shares, or of 1 Y share + 1 Z share. As the dividend payments are determined by a random process, it is not only the content of the portfolio which determines your payment, but also luck. Which event (ECU 8 or ECU 0) occurs in the case of the two shares is determined separately by drawing lots for each round of the experimental survey.

You can make the portfolio decisions yourself or delegate them to an algorithm (robo advisor). The robo advisor is specialized in making meaningful portfolio decisions and takes all of the relevant information into account in an optimal way. However, the robo advisor also does not know which random event (ECU 8 or ECU 0) will occur as the dividend of the shares. In other words, even when the robo advisor is used, luck determines the payment to a certain extent.

Now make your choice!

- o I will let the robo advisor decide;

I will decide myself and choose:

- o 2 Y shares;
- o 2 Z shares;
- o 1 Y share + 1 Z share.

Appendix C.2: Task 2 (Treatment 'Self')

There are two shares to choose from: share X and share Q. The dividend payments of the two companies are independent random processes with two possible configurations: ECU 4 and ECU 0, and with an expected value of ECU 2. In the table you can see how high the dividend payments of the two shares were in the last 10 years.

Table C-2: Dividend payments of the shares in task 2 of treatment 'Self'

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Share X	ECU 0	ECU 0	ECU 4	ECU 0	ECU 0	ECU 0	ECU 4	ECU 4	ECU 4	ECU 4	?
Share Q	ECU 0	ECU 4	ECU 4	ECU 4	ECU 0	ECU 4	ECU 0	ECU 0	ECU 4	ECU 0	?

You may choose four shares. As payment you receive the risk-adjusted dividends of the four selected shares. The risk-adjusted dividend corresponds to the dividend payment divided by the variance of the dividend payments of the selected portfolio. Depending on

the portfolio selected, you thus receive the risk-adjusted dividends of 4 X shares, of 4 Q shares, of 3 X shares + 1 Q share, of 3 Q shares + 1 X share, or of 2 X shares + 2 Q shares. As the dividend payments are determined by a random process, it is not only the content of the portfolio which determines your payment, but also luck. Which event (ECU 4 or ECU 0) occurs in the case of the two shares is determined separately by drawing lots for each round of the experimental survey.

You can make the portfolio decisions yourself or delegate them to an algorithm (robo advisor). The robo advisor is specialized in making meaningful portfolio decisions and takes all of the relevant information into account in an optimal way. However, the robo advisor also does not know which random event (ECU 4 or ECU 0) will occur as the dividend of the shares. In other words, even when the robo advisor is used, luck determines the payment to a certain extent.

Now make your choice!

- I will let the robo advisor decide;

I will decide myself and choose:

- 4 X shares;
- 4 Q shares;
- 3 X shares + 1 Q share;
- 3 Q shares + 1 X share;
- 2 Q shares + 2 X shares.

Appendix C.3: Task 3 (Treatment 'Self')

There are two shares from a specific sector of industry to choose from (share K and share L). In the table you can see how high the dividend payments of the two shares were in the last 10 years. When business is good in the sector, the dividend of share K is ECU 6, and that of share L is ECU 7. When business is poor in the sector, the dividend of share K is ECU 2, and that of share L is ECU 1. The business situation in the sector can vary from year to year and thus has to be viewed as a random process: the probability of the business situation being either good or poor in 2022 is 50% in each case.

Table C-3: Dividend payments of the shares in task 3 of treatment 'Self'

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Share K	ECU 2	ECU 6	ECU 2	ECU 6	ECU 6	ECU 6	ECU 2	ECU 6	ECU 2	ECU 2	?
Share L	ECU 1	ECU 7	ECU 1	ECU 7	ECU 7	ECU 7	ECU 1	ECU 7	ECU 1	ECU 1	?

You may choose two shares. As payment you receive the risk-adjusted dividends of the two selected shares. The risk-adjusted dividend corresponds to the dividend payment divided by the variance of the dividend payments of the selected portfolio. Depending on the portfolio selected, you thus receive the risk-adjusted dividends of 2 K shares, of 2 L shares, or of 1 K share + 1 L share. As the dividend payments are determined by a random process, it is not only the content of the portfolio which determines your payment, but also luck. Which event (good or poor economic situation in the sector) occurs in the case of the two shares is determined separately by drawing lots for each round of the experimental survey.

You can make the portfolio decisions yourself or delegate them to an algorithm (robo advisor). The robo advisor is specialized in making meaningful portfolio decisions and takes all of the relevant information into account in an optimal way. However, the robo advisor also does not know which random event (good or poor economic situation in the sector) will occur as the dividend of the shares. In other words, even when the robo advisor is used, luck determines the payment to a certain extent.

Now make your choice!

- I will let the robo advisor decide;

I will decide myself and choose:

- 2 K shares;
- 2 L shares;
- 1 K share + 1 L share.

Appendix C.4: Task 4 (Treatment 'Self')

There are two shares from a specific sector of industry to choose from (share M and share P). In the table you can see how high the dividend payments of the two shares were in the last 10 years. When business is good in the sector, the dividend of share M is ECU 4, and that of share P is ECU 3. When business is poor in the sector, the dividend of share M is ECU 0, and that of share P is ECU 1. The business situation in the sector can vary from year to year and thus has to be viewed as a random process: the probability of the business situation being either good or poor in 2022 is 50% in each case.

Table C-4: Dividend payments of the shares in task 4 of treatment 'Self'

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Share M	ECU 4	ECU 0	ECU 4	ECU 0	ECU 0	ECU 0	ECU 4	ECU 4	ECU 0	ECU 4	?
Share P	ECU 3	ECU 1	ECU 3	ECU 1	ECU 1	ECU 1	ECU 3	ECU 3	ECU 1	ECU 3	?

You may choose four shares. As payment you receive the risk-adjusted dividends of the four selected shares. The risk-adjusted dividend corresponds to the dividend payment divided by the variance of the dividend payments of the selected portfolio. Depending on the portfolio selected, you thus receive the risk-adjusted dividends of 4 M shares, of 4 P shares, of 3 M shares + 1 P share, of 3 P shares + 1 M share, or of 2 M shares + 2 P shares. As the dividend payments are determined by a random process, it is not only the content of the portfolio which determines your payment, but also luck. Which event (good or poor economic situation in the sector) occurs in the case of the two shares is determined separately by drawing lots for each round of the experimental survey.

You can make the portfolio decisions yourself or delegate them to an algorithm (robo advisor). The robo advisor is specialized in making meaningful portfolio decisions and takes all of the relevant information into account in an optimal way. However, the robo advisor also does not know which random event (good or poor economic situation in the sector) will occur as the dividend of the shares. In other words, even when the robo advisor is used, luck determines the payment to a certain extent.

Now make your choice!

- I will let the robo advisor decide;

I will decide myself and choose:

- 4 M shares;
- 4 P shares;
- 3 M shares + 1 P share;
- 3 P shares + 1 M share;
- 2 M shares + 2 P shares.

Appendix C.5: Task 1 (Treatment 'Representative')

There are two shares to choose from: share Y and share Z. The dividend payments of the two companies are independent random processes with two possible configurations: ECU 8 and ECU 0, and with an expected value of ECU 4. In the table you can see how high the dividend payments of the two shares were in the last 10 years.

Table C-5: Dividend payments of the shares in task 1 of treatment 'Representative'

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Share Y	ECU 8	ECU 0	ECU 8	ECU 8	ECU 8	ECU 0	ECU 8	ECU 0	ECU 0	ECU 0	?
Share Z	ECU 0	ECU 0	ECU 8	ECU 0	ECU 8	ECU 8	ECU 0	ECU 0	ECU 8	ECU 8	?

You may choose two shares. As compensation, the risk-adjusted dividends are paid from the two selected shares. The risk-adjusted dividend corresponds to the dividend payment divided by the variance of the dividend payments of the selected portfolio. Depending on the portfolio selection, the risk-adjusted dividend of 2 Y shares, of 2 Z shares, or of 1 Y share + 1 Z share is paid out. As the dividend payments are determined by a random process, it is not only the content of the portfolio which determines your payment, but also luck. Which event (ECU 8 or ECU 0) occurs in the case of the two shares is determined separately by drawing lots for each round of the experimental survey.

You can make the portfolio decisions yourself or delegate them to an algorithm (robo advisor). The robo advisor is specialized in making meaningful portfolio decisions and takes all of the relevant information into account in an optimal way. However, the robo advisor also does not know which random event (ECU 8 or ECU 0) will occur as the dividend of the shares. In other words, even when the robo advisor is used, luck determines the payment to a certain extent.

The payment which you achieve with your decision is received by one of the other participants and not by you. This other participant might ask you to justify your choices, so you should think carefully about the decisions you make.

Now make your choice!

- I will let the robo advisor decide;

I will decide myself and choose:

- 2 Y shares;
- 2 Z shares;
- 1 Y share + 1 Z share.

Appendix C.6: Task 2 (Treatment 'Representative')

There are two shares to choose from: share X and share Q. The dividend payments of the two companies are independent random processes with two possible configurations: ECU 4 and ECU 0, and with an expected value of ECU 2. In the table you can see how high the dividend payments of the two shares were in the last 10 years.

Table C-6: Dividend payments of the shares in task 2 of treatment 'Representative'

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Share X	ECU 0	ECU 0	ECU 4	ECU 0	ECU 0	ECU 0	ECU 4	ECU 4	ECU 4	ECU 4	?
Share Q	ECU 0	ECU 4	ECU 4	ECU 4	ECU 0	ECU 4	ECU 0	ECU 0	ECU 4	ECU 0	?

You may choose four shares. As compensation, the risk-adjusted dividends are paid from the four selected shares. The risk-adjusted dividend corresponds to the dividend payment divided by the variance of the dividend payments of the selected portfolio. Depending on the portfolio selection, the risk-adjusted dividend of 4 X shares, of 4 Q shares, of 3 X shares + 1 Q share, of 3 Q shares + 1 X share, or of 2 X shares + 2 Q shares is paid out. As the dividend payments are determined by a random process, it is not only the content of the portfolio which determines your payment, but also luck. Which event (ECU 4 or ECU 0) occurs in the case of the two shares is determined separately by drawing lots for each round of the experimental survey.

You can make the portfolio decisions yourself or delegate them to an algorithm (robo advisor). The robo advisor is specialized in making meaningful portfolio decisions and takes all of the relevant information into account in an optimal way. However, the robo advisor also does not know which random event (ECU 4 or ECU 0) will occur as the dividend of the shares. In other words, even when the robo advisor is used, luck determines the payment to a certain extent.

The payment which you achieve with your decision is received by one of the other participants and not by you. This other participant might ask you to justify your choices, so you should think carefully about the decisions you make.

Now make your choice!

- I will let the robo advisor decide;

I will decide myself and choose:

- 4 X shares;
- 4 Q shares;
- 3 X shares + 1 Q share;
- 3 Q shares + 1 X share;
- 2 Q shares + 2 X shares.

Appendix C.7: Task 3 (Treatment 'Representative')

There are two shares from a specific sector of industry to choose from (share K and share L). In the table you can see how high the dividend payments of the two shares were in the last 10 years. When business is good in the sector, the dividend of share K is ECU 6, and that of share L is ECU 7. When business is poor in the sector, the dividend of share K is ECU 2, and that of share L is ECU 1. The business situation in the sector can vary from year to year and thus has to be viewed as a random process: the probability of the business situation being either good or poor in 2022 is 50% in each case.

Table C-7: Dividend payments of the shares in task 3 of treatment 'Representative'

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Share K	ECU 2	ECU 6	ECU 2	ECU 6	ECU 6	ECU 6	ECU 2	ECU 6	ECU 2	ECU 2	?
Share L	ECU 1	ECU 7	ECU 1	ECU 7	ECU 7	ECU 7	ECU 1	ECU 7	ECU 1	ECU 1	?

You may choose two shares. As compensation, the risk-adjusted dividends are paid from the two selected shares. The risk-adjusted dividend corresponds to the dividend payment divided by the variance of the dividend payments of the selected portfolio. Depending on the portfolio selection, the risk-adjusted dividend of 2 K shares, of 2 L shares, or of 1 K share + 1 L share is paid out. As the dividend payments are determined by a random process, it is not only the content of the portfolio which determines your payment, but also luck. Which event (good or poor economic situation in the sector) occurs in the case of the two shares is determined separately by drawing lots for each round of the experimental survey.

You can make the portfolio decisions yourself or delegate them to an algorithm (robo advisor). The robo advisor is specialized in making meaningful portfolio decisions and takes all of the relevant information into account in an optimal way. However, the robo advisor also does not know which random event (good or poor economic situation in the sector) will occur as the dividend of the shares. In other words, even when the robo advisor is used, luck determines the payment to a certain extent.

The payment which you achieve with your decision is received by one of the other participants and not by you. This other participant might ask you to justify your choices, so you should think carefully about the decisions you make.

Now make your choice!

- I will let the robo advisor decide;

I will decide myself and choose:

- 2 K shares;
- 2 L shares;
- 1 K share + 1 L share.

Appendix C.8: Task 4 (Treatment 'Representative')

There are two shares from a specific sector of industry to choose from (share M and share P). In the table you can see how high the dividend payments of the two shares were in the last 10 years. When business is good in the sector, the dividend of share M is ECU 4, and that of share P is ECU 3. When business is poor in the sector, the dividend of share M is ECU 0, and that of share P is ECU 1. The business situation in the sector can vary

from year to year and thus has to be viewed as a random process: the probability of the business situation being either good or poor in 2022 is 50% in each case.

Table C-8: Dividend payments of the shares in task 4 of treatment 'Representative'

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Share M	ECU 4	ECU 0	ECU 4	ECU 0	ECU 0	ECU 0	ECU 4	ECU 4	ECU 0	ECU 4	?
Share P	ECU 3	ECU 1	ECU 3	ECU 1	ECU 1	ECU 1	ECU 3	ECU 3	ECU 1	ECU 3	?

You may choose four shares. As compensation, the risk-adjusted dividends are paid from the four selected shares. The risk-adjusted dividend corresponds to the dividend payment divided by the variance of the dividend payments of the selected portfolio. Depending on the portfolio selection, the risk-adjusted dividend of 4 M shares, of 4 P shares, of 3 M shares + 1 P share, of 3 P shares + 1 M share, or of 2 M shares + 2 P shares is paid out. As the dividend payments are determined by a random process, it is not only the content of the portfolio which determines your payment, but also luck. Which event (good or poor economic situation in the sector) occurs in the case of the two shares is determined separately by drawing lots for each round of the experimental survey.

You can make the portfolio decisions yourself or delegate them to an algorithm (robo advisor). The robo advisor is specialized in making meaningful portfolio decisions and takes all of the relevant information into account in an optimal way. However, the robo advisor also does not know which random event (good or poor economic situation in the sector) will occur as the dividend of the shares. In other words, even when the robo advisor is used, luck determines the payment to a certain extent.

The payment which you achieve with your decision is received by one of the other participants and not by you. This other participant might ask you to justify your choices, so you should think carefully about the decisions you make.

Now make your choice!

- o I will let the robo advisor decide;

I will decide myself and choose:

- o 4 M shares;
- o 4 P shares;
- o 3 M shares + 1 P share;
- o 3 P shares + 1 M share;
- o 2 M shares + 2 P shares.

Chapter VI

Willingness to Use Algorithms Varies with Social Information on Weak vs. Strong Adoption: An Experimental Study on Algorithm Aversion

Contribution Jan René Judek: 100%

Submitted to:
Journal of Behavioral Decision Making

Also published as:

Discussion Paper in Sonderforschungsgruppe Institutionenanalyse (sofia), No. 22-5, Darmstadt, December 2022.

Working Paper in Wolfsburg Working Papers (WWP), Ostfalia University of Applied Sciences, No. 23-01, Wolfsburg, January 2023.

Abstract

The process of decision-making is increasingly supported by algorithms in a wide variety of contexts. However, the phenomenon of algorithm aversion conflicts with the development of the technological potential that algorithms bring with them. Economic agents tend to base their decisions on those of other economic agents. Therefore, this experimental approach examines the willingness to use an algorithm when making stock price forecasts when information about the prior adoption of an algorithm is provided. It is found that decision makers are more likely to use an algorithm if the majority of preceding economic agents have also used it. Willingness to use an algorithm varies with

social information about prior weak or strong adoption. In addition, the affinity for technology interaction of the economic agents shows an effect on decision behavior.

Keywords

Algorithm aversion; Algorithmic decision-making; Herding behavior; Decision aids; Forecasting; Behavioral finance; Experiments.

JEL Classification

D81; D91; G17; G41; O33.

1 Literature Review and Hypothesis Development

In a wide variety of domains, such as asset management (Niszczota & Kaszás, 2020; Méndez-Suárez, García-Fernández & Gallardo, 2019), justice (Ireland, 2020; Simpson, 2016), medicine (Beck et al., 2011; Ægisdóttir et al., 2006; Grove et al., 2000), sports (Pérez-Toledano et al., 2019), or predictive policing (Mohler et al., 2015), stochastic models or algorithms are increasingly used to make predictions. The forecasting performance of these models is often superior to the forecasting performance of humans (Castelo, Bos & Lehmann, 2019; Youyou, Kosinski & Stillwell, 2015; Dawes, Faust & Meehl, 1989; Meehl, 1954). Algorithms can identify complex connections in large data sets where humans reach their cognitive limits. Nonetheless, it appears that rejection of predictions by automated methods is widespread (for a literature review, see Alvarado-Valencia & Barrero, 2014), and people often opt out of using superior algorithms and instead choose less accurate predictions by humans (Dietvorst, Simmons & Massey, 2015; Önkal et al., 2009; Highhouse, 2008).

The negative attitude towards algorithms is referred to as algorithm aversion. This describes the fact that economic agents refrain from using an algorithm as soon as they realize that it is superior but still not error-free (Prahl & Van Swol, 2017; Dietvorst, Simmons & Massey, 2015). While people often respond to algorithm prediction errors by showing a strong rejection attitude toward them, it is less pronounced for prediction errors made by humans (Dietvorst, Simmons & Massey, 2015). Economic agents underestimate the accuracy of stochastic models and prefer predictions made by humans (Önkal et al., 2009). Given the power of algorithms in forecasting, algorithm aversion is particularly harmful by not using them, economic agents result in using inferior forecasts. In other words, by using human forecasts instead of algorithmic forecasts, it reduces the chance of success. Although algorithms outperform the quality of human forecasts, the maximum benefit can only be achieved in the long run if economic agents give preference to algorithmic forecasts over human forecasts (for detailed literature reviews on algorithm aversion, see Mahmud et al., 2022; Burton, Stein & Jensen, 2020; Jussupow, Benbasat & Heinzl, 2020).

Algorithm aversion occurs primarily when economic agents interact with algorithms that do not make error-free predictions, and thus economic agents occasionally are given bad advice (Prahl & Van Swol, 2017; Dietvorst, Simmons & Massey, 2015). In the context of research on algorithm aversion, the decision-making behavior of economic agents is considered in different contexts. For example, the perceived objectivity of a task affects the willingness to use an algorithm. Economic agents are more willing to use an algorithm if it performs an apparently objective task rather than an apparently subjective task. However, perceived objectivity is malleable via description, and as the objectivity of a task increases, so does the willingness to use an algorithm (Castelo, Bos & Lehman, 2019). The response time of an algorithm also has an impact on the willingness of economic agents to use it. Forecasts generated slowly by algorithms are perceived as less reliable and therefore used less frequently than forecasts that are generated quickly (Efendić, van de Calseyde & Evans, 2020).

Economic agents who gain experience with an algorithm by working on incentivized, similar tasks under regular feedback learn to better assess the limits of their own abilities and use an algorithm more often (Filiz et al., 2021). Another approach shows that perceived learning from mistakes by algorithms and humans has an impact on algorithm aversion. After making mistakes, algorithms are perceived as less capable of learning compared to humans. However, if evidence is provided that an algorithm can learn from mistakes, this leads to higher trust and more frequent use of the algorithm (Reich, Kaju & Maglio, 2022).

There are other ways to mitigate algorithm aversion such as humanizing algorithms (Hodge, Mendoza & Sinha, 2021; Castelo, Bos & Lehmann, 2019), providing different explanations of the automated forecast result (Ben David, Resheff & Tron, 2021), or providing a suitable representation of the automated forecast result (Kim, Giroux & Lee, 2021). If economic agents are granted a possibility to influence the forecast result in the form of a subsequent adjustment, the willingness to use it increases significantly. This holds true even when the possibilities for adjusting the forecast result are severely limited (Dietvorst, Simmons & Massey, 2018).

Economic agents tend to align their behavior with the behavior of other economic agents (Spyrou, 2013; Raafat, Chater & Frith, 2009). Social orientation to the behavior of others is referred to as herding behavior. This describes the phenomenon that economic agents follow the actions of other economic agents (a herd), regardless of whether the actions are rational or irrational (Baddeley et al., 2012). Non-rational herd behavior occurs when economic agents blindly mimic the actions of other economic agents and largely forgo the incorporation of rational considerations into decision making (Baddeley et al., 2012; Devenow & Welch, 1996). The ability to observe the decisions of other economic agents (for example, the investment decision of a colleague) can also lead to herd behavior (Devenow & Welch, 1996). Thus, by imitating the actions of others, the behavior of many individual economic agents can become aligned (Spyrou, 2013; Raafat, Chater & Frith, 2009; Hirshleifer & Teoh, 2003).

Herd behavior is observable in stock markets; for example, during stock market crashes, which is studied in the context of investment and financial decisions (Mavruk, 2022; Deng, 2013; Baddeley et al., 2012; Bikhchandani & Sharma, 2000). The events surrounding GameStop's stock from the winter of 2020 into the spring of 2021 demonstrated the powerful impact of herd behavior. Retail investors initiated a short squeeze of institutional investors who had bet on a decline in the stock price. As a result, GameStop's stock price jumped from about $10 in October 2020 to as high as $480 in January 2021, leading to substantial losses on the part of institutional investors who had bet on a decline in the stock price (Lyócsa, Baumöhl & Výrost, 2021; Vasileiou, Bartzou & Tzanakis, 2021; Chohan, 2021). In this context, Betzer and Harries (2022) show a positive relationship between coordinated activities on social media and various trading measures.

The present study focuses on the decision-making behavior of economic agents who have an algorithm at their disposal during the decision process. It is possible that the influence of other economic agents' decisions may have an effect on the extent of algorithm aversion. This paper therefore aims to investigate whether economic agents interacting with algorithms are influenced by the decisions of others and adapt their own decisions to the behavior of others. To this end, an incentivized economic experiment is conducted in which economic agents receive information about the (low or high) willingness to use an algorithm from previously deciding economic agents before deciding whether to use an algorithm themselves. It is of interest whether economic agents mimic each other's behavior and are more or less willing to use an algorithm. Convergent social behavior (Raafat, Chater & Frith, 2009) could lead to economic agents being more willing to use an algorithm if prior decision makers have chosen to use an algorithm by a majority (high utilization rate) and vice versa. In areas such as social commerce, it has also been shown that information about what others are doing can increase trust in new technologies and drive sales (Hajli et al., 2014; Amblee & Bui, 2011). Alexander, Blinder and Zak (2018) show that the availability of social information about the use of an algorithm may have an impact on the willingness to use it. Therefore, it is hypothesized:

H$_1$: Economic agents who receive information about prior high adoption of an algorithm are more likely to use an algorithm than economic agents who receive information about prior low adoption.

This could have interesting implications for practice: algorithm aversion could be reduced by providing economic agents with information about the high willingness to use the algorithm. If this contributes to a more frequent use of a powerful algorithm, an increase in economic efficiency can be enabled at the same time. Considering the forecasting success of algorithms, it is advisable to use them over forecast predictions of humans in most cases (Dietvorst, Simmons & Massey, 2015; Önkal et al., 2009). Nevertheless, many economic agents are reluctant to use an algorithm, thereby reducing the quality of their forecasts. They deliberately forgo the use of a superior algorithm at the expense of their forecasting success, preferring their own forecasts (Burton, Stein & Jensen, 2020).

2 Research Methods

Participants

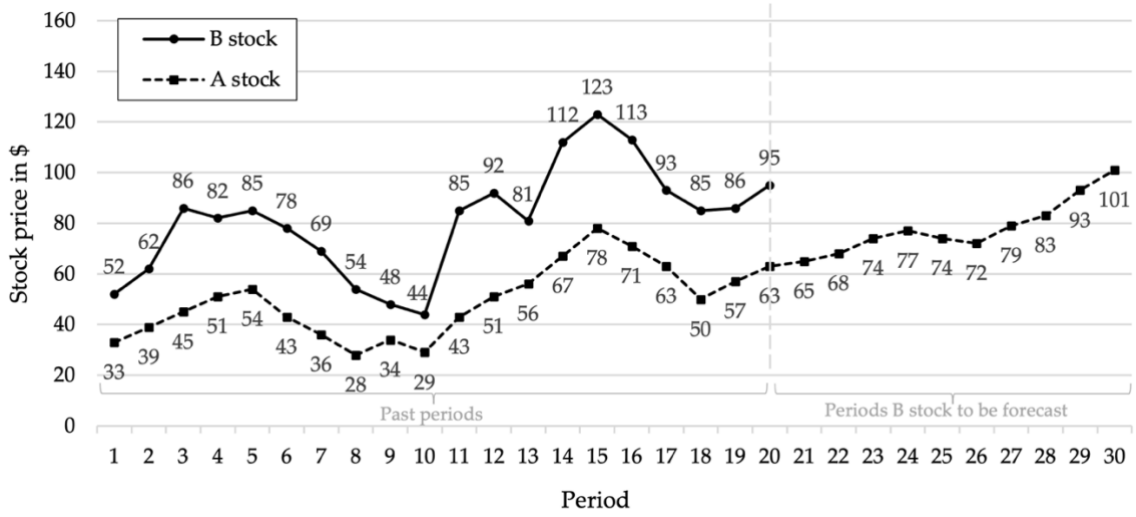
For participation in the economic experiment, 285 subjects were recruited online via Amazon Mechanical Turk (MTurk) and Cloud Research. 31 subjects were excluded from the analysis due to incorrect answering of at least one comprehension question (on a maximum of two attempts) or failure to pass an attention check. This leaves 254 subjects as the sample, of which 50.4% are female and 49.6% are male. The mean age is 40.6 years (σ_{age} = 10.97). The experiment was programmed as a survey in Qualtrics. The survey was

administered on November 28, 2022. The average completion time was 7.02 minutes. Subjects received a fixed show-up fee of $0.30 and a performance-based bonus that could be as high as $1.67.

Design

To conduct the economic experiment, a task on stock price forecasting was designed. Forecasting tasks in this domain have also been used in other studies examining decision behavior in cooperation with algorithms or stochastic models (Gubaydullina et al., 2022; Castelo, Bos & Lehman, 2019; Önkal et al., 2009). Subjects are told that the task is to make ten stock price predictions. Subjects are provided with the stock price of the A stock for periods 1 to 30 on the one hand and the stock price of the B stock for periods 1 to 20 on the other hand (Figure 1). The forecast object is the stock price of the B stock in periods 21 to 30. Subjects are further told that the companies of the A stock and the B stock operate in the same industry and are therefore closely related. That is, the success of the A-stock company is closely related to the success of the B-stock company. As a result, a rising price of the A stock is likely to be accompanied by a rising price of the B stock, and vice versa. Thus, subjects can draw conclusions about the development of the price of the B stock from the development of the price of the A stock. In fact, the prices of the A stock and the B stock have a correlation coefficient of 0.94 in periods 1 to 20.

Figure 1: Stock price development of the A stock and B stock



Next, the subjects learn about the incentive model, which consists of two components. On the one hand, a fixed show-up fee of $0.30 is paid and, on the other hand, a performance-related bonus is paid, which is based on the accuracy of the forecasts made and is higher the more accurate the forecasts are. To determine the amount of the performance-related bonus, the percentage deviation from the actual stock price is calculated for each individual forecast and paid on a sliding scale (Table 1). A forecast is only rewarded if it deviates by a maximum of 15 percent from the actual stock price. In this way, a maximum payment of $1.97 can be achieved. The compensation is earned in

Coins during the experiment and exchanged at a conversion rate of 300 Coins = $1 at the end of the experiment.

Table 1: Grading of the performance-related bonus by accuracy of each forecast

Maximum Deviation in %	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	>15
Bonus in Coins	50	47	43	40	37	33	30	27	23	20	17	13	10	7	3	0

Subsequently, the subjects are informed that, in addition to the possibility of making their own stock price forecasts, a forecasting computer (algorithm) is available. The subjects are informed that in the past the stock price forecasts of the algorithm deviated by a maximum of 10 percent from the actual stock price in 6 out of 10 cases. In addition, subjects receive the information about the low versus high acceptance of the algorithm from the pre-survey, depending on the treatment. They are informed that the algorithm has the same information about the stock price trends of the A stock and the B stock as the subjects. Before submitting their forecasts, the subjects have a one-time choice of whether their own stock price forecasts or the algorithm's stock price forecasts should be used to determine the performance-based bonus. This approach is in line with other studies on algorithm aversion (Dietvorst, Simmons & Massey, 2018; Dietvorst, Simmons & Massey, 2015). The order for displaying the two options is randomized. However, regardless of the choice, subjects must make their own stock price predictions.

The study is designed as a between-subjects design. Subjects are randomly assigned to one of two treatments. In Treatment 1 (social information about low acceptance), subjects are informed about the low acceptance to use the algorithm by other economic agents in the pre-survey, in addition to the accuracy of the algorithm. In Treatment 2 (social information about high acceptance), the subjects are informed about the high acceptance of the algorithm by other economic agents in the pre-survey, in addition to the accuracy of the algorithm.

The operation of the forecasting calculator (algorithm) is based on a linear OLS regression with the stock prices of the A stock and the B stock in periods 1 to 20 (in-sample range). The resulting regression equation ($K_{B_t} = 1.43K_{A_t} + 10.47$) is used to forecast the prices of the B stock in the out-of-sample range of periods 21 to 30, taking the price of the A stock as the independent variable. For example, to predict the B stock price in the first period to be forecast (period 21), the A stock price of $65 (Figure 1) is substituted into the regression equation ($K_{B_{21}} = 1.43 \times 65 + 10.47$). This results in a forecast of the forecasting calculator of $103.4 for the price of the B stock in period 21 and so on.

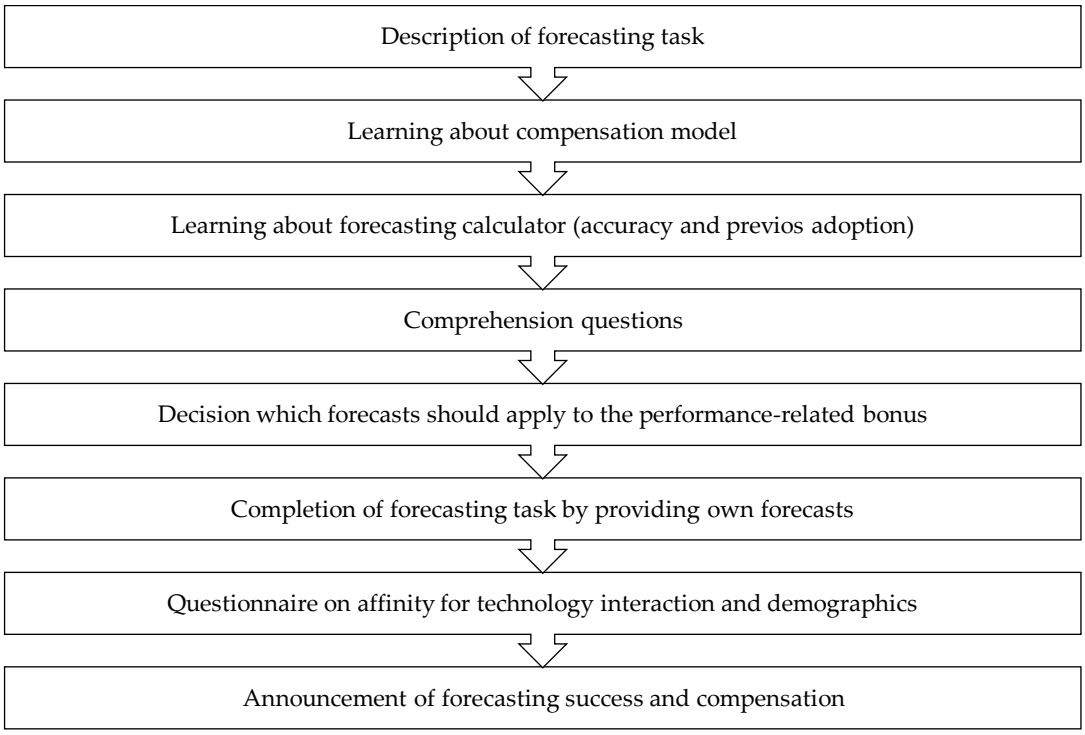
To determine the prior low or high acceptance rate of the algorithm reported in the main study, a pre-survey was conducted in the same setting. The subjects of the pre-survey (n = 29; $\bar{x}_{Age}$ = 35.48; σ_{Age} = 10.47; 41.4% female) were divided into two halves

according to the achieved score after answering a questionnaire to assess affinity for technology interaction (Franke, Attig & Wessel, 2019). For the top half, 35.71% of subjects used the available algorithm and in the bottom half, 71.43% of subjects used it.

Procedure

Subjects first learn about the forecasting task and the forecasting object by reading the instructions (complete procedure see Figure 2). The graphical development of the available stock prices of the A stock and the B stock is shown (Figure 1). Subjects are given information about the compensation model. Subjects are informed that a forecasting calculator can be used. They receive information about the accuracy of the forecasts of the forecasting calculator and, depending on the treatment, about the previous low or high adoption. Subsequently, the subjects answer some comprehension questions to make sure that the task has been understood. A maximum of two attempts are available for answering. In the next step, the subjects decide whether their own forecasts or the forecasts of the forecasting computer are to be used to determine the performance-related bonus. Regardless of the decision, the subjects then complete the forecasting task and submit ten of their own stock price forecasts. Subsequently, subjects answer a questionnaire consisting of nine items to assess affinity for technology interaction (Franke, Attig & Wessel, 2019) and a short demographic questionnaire. For the attention check, the selection of an option is specified in an additional question. If this is not selected, the control is not passed. Last, subjects are informed about the success of the forecasts used and the compensation achieved.

Figure 2: Procedure of the study



3 Results

Forecast accuracy

The analysis of the accuracy of the forecasts made, which were chosen as the basis for the compensation, shows that the stock price forecasts of the forecast calculator are more accurate than the stock price forecasts of the subjects (Table 2). While the average deviation between the predicted and actual stock price of the subjects' forecasts who chose their own forecast as the basis for compensation is \$20.51 (or 18.51%), the forecasts of the forecasting calculator have a forecast error of only \$10.90 (or 8.56%) (absolute forecast error: $t(252) = 16.21$; $p < 0.001$; relative forecast error: $t(252) = 14.19$; $p < 0.001$). The subjects' forecasts show a forecast error up to 88% higher than the forecasts of the forecasting calculator. Also, when considering the bonus paid for forecast accuracy, it can be seen that using the forecasting calculator results in a bonus that is approximately 63% higher ($t(252) = 17.47$; $p < 0.001$). Thus, in terms of forecasting success and the resulting bonus, it is advisable to use the forecasting calculator. On average, the forecasts given by the subjects lead to a lower forecast success and thus to a lower performance-related bonus.

Table 2: Accuracy of forecasts: Own forecasts vs. forecasting calculator

	Basis of performance-related bonus		t-test
	Own forecasts	Forecasting calculator (algorithm)	
Ø absolute forecast error [in \$]	20.51	10.90	$t(252) = 16.21$; $p < 0.001$; $d = 2.06$
Ø relative forecast error [in %]	18.51	8.56	$t(252) = 14.19$; $p < 0.001$; $d = 1.80$
Ø performance-related bonus [in \$]	0.51	0.83	$t(252) = 17.47$; $p < 0.001$; $d = 2.22$

Willingness to use the algorithm

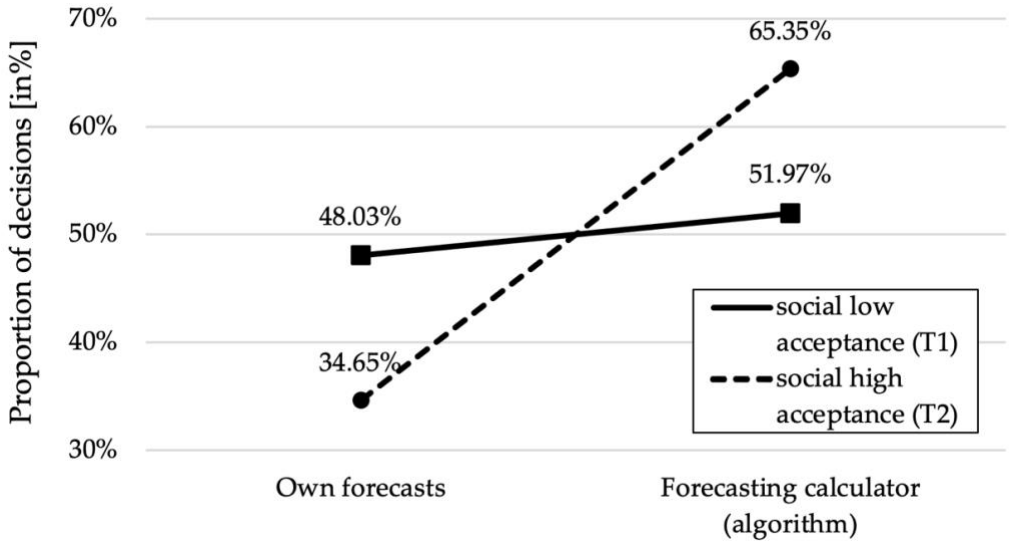
Despite the higher accuracy of the algorithm's stock price forecasts, a large proportion of the subjects refrain from using the algorithm as the basis for determining the performance-related bonus. Overall, 41.34% of the subjects prefer their own stock price forecasts to the forecasts of the forecasting calculator, thereby reducing their forecast success.

Table 3: Decisions in the presence of social information about low vs. high acceptance

	Total	Forecasting calculator (algorithm)		Own forecasts	
	n	n	%	n	%
Social low acceptance (T1)	127	66	51.97%	61	48.03%
Social high acceptance (T2)	127	83	65.35%	44	34.65%

The hypothesis H_1 states that subjects who are informed about a prior high acceptance of an algorithm use it more often than subjects who receive information about a prior low acceptance. The only difference between the two treatments is the information about the low or high adoption of the forecasting calculator, which is obtained from the pre-survey. In fact, the subjects' decision behavior to use the forecasting calculator differs between the treatments. When informed about the low acceptance of an algorithm (T1), 51.97% of subjects use the forecasting calculator to determine the performance-based bonus (Table 3; Figure 3). In contrast, when informed about the high acceptance of an algorithm (T2), 65.35% of the subjects use the forecasting calculator (χ^2 (n = 254) = 4.69; p = 0.030). Thus, H_1 cannot be rejected. Social information about frequent use of the algorithm leads subjects to use the algorithm significantly more often. Information about prior willingness to use an algorithm from other economic agents has an impact on the decision to use an algorithm.

Figure 3: Decisions in the presence of social information about low vs. high acceptance



Further analyses show that the choice to use the algorithm is influenced by gender. In both, Treatment 1 (χ^2 (n = 127) = 3.69; p = 0.055) and Treatment 2 (χ^2 (n = 127) = 8.14; p = 0.004), women use the algorithm more frequently than men. The comparison of treatments by gender shows that the effect is mainly driven by women (χ^2 (n = 128) = 3.20; p = 0.073) and less by men (χ^2 (n = 126) = 0.70; p = 0.402). While at low adoption (T1) 61.40% of women use the algorithm, at high adoption (T2) 76.06% of women already use the algorithm. For men, on the other hand, at T1 (and T2, respectively), 44.29% (51.79%) use the algorithm (Table 4). In contrast, the age of the subjects shows no statistically significant effect on decisions (t(252) = 1.97; p = 0.278).

Table 4: Decision-making behavior by gender

	Gender	Forecasting calculator (algorithm)		Own forecast	
		n	%	n	%
Social low acceptance (T1)	male	31	44.29%	39	55.71%
	female	35	61.40%	22	38.60%
Social high acceptance (T2)	male	29	51.79%	27	48.21%
	female	54	76.06%	17	23.94%

Regarding affinity for technology interaction (ATI), the subjects had a mean ATI score of 3.84 overall. An ATI score of 1.00 corresponds to a low affinity for technology interaction and an ATI score of 6.00 to a high affinity. Considering the ATI score in general, there are almost no differences between the treatments: In both treatments, the proportion of subjects showing a low ATI score is about 30% and the proportion of subjects showing a high ATI score is about 70% (Table 5). Thus, most subjects exhibit a high affinity for technology interaction. On the other hand, the differences in the decision behavior to use the algorithm considering the ATI score are notable. In particular, subjects who have a low ATI score are more likely to use the forecasting calculator. While 64.86% of subjects who have a low ATI score use the algorithm when given information about low acceptance, as many as 84.21% of subjects choose the algorithm when given information about high acceptance (χ^2 (n = 75) = 3.71; p = 0.054). Subjects who have a high ATI score use the algorithm in 46.67% of cases when they receive information about low acceptance, and in 57.30% of cases when they receive information about high acceptance (χ^2 (n = 179) = 2.03; p = 0.154). Thus, in particular, subjects (84.21%) who have low ATI are more likely to be influenced and persuaded to use the algorithm by social information about high acceptance than subjects (57.30%) who have high ATI (χ^2 (n = 127) = 8.52; p = 0.004).

Table 5: Decision-making behavior by ATI score

	ATI score*	Total		thereof use algorithm	thereof use own forecasts
		n	%	%	%
Social low acceptance (T1)	≤ 3.5	37	29.13%	64.86%	35.14%
	> 3.5	90	70.87%	46.67%	53.33%
Social high acceptance (T2)	≤ 3.5	38	29.92%	84.21%	15.79%
	> 3.5	89	70.08%	57.30%	42.70%

*The ATI score can take values from 1.00 (low ATI) to 6.00 (high ATI).

4 Discussion

Aversion to using algorithms, which may be more successful on average is costly in a forecasting process because algorithmic offerings go unused and decision makers do not benefit from higher accuracy that predictions from algorithms often provide (Reich, Kaju & Maglio, 2022). Algorithm aversion is shown to decrease due to the ability to customize algorithmic prediction (Dietvorst, Simmons & Massey, 2018). Nevertheless, a conflict of interests arises here: overall, while the possibility of adjustment increases the acceptance to use algorithms, the adjustments made simultaneously decrease the quality of the final decisions (Sele & Chugunova, 2022).

The results of the present study show that a reduction in algorithm aversion can be achieved even without adjusting the algorithmic prediction. Thus, any worsening of the final decisions due to adjustments can be ruled out. Economic agents also tend to be guided by the decisions of other economic agents in the process of algorithmic decision making, which is consistent with findings on herd behavior (Spyrou, 2013). Economic agents who are informed about an algorithm's prior strong adoption by other economic agents in addition to its accuracy are significantly more likely to use an algorithm than economic agents who are informed about its prior weak adoption in addition to its accuracy. However, the results also show that this effect occurs mainly among women and less among men. There is evidence that overconfidence can exert an influence on algorithm aversion (Filiz et al., 2021). Spiwoks and Bizer (2018) show that men's and women's judgments diverge sharply when making stock price predictions, and women tend to be more underconfident. This could also affect the decisions to use an algorithm in the present study

Alexander, Blinder, and Zak (2018) examine willingness to use an automated aid in solving a maze in four treatments (no information, information about accuracy, and information about low or high social acceptability). In the two treatments that provide information about the social acceptance of the automated aid, subjects are presented with a social acceptance of 54% and 70%, respectively. As a result, it appears that social information about acceptance (regardless of the extent of acceptance), that is, knowledge that others have used the assistive device, is most likely to persuade economic agents to use the assistive device themselves (Alexander, Blinder & Zak, 2018). However, it is important to note as a limitation that the study was conducted with a small number of subjects who were assigned to two of four treatments. In addition, the technical aid is not the identifiable best option and subjects must use a part of their earnings to use the aid. Last, this is not a classic prediction task that can be either automated or performed by a human, but rather an assistive device that can facilitate solving the maze on its own. The present results, unlike the results of Alexander, Blinder, and Zak (2018), show that social information about high acceptance is particularly likely to persuade economic agents to use an algorithm. In contrast, when social information is about low adoption, significantly fewer economic agents are willing to use an algorithm. Thus, the willingness to use an algorithm varies with information about weak or strong adoption.

The present results also indicate that primarily economic agents with a low ATI are influenced by the decisions of other economic agents. While tech-savvy economic agents are only slightly influenced by prior adoption information, non- or low-tech-savvy economic agents are significantly more likely to use an algorithm when given information about strong adoption than when given information about weak adoption. This could indicate that the more uninformed economic agents are about the capabilities of an algorithm, the more likely they are to trust the presumed "swarm intelligence" of society or the decisions of previous decision makers.

Tech-savvy economic agents are by no means in favor of the use of technology of any kind. However, tech-savvy economic agents might have a higher awareness of when it makes sense to use technology and when to refrain from using it. Establishing a connection between the actual motivations for using an algorithm (e.g., accuracy, time pressure, habit, etc.) and algorithm aversion, as well as identifying additional ways to reduce algorithm aversion, is left to future research. The present study uses a stock price prediction task. Although stock price prediction can generally be considered a difficult task, the design of this study allows even non-experts to make a prediction. Nevertheless, it is possible that there are areas where decision makers are more likely to use algorithms than in other areas and these circumstances may affect outcomes. Factors other than ATI or gender may exert an influence on economic agents' decisions to use an algorithm. However, this study provides preliminary insights into ATI, which may have an impact on algorithm aversion.

5 Conclusion

Algorithm aversion causes economic agents to refrain from using superior algorithms as soon as they realize that algorithms may be prone to error. Own forecasts are preferred to the forecasts of algorithms, which can lead to lower forecasting success overall. In the present study, it can be seen in stock price forecasts that a forecasting computer using a simple linear regression to generate its forecast is superior to the subjects' own forecasts. Nevertheless, a large proportion of subjects refrain from using the forecast calculator and use subpar forecasts of their own.

The present study shows that the decision to use an algorithm takes into account the prior behavior of other economic agents in the decision-making process. If economic agents are informed not only about the accuracy of an algorithm, but also about a high adoption among other economic agents, they are significantly more likely to use an algorithm than if they are informed about a previous low adoption. Information about the weak or strong acceptance of an algorithm, i.e., about how other economic agents have decided, has an impact on algorithm aversion. Similarly, providing information about strong adoption leads to a reduction of algorithm aversion. Economic agents decide to use an algorithm by a majority, when earlier decisive economic agents also decided to use an algorithm by a majority. Economic agents who have a low affinity for technology

interaction (ATI) are more likely to use the algorithm than economic agents who have a high ATI due to the social information of strong adoption.

In summary, the willingness to use an algorithm varies with social information about weak or strong adoption. Thus, providing information about the high willingness of other economic agents to use an algorithm can help increase the willingness to use an algorithm in economic practice. This may contribute to an improvement in overall forecast quality and an increase in economic efficiency. Nevertheless, more research is needed to identify causes and further ways to reduce algorithm aversion.

Acknowledgements

I would like to thank Markus Spiwoks, Ibrahim Filiz, and Marco Lorenz for constructive comments and helpful discussions during the preparation of the study as well as Albert Heinecke for his support throughout the project. I also thank Galen Jiang for valuable advice in reviewing my translation.

References

Ægisdóttir, S., White, M.J., Spengler, P.M., Maugherman, A.S., Anderson, L.A., Cook, R.S., Nichols, C.N., Lampropoulos, G., Walker, B.S., Cohen, G.R., & Rush, J.D. (2006). The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction. *The Counseling Psychologist*, 34(3), 341-382.

Alexander, V., Blinder, C., & Zak, P.J. (2018). Why trust an algorithm? Performance, cognition, and neurophysiology. *Computers in Human Behavior*, 89, 279-288.

Alvarado-Valencia, J.A., & Barrero, L.H. (2014). Reliance, trust and heuristics in judgmental forecasting. *Computers in Human Behavior*, 36, 102-113.

Amblee, N., & Bui, T.X. (2011). Harnessing the Influence of Social Proof in Online Shopping: The Effect of Electronic Word of Mouth on Sales of Digital Microproducts. *International Journal of Electronic Commerce*, 16(2), 91-114.

Baddeley, M., Burke, C.J., Schultz, W., & Tobler, P.N. (2012). Herding in Financial Behaviour: A Behavioural and Neuroeconomic Analysis of Individual Differences. https://doi.org/10.17863/CAM.1041.

Beck, A., Sangoi, A., Leung, S., Marinelli, R. J., Nielsen, T., Vijver, M. J., West, R., Rijn, M.V., & Koller, D. (2011). Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival. *Science Translational Medicine*, 3(108), 108-113.

Ben David, D., Resheff, Y.S., & Tron, T. (2021). Explainable AI and Adoption of Algorithmic Advisors: an Experimental Study. *ArXiv, abs/2101.02555*.

Betzer, A., Harries, J.P. (2022). How online discussion board activity affects stock trading: the case of GameStop. *Financial Markets and Portfolio Management*, 36(4), 443-472.

Bikhchandani, S., & Sharma, S.K. (2000). Herd Behavior in Financial Markets. *IMF Staff Papers*, 47(3), 279-310.

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

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

Chohan, U.W., YOLO Capitalism (2022). Available at SSRN 3775127.

Dawes, R.M., Faust, D., & Meehl, P.E. (1989). Clinical versus actuarial judgment. *Science*, 243(4899), 1668-1674.

Deng, G. (2013). The Herd Behavior of Risk-Averse Investor Based on Information Cost. *Journal of Financial Risk Management*, 2(4), 87-91.

Devenow, A., & Welch, I. (1996). Rational herding in financial economics. *European Economic Review, 40(3-5)*, 603-615.

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.

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), 114-126.

Efendić, E., Van de Calseyde, P.P., & Evans, A.M. (2020). Slow response times undermine trust in algorithmic (but not human) predictions. *Organizational Behavior and Human Decision Processes*, 157, 103-114.

Filiz, I., Judek, J.R., Lorenz, M., & Spiwoks, M. (2021). Reducing algorithm aversion through experience. *Journal of Behavioral and Experimental Finance*, 31, 100524.

Franke, T., Attig, C., & Wessel, D. (2019). A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction*, 35(6), 456-467.

Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E., & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychological Assessment*, 12(1), 19-30.

Gubaydullina, Z., Judek, J.R., Lorenz, M., & Spiwoks, M. (2022). Comparing Different Kinds of Influence on an Algorithm in Its Forecasting Process and Their Impact on Algorithm Aversion. *Businesses*, 2(4), 448-470.

Hajli, N., Lin, X., Featherman, M., & Wang, Y. (2014). Social Word of Mouth: How Trust Develops in the Market. *International Journal of Market Research*, 56(5), 673-689.

Highhouse, S. (2008). Stubborn Reliance on Intuition and Subjectivity in Employee Selection. *Industrial and Organizational Psychology*, 1(3), 333-342.

Hirshleifer, D., & Hong Teoh, S. (2003). Herd behaviour and cascading in capital markets: A review and synthesis. *European Financial Management*, 9(1), 25-66.

Hodge, F.D., Mendoza, K.I., & Sinha, R.K. (2021). The effect of humanizing robo-advisors on investor judgments. *Contemporary Accounting Research*, 38(1), 770-792.

Ireland, L. (2020). Who errs? Algorithm aversion, the source of judicial error, and public support for self-help behaviors. *Journal of Crime and Justice*, 43(2), 174-192.

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

Kim, J., Giroux, M., & Lee, J.C. (2021). When do you trust AI? The effect of number presentation detail on consumer trust and acceptance of AI recommendations. *Psychology & Marketing*, 38(7), 1140-1155.

Lyócsa, Š., Baumöhl, E., & Výrost, T. (2021). YOLO trading: Riding with the herd during the GameStop episode. *Finance Research Letters*, 46(A), 102359.

Mahmud, H., Islam, A.N., Ahmed, S.I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, 175, 121390.

Mavruk, T. (2022). Analysis of herding behavior in individual investor portfolios using machine learning algorithms. *Research in International Business and Finance*, 62, 101740.

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

Méndez-Suárez, M., García-Fernández, F., & Gallardo, F. (2019). Artificial Intelligence Modelling Framework for Financial Automated Advising in the Copper Market. *Journal of Open Innovation: Technology, Market, and Complexity*, 5(4), 81.

Mohler, G.O., Short, M.B., Malinowski, S., Johnson, M.E., Tita, G.E., Bertozzi, A., & Brantingham, P.J. (2015). Randomized Controlled Field Trials of Predictive Policing. *Journal of the American Statistical Association*, 110, 139-1411.

Niszczota, P., & Kaszás, D. (2020). Robo-investment aversion. *PLoS ONE*, 15(9), 0239277, 1-19.

Önkal, D., Goodwin, P., Thomson, M.E., Gönül, S., & Pollock, A.C. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, 22(4), 390-409.

Pérez-Toledano, M., Rodriguez, F.J., García-Rubio, J., & Ibáñez, S.J. (2019). Players' selection for basketball teams, through Performance Index Rating, using multiobjective evolutionary algorithms. *PLoS ONE,* 14(9), 0221258, 1-20.

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

Raafat, R.M., Chater, N., & Frith, C. (2009). Herding in humans. *Trends in Cognitive Sciences*, 13(10), 420-428.

Reich, T., Kaju, A., & Maglio, S.J. (2022). How to overcome algorithm aversion: Learning from mistakes. *Journal of Consumer Psychology,* ahead-of-print, 1-18.

Sele, D., & Chugunova, M. (2022). Putting a Human in the Loop: Increasing Uptake, but Decreasing Accuracy of Automated Decision-Making. *Max Planck Institute for Innovation & Competition Research Paper No. 22-20*. Available at SSRN 4285645.

Simpson, B. (2016). Algorithms or advocacy: does the legal profession have a future in a digital world? *Information & Communications Technology Law*, 25(1), 50-61.

Spiwoks, M., & Bizer, K. (2018). On the Measurement of Overconfidence: An Experimental Study. *International Journal of Economics and Financial Research*, 4(1), 30-37.

Spyrou, S.I. (2013). Herding in financial markets: a review of the literature. *Review of Behavioral Finance*, 5, 175-194.

Vasileiou, E., Bartzou, E., & Tzanakis, P. (2021). Explaining Gamestop Short Squeeze using Intraday Data and Google Searches. Available at SSRN 3805630.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4), 1036-1040.

Chapter VII

Interest Rate Forecasts in Latin America

Co-authored by Ibrahim Filiz, Marco Lorenz, and Markus Spiwoks
Contribution Jan René Judek: 45%

Published:
Journal of Economic Studies, Vol. 49, Issue 5, 920-936. (July 2022)
https://doi.org/10.1108/JES-02-2021-0108

Also published as:

Discussion Paper in Sonderforschungsgruppe Institutionenanalyse (sofia), No. 20-5, Darmstadt, September 2020.

Working Paper in Wolfsburg Working Papers (WWP), Ostfalia University of Applied Sciences, No. 20-02, Wolfsburg, October 2020.

Abstract

Purpose – This paper aims to assess the quality of interest rate forecasts for the money markets in Argentina, Brazil, Chile, Mexico and Venezuela for the period between 2001 and 2019. Future interest rate trends are of key significance for many business-related decisions. Thus, reliable interest rate forecasts are essential, for example, for banks that make profits by carrying out maturity transformations.

Design/methodology/approach – The data that we analyze were collected by Consensus Economics through a monthly survey with over 120 renowned economists and were published between 2001 and 2019 in the journal *Latin American Consensus Forecasts*. We use the Diebold-Mariano test, the sign accuracy test, the TOTA coefficient and the unbiasedness test to determine the precision and biasedness of the forecasts.

Findings – Our research reveals that the forecasting work carried out in Brazil, Chile and Mexico is remarkably successful. The quality of forecasts from Argentina and Venezuela, on the other hand, is significantly poorer.

Originality/value – Over 50 studies have already been published with regard to the accuracy of interest rate forecasts, emphasizing the importance of the topic. However, interest rate forecasts for Latin American money markets have hardly been considered thus far. Our paper closes this research gap. Overall, the analyzed database amounts to a total of 209 forecast time series with 28,451 individual interest rate forecasts. This study is thus far more comprehensive than all previous studies.

Paper type – Research paper.

Keywords

Forecast accuracy; Interest rate forecasts; Maturity transformation; Survey forecasts; Topically-orientated trend adjustments.

JEL Classification

E44; E47; F37; G15; G17; G21.

1 Introduction

Future interest rate trends are of key significance for many business-related decisions. This is why banks, investment companies and economic research institutes regularly draw up interest rate forecasts. Whereas the interest rates of bonds with several years of residual maturity are predominantly monitored by portfolio managers, very short-term interest rate trends play a significant role for banks which carry out maturity transformations in the lending business. A bank can provide a loan with a payback period of a year. The necessary procurement of funds can be achieved by the bank receiving 12 consecutive deposits with a maturity of one month. In the process, the bank earns the usual profit margin which results from charging its borrowers higher interest rates than it is prepared to pay for customer deposits. Given a normal yield curve, banks also earn from the fact that short-term deposits are rewarded with lower interest rates than long-term ones. This form of maturity transformation plays a major role in making a profit in the lending business.

However, this approach bears risks. If interest rates for short-term deposits rise considerably, maturity transformation can lead to serious losses under certain circumstances. Such events can endanger the existence of financial institutions as the state rescue of the European-based bank Hypo Real Estate or the federal takeover of US mortgage corporations Fannie Mae and Freddie Mac in the aftermath of the credit default crisis of 2007 have emphasized. Banks which carry out maturity transformations are thus dependent on generating interest rate forecasts for the short end of the yield curve which are generally reliable. Unexpected changes in the interest rates can mean that immediate action must be taken to ensure liquidity. Conclusively, it is important that scholars evaluate the reliability of interest rate forecasts on a continuous basis.

In our study, we focus on a data basis which has not yet been analyzed. We examine interest rate forecasts for the money markets in Argentina, Brazil, Chile, Mexico, and Venezuela which were published in the period 2001–2019 in the monthly journal *Latin American Consensus Forecasts*. Each month, Consensus Economics surveys more than 700 leading economists and business scientists for their forecasts on various economic indicators for over 85 countries. While the extensive databases for the regions North America (cf. Gubaydullina et al., 2011; Mose, 2005), Europe (cf. Kunze & Gruppe, 2014; Chortareas et al., 2012) and Asia–Pacific (cf. Filiz et al., 2019; Jongen et al., 2011) were already part of prior research, this study is the first to examine the *Latin American Consensus Forecasts* database.

We use an established and comprehensive set of instruments to assess the quality of the forecasts: We make a comparison between experts' forecasts and naïve forecasts, in the course of which we apply the Diebold-Mariano test. In addition, we use the sign accuracy test to examine the direction of the forecasts and the unbiasedness test to assess the rationality of the forecasts. Furthermore, we add to the established instruments by applying the TOTA coefficient because it provides additional information on how close the forecasts are to the level prevailing when the forecasts were issued. We believe that

this more sophisticated approach is necessary in order to get a deeper understanding of the reliability of the interest rate forecasts.

In doing so, we differentiate between the forecasting results of the individual institutions which participated in the surveys, which were carried out on a monthly basis. In this way, we are thus not limiting ourselves to the analysis of consensus forecasts. The aim of our study is to find out whether or not the individual institutions' forecasts from the *Latin American Consensus Forecasts* database can be used for a business model based on maturity transformation.

In the next chapter, we provide an overview of the existing literature. In the following two chapters, we explain the data basis and the methodology used. In the penultimate chapter, the results are presented, and the final chapter consists of a brief summary.

2 Literature Review

Pesando (1979) is showing that in efficient markets with invariant term premiums, interest rate forecasts with a long horizon will exhibit the characteristics of a random walk. It is thus not to be expected that long-term interest rate forecasts differ significantly from naïve forecast in terms of their quality. However, when it comes to short-term interest rate forecasts, results are mixed. For example, Dua (1988) concludes that the "absolute" forecast accuracy, measured by the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), increases with a decrease in the forecast horizon, whereas the "relative" accuracy, measured by the Theil coefficient, is unaffected. The findings are in line with those of Gosnell and Kolb (1997), who show that the accuracy of short-term interest rate forecasts made by banks can be superior to the accuracy of the respective naïve forecast. Miah et al. (2016) state that short-term interest rate forecasts are more accurate in emerging economies than in developed ones. They attribute their finding to higher inflation rates in emerging markets.

The synoptic literature survey in Filiz et al. (2019) reveals that the success of interest rate forecasts has been examined by at least 50 studies in the past four decades, with advanced economies clearly being in the focus of scholars. Many of these studies focused on forecasts of US interest rate trends, but the reliability of European and Asian interest rate forecasts is also comprehensively reviewed in these studies. Interest rate forecasts for Latin American money markets, however, have hardly been considered until now. Three studies examine Brazilian interest rate forecasts: Tabak and Feitosa (2008) analyze Brazilian interest rate forecasts for the period 1982–2002 and place particular emphasis on the Diebold-Mariano test. Baghestani and Marchon (2012), on the other hand, assess Brazilian forecasts from the period 2003–2008 and focus on the unbiasedness test. Knüppel and Schultefrankenfeld (2013) examine Brazilian interest rate forecasts in the period 1999–2011 and use Theil's U among other tools. All three studies make a largely positive assessment of the forecasts they examine.

Until now, there has only been one study dealing with interest rate forecasts in several Latin American countries: In their groundbreaking study, Miah et al. (2016) take a look at interest rate forecasts involving Argentina, Brazil, Colombia, Mexico and Venezuela. They analyze forecasts based on surveys which were published on the website Fx4casts.com for the period 2001–2012 and deploy the efficiency test and the unbiasedness test. They conclude that the interest rate forecasts in question can generally be viewed as efficient but biased.

Nonetheless, in our view, there is still a lot more to explore in order to enable an even more comprehensive verdict on interest rate forecasts in Latin America. First of all, the mentioned study dates back to 2016 and covers a smaller timespan. As stated, to enable meaningful conclusions for practitioners, it is important to assess the accuracy of interest rate forecasts on a continuous basis. In addition, our study is the first one to examine interest rate forecasts in Chile, a country which has not been in the scope of researchers thus far.

Moreover, previous studies have focused on consensus forecasts, which are obtained by aggregating various forecasts into one. By considering forecasts on the level of individual forecasting institutions, it can be detected if particular institutions are more successful than others that are looking at the same market. Finally, we believe that applying a new methodology that covers the TOTA coefficient helps us create valuable insights regarding the reliability of the forecasts, as we will demonstrate in chapter 4.

3 Data basis

The interest rate forecasts for the countries Argentina, Brazil, Chile, Mexico and Venezuela, which are considered here, originate from the journal *Latin American Consensus Forecasts*. Since 1994, this journal has – initially every two months – published forecasts on various economically relevant benchmarks such as gross domestic product, private consumption, capital investment, industrial production or inflation. Since April 2001, the journal has been published on a monthly basis and also deals with the field of interest rate forecasts.

Latin American Consensus Forecasts differentiates between two forecast horizons. In the journal, the forecasts are sometimes described as three-month and twelve-month forecasts. In reality, however, the forecast horizons are four and thirteen months. This can be seen in the following example: in the edition of January 2018, which was available around mid-January, the forecasts for the end of April 2018 and the end of January 2019 are published. The forecasts themselves are handed in by the participating institutions at the beginning of January. The actual period of time from the beginning of January 2018 to the end of April 2018, however, is four months, while the period of time from the beginning of January 2018 to the end of January 2019 is 13 months.

We examine the forecasts which were published there in the period from April 2001 to December 2019. We evaluate a total of 209 time series with 28,451 interest rate forecasts. There is a detailed overview in Table 1. We analyze all the forecast time series which contain at least 59 items of data. We do not take time series with less than 59 observations into consideration. Under certain circumstances, time series which are too short or contain large gaps can lead to inconclusive results in the procedures used here to measure the quality of forecasts. However, there are three exceptions: three times series with a forecast horizon of 13 months exhibit less than 59 observations, but we included these three-time series nevertheless because the respective forecasters are represented in the forecast horizons of four months with time series which contain more than 59 observations. In order to round off the results, it seemed meaningful to make these three exceptions. Documentation on merging of time series from related institutions can be found in the supplementary online file.

Table 1: Data basis from the journal *Latin American Consensus Forecasts* as used in the study

Country, subject of the forecast	Forecast horizon	Number of time series analyzed	Number of forecasts analyzed	Results in the table
Argentina, 30 days deposit rate	4 M	21	2,870	2
	13 M	21	2,485	2
Brazil, financing overnight rate (SELIC)	4 M	23	3,363	3
	13 M	23	3,310	3
Chile, monetary policy rate	4 M	22	3,093	4
	13 M	22	3,061	4
Mexico, 28 days closing rate (CETES)	4 M	24	3,445	5
	13 M	24	3,403	5
Venezuela, 30 days deposit rate	4 M	15	1,816	6
	13 M	14	1,605	6
∑		209	28,451	

4 M = 4 months, 13 M = 13 months

4 Methods

In order to evaluate the interest rate forecasts, we use the Diebold-Mariano test, the sign accuracy test, the TOTA coefficient and the unbiasedness test (cf. Filiz et al., 2019). A comparison with the naïve forecast (i.e., everything remains as it is) is still the most significant benchmark for the analysis of capital market forecasts. Given that naïve forecasts are always available as a cost-free alternative, one should expect experts' forecasts to be clearly better.

Simple measurements of forecast quality (such as Mean Absolute Error or Mean Squared Error) enable us to make a comparison with a naïve forecast. However, these simple approaches do not permit an assessment of statistical significance. This deficit is avoided by using the Diebold-Mariano test (Diebold & Mariano, 1995), which has become the state of the art (cf. Kunze et al., 2017; Baghestani & Danila, 2014; Beechey & Österholm, 2014). To do so, we calculate the Mean Squared Error (MSE) for the time series of the expert prognoses and for the time series of the naïve forecasts. The test statistics of the Diebold-Mariano test are defined as follows:

$$DM = \frac{\frac{1}{T}\sum\left(V(P_{t1}) - V(P_{t2})\right)}{\sqrt{\hat{y}\ d/T}}$$

T	= number of observations
V	= loss function
P_1	= naïve forecast
P_2	= expert forecast
$\sqrt{\hat{y}\ d/T}$	= joint spread of the two loss functions

The null hypothesis tested in this way is that the naïve forecast (P_1) and the expert forecast (P_2) have the same accuracy. Neither one of the two alternatives thus provides a clearly better result. The numerator is the mean deviation between the loss functions V of the two forecast approaches to be compared. Normally a squared loss function is assumed; in other words, the squared errors of the two forecast approaches are compared (P_1 and P_2). The denominator is the joint spread of the two loss functions. This is estimated on the basis of the long-term autocovariances of the loss functions. In the case of large samples, this test value is asymptotically normally distributed.

The Diebold-Mariano test is usually carried out with standard kernel density estimates. However, in exceptional cases, this can lead to individual intrinsic values being smaller than or equal to zero. As a result, the entire matrix is no longer positive definite, which, however, is a necessary precondition for carrying out the Diebold-Mariano test. In these cases, the Bartlett kernel proposed by Newey and West (1987) is used.

The sign accuracy test (Merton, 1981; Henriksson & Merton, 1981) is another widespread tool for evaluating forecasts. In this procedure, the extent of a forecasted change is not the issue. It only examines whether the general direction of the forecasts (rising or falling) is correct. The forecasts are then entered into a 2 × 2 matrix. On the one hand, a differentiation is made between whether an interest rate increase or an interest rate fall was forecast; on the other hand, a differentiation is also made between whether an interest rate rise or an interest rate fall has actually occurred. The principal diagonal in the 2 × 2 matrix indicates the forecasts which are correct regarding the trend direction. The secondary diagonal indicates the forecasts which are incorrect regarding the trend direction. A chi-squared test is now applied to examine whether the distribution frequency of the four fields is significantly different from a random walk forecast (cf. Diebold & Lopez, 1996; Joutz & Stekler, 2000). If this is the case, a comparison between the number of observations in the principal diagonals and the secondary diagonals must be carried out to establish whether the forecasts are significantly better or significantly worse than a random walk forecast.

In order to answer the question of whether forecasters have oriented themselves toward current levels when drawing up interest rate forecasts, the topically orientated trend adjustment (TOTA) coefficient is used as a statistical benchmark (Andres & Spiwoks, 1999). TOTA is present when forecasts reflect the present more strongly than the future. In the most unfavorable case, the future-oriented character of such forecasts may be lost entirely. TOTA can be observed almost without exception in capital market forecasts of all kinds worldwide (cf. Spiwoks et al., 2015) and leads to the verdict that forecasts are biased (cf. Spiwoks et al., 2010).

The TOTA coefficient is the quotient of two coefficients of determination (R_A^2 and R_B^2). The R_A^2 measures the correlation between the forecasts at the time of their validity and the actual events. The R_B^2 measures the correlation between the forecasts at the time of their appearance and the actual events. The TOTA coefficient takes the following form:

$$TOTA\ coefficient = \frac{R^2{}_{\text{forecasts (validity date); actual events}}}{R^2{}_{\text{forecasts (issue date); actual events}}} = \frac{R^2{}_A}{R^2{}_B}$$

If the TOTA coefficient has a value of < 1, TOTA is given, and forecasts reflect the present more strongly than the future.

Finally, we use the unbiasedness test to check whether the forecast errors are systematic. According to the theory of rational expectations, this should not be the case. Even though the unbiasedness test is one of the most traditional tools for assessing forecast quality (cf. Friedman, 1980), it is still widely used today (Fassas et al., 2021; Filiz et al., 2019; Miah et al., 2016; Baghestani et al., 2015). In the context of the unbiasedness test, many different procedures can be applied depending on the research question and

the structure of the database. For example, Davies and Lahiri (1995) recommend pooling forecasts in order to get a clearer picture of rationality along the entire forecast path. Following their recommendation, Fassas et al. (2021) examine the unbiasedness of forecasts with a panel regression, taking into consideration 18 different forecasts with horizons from 1 to 18 months for each actual datapoint. While this approach can produce insightful results, it is most efficient when the database contains various different forecasts horizons. As the forecasts in our database are only made for two different horizons, we apply the Mincer-Zarnowitz regression (Mincer & Zarnowitz, 1969). The Mincer-Zarnowitz regression takes the following form:

$$A_t = \alpha + \beta P_t + u_t$$

A_t	= event that actually occurred in time t (dependent variable)
α	= constant
β	= coefficient of the respective forecasts
P_t	= forecast of the actual event in time t
u_t	= error term in time t

Based on this equation, forecasts are considered unbiased if α is not significantly different to 0, and β is not significantly different to 1. In addition, the error term u_t may not be autocorrelated.

Forecasts are considered unbiased when, with a low probability of error, the joint hypothesis of $\alpha = 0$ und $\beta = 1$ does not have to be rejected. This is checked by using the Wald test. A further condition is the absence of autocorrelations in the value of the error term u_t, which is examined with the Durbin-Watson test. If, according to these criteria, a forecast time series is based on rational expectations, Granger and Newbold (1973) argue that this by no means signifies that the forecasts are perfect. They merely do not exhibit systematic errors.

The TOTA coefficient and the unbiasedness test are closely related. If a forecast time series is characterized by the phenomenon of TOTA, the forecast error u_t is normally not randomly distributed (cf. Spiwoks et al., 2010). Forecast time series which have a TOTA coefficient of < 1 are therefore normally biased.

In contrast to our methodology, the widely used efficiency test is not a very difficult hurdle for interest rate forecasts because it only examines whether the information contained in the most recent interest rate data before the forecast is made has been given sufficient consideration in the forecasts. If this information content is zero, which is very frequently the case, it is of course not possible to take this into account insufficiently. Forecast time series which pass the efficiency test can thus in no way be considered reliable.

5 Results

In the forecast of the 30 days deposit rate in Argentina (Table 2), there are at least some successes at a forecast horizon of four months. Only three of the 21 forecasts analyzed (14.3%) are significantly more successful than a naïve forecast, but 17 of the 21 forecast time series (81.0%) predict the future trend (rising or falling) notably better than a random walk forecast.

However, the forecasts are somewhat poorer at a forecast horizon of 13 months. Only two out of 21 forecast time series (9.5%) reveal themselves to be significantly more reliable than a naïve forecast. Eleven out of 21 forecast time series (52.4%) predict the future trend (rising or falling) significantly better than a random walk forecast.

The results of the TOTA coefficient at a forecast horizon of four months as well as with a forecast horizon of 13 months are rather sobering. All 42 forecast time series (100%) tend to reflect the present rather than the future. The forecast time series thus lag behind actual interest rate movements by a period which is roughly equivalent to the forecast horizon (see Figure 1). It is therefore unsurprising that only two of the 42 forecast time series (4.8%) prove to be unbiased (Spiwoks et al., 2010).

The experts were highly successful with their forecasts of the financing overnight rate in Brazil (SELIC) (Table 3). At a forecast horizon of four months, 18 of the 23 forecast time series analyzed (78.3%) are significantly better than the corresponding time series of naïve forecasts. The sign accuracy test shows an even better result. All 23 forecast time series (100%) predict the future interest rate trend (rising or falling) significantly better than a random walk forecast. These results are in line with previous evidence that the accuracy of interest rate forecasts in the Brazilian market is well above average (Knüppel & Schultefrankenfeld, 2013; Baghestani & Marchon, 2012; Tabak & Feitosa, 2008).

The various preceding studies on interest rate forecasts from around the world show that the longer the forecast horizon is, the greater the challenge for forecasters (Filiz et al., 2019). It is thus not surprising that the results are somewhat less impressive at a forecast horizon of 13 months. Nevertheless, seven of the 23 forecast time series (30.4%) are significantly more successful than the corresponding time series of naïve forecasts. Furthermore, 19 of the 23 forecast time series (82.6%) predict the future interest rate trend (rising or falling) significantly better than a random walk forecast.

A very unusual result can also be seen in the TOTA coefficients. Among the forecasts with a horizon of four months, nine of the 23 forecast time series (39.1%) do not exhibit TOTA. This means that these time series do not reflect the present more strongly than the future in their forecasts. This is surprising because capital market forecast time series which do not exhibit TOTA are rare (Spiwoks et al., 2015). However, at a forecast horizon of 13 months, the customary picture is restored. All 23 forecast time series (100%) exhibit TOTA. At a forecast horizon of four months and also at a horizon of 13 months, the unbiasedness test reveals itself to be the customary high hurdle for forecasters. Not one

Table 2: Argentinian 30 days deposit rate

	Forecast horizon 4 months								Forecast horizon 13 months							
		Diebold-Mariano test		Sign accuracy test		TOTA coeff.	Unbiasedness test			Diebold-Mariano test		Sign accuracy test		TOTA coeff.	Unbiasedness test	
Institution	#	Res	p value	Res	p value		F test p val.	DWT p val.	#	Res	p value	Res	p value		F test p val.	DWT p val.
Abeceb.com	65	-	0.098	o	0.055	0.897	0.000	0.000	64	o	0.278	+	0.032	0.150	0.019*	0.000
Análisis de Coyuntura	176	o	0.213	+	0.004	0.903	0.000	0.004	174	o	0.413	+	0.001	0.605	0.000*	0.000
ALPHA	204	o	0.484	+	0.005	0.888	0.472*	0.000	180	o	0.459	+	0.047	0.270	0.195*	0.000
Banco Credicoop	158	+	0.071	+	0.000	0.871	0.000	0.000	158	o	0.585	+	0.007	0.456	0.000	0.000
Banco Galicia	131	+	0.075	+	0.000	0.911	0.000*	0.000	94	o	0.919^	o	0.162	0.717	0.000	0.000
BBVA	192	-	0.006	+	0.046	0.498	0.000	0.000	174	o	0.974	+	0.005	0.038	0.052*	0.000
Datarisk	100	o	0.660	+	0.010	0.835	0.000	0.000	99	+	0.054	+	0.000	0.758	0.000	0.000
Deutsche Bank Research	59	o	0.542	+	0.033	0.631	0.038°	0.092	59	o	0.594	o	0.055	0.001	0.004°	0.175
Eco Go Consultores	127	-	0.000	+	0.005	0.886	0.000	0.021	118	o	0.507	+	0.005	0.749	0.000	0.000
Ecolatina	159	o	0.510	o	0.387	0.885	0.000	0.000	159	o	0.368	o	0.237	0.413	0.000	0.000
Econometrica	147	o	0.784	+	0.000	0.805	0.018*	0.000	124	o	0.390	+	0.001	0.325	0.036*	0.000
Econviews	129	o	0.169	+	0.007	0.900	0.000*	0.000	129	o	0.227	o	0.169	0.317	0.000	0.000
Espert & Asociados	132	-	0.016	o	0.437	0.230	0.000	0.169	54	o	0.179	o	0.554	0.261	0.000°	0.000
FIEL	171	o	0.167	+	0.004	0.696	0.000	0.003	140	o	0.248	o	0.063	0.231	0.000	0.021
IHS Markit	76	o	0.892	+	0.007	0.824	0.000°	0.539	76	+	0.051	+	0.002	0.459	0.000	0.000
M A Broda & Asociados	161	o	0.266	+	0.005	0.489	0.001*	0.000	142	o	0.365	o	0.075	0.157	0.000°	0.000
Macroview S.A.	161	o	0.258	+	0.000	0.785	0.122*	0.888	35	o	0.671	o	0.074	0.078	0.526°	0.152
Orlando Ferreres & Asoc	110	o	0.113	o	0.319	0.872	0.000	0.000	94	o	0.208	o	0.335	0.680	0.000°	0.000
Oxford Economics	69	o	0.110	+	0.041	0.803	0.000	0.000	69	o	0.609	o	0.438	0.254	0.000	0.000
Santander Investment	118	o	0.881	+	0.038	0.714	0.000	0.015	118	o	0.349	+	0.015	0.139	0.000*	0.000
Consensus (Mean)	225	+	0.020	+	0.000	0.612	0.000	0.000	225	o	0.600	+	0.008	0.138	0.000	0.000

= Number of observations; TOTA coeff. = TOTA coefficient; Res. = result; o = no significant result; - = significantly worse than a naïve or random walk forecast; + = significantly better than a naïve or random walk forecast; p val. = p value; DWT = Durbin-Watson test; ° = heteroscedasticity could not be proven, so the p value was determined with simple standard errors; * = p values which have changed due to estimation with robust standard errors; ^ = calculation with the Bartlett kernel.

of the 46 forecast time series (0.0%) can be considered unbiased. This signifies that the forecasts contain systematic errors, not just random ones.

The experts were also highly successful when forecasting the monetary policy rate in Chile (Table 4). At a forecast horizon of four months, just under half of the forecast time series (45.5%) are significantly better than the corresponding time series of naïve forecasts. The sign accuracy test even shows that all 22 forecast time series (100%) predict the interest rate trend (rising or falling) significantly better than a random walk forecast would.

At a forecast horizon of 13 months, the forecasters were still notably successful. Five out of 22 forecast time series (22.7%) reveal themselves to be significantly more reliable than the corresponding time series of naïve forecasts, while 20 out of 22 forecast time series (90.9%) predict the future interest rate trend (rising or falling) significantly better than a random walk forecast.

However, 40 out of the 44 forecast time series on the monetary policy rate in Chile (90.9%) are characterized by TOTA. All 44 forecast time series (100%) prove to be biased.

The successes achieved in the forecasts of the 28 days closing rate (CETES) in Mexico are at a comparable level (Table 5). At a forecast horizon of four months, nine of the 24 forecast time series analyzed (37.5%) predict the future interest rate trend significantly better than the corresponding naïve forecasts. A total of 23 out of 24 forecast time series (95.8%) predict the future interest rate trend (rising or falling) significantly more precisely than a random walk forecast.

When considering the forecast horizon of 13 months, it is revealed that 10 out of 24 forecast time series (41.7%) estimate future interest rate trends significantly more precisely than naïve forecasts. Fifteen out of 24 forecast time series (62.5%) predict the future interest rate trend (rising or falling) significantly more precisely than a random walk forecast.

However, it can be noted that all 48 forecast time series (100%) for the 28 days closing rate (CETES) in Mexico are characterized by TOTA. They thus reflect the present rather than the future. This is also mirrored by the unbiasedness test. Only one of the 48 forecast time series (2.1%) proved to be unbiased.

The forecasters were less successful in their predictions of interest rate trends in Venezuela (Table 6). At a forecast horizon of four months, only two of the 15 forecast time series analyzed (13.3%) are significantly better than a naïve forecast. Nevertheless, nine out of 15 forecast time series (60%) predict the future interest rate trend (rising or falling) significantly more precisely than a random walk forecast.

By contrast, the results are considerably less impressive at a forecast horizon of 13 months. Not one of the 14 forecast time series (0.0%) proved to be significantly superior to a naïve forecast, and only three out of 14 forecast time series (21.4%) predict the future interest rate trend significantly more precisely than a random walk forecast.

Table 3: Brazilian financing overnight rate SELIC

	Forecast horizon 4 months								Forecast horizon 13 months							
		Diebold-Mariano test		Sign accuracy test		TOTA coeff.	Unbiasedness test			Diebold-Mariano test		Sign accuracy test		TOTA coeff.	Unbiasedness test	
Institution	#	Res	p value	Res	p value		F test p val.	DWT p val.	#	Res	p value	Res	p value		F test p val.	DWT p val.
Banco Fator	70	o	0.137	+	0.000	0.935	0.000	0.000	70	o	0.746	o	0.383	0.832	0.000	0.000
Banco Votorantim	194	+	0.006	+	0.000	1.017	0.084*	0.000	194	o	0.298	+	0.000	0.718	0.000	0.000
BofA - Merrill Lynch	87	o	0.783	+	0.000	0.980	0.000*	0.000	84	o	0.902	o	0.157	0.855	0.000	0.000
Barclays	129	+	0.025	+	0.000	1.009	0.001*	0.000	128	+	0.032	+	0.000	0.704	0.000	0.000
BBVA	116	+	0.018	+	0.000	0.956	0.502°	0.000	116	o	0.182	+	0.000	0.565	0.000°	0.000
Capital Economics	72	+	0.030	+	0.000	1.060	0.207°	0.000	72	o	0.228	+	0.001	0.830	0.412°	0.000
Datalynk	206	+	0.000	+	0.000	0.968	0.065*	0.000	206	+	0.079	+	0.000	0.613	0.000	0.000
Deutsche Bank	124	+	0.011	+	0.000	0.974	0.009*	0.000	122	o	0.484	+	0.026	0.512	0.000	0.000
Dresdner Kleinwort	83	o	0.169	+	0.000	0.760	0.000°	0.000	67	o	0.914	o	0.301	0.490	0.000°	0.000
Eaton	215	o	0.149	+	0.000	0.917	0.259°	0.000	215	+	0.028	+	0.000	0.624	0.084°	0.000
HSBC (Lloyds TSB Brazil)	141	+	0.055	+	0.000	0.956	0.259*	0.000	139	o	0.291	+	0.000	0.551	0.000°	0.000
IDEAglobal	69	o	0.124	+	0.000	0.899	0.000*	0.000	69	o	0.528	+	0.050	0.369	0.000	0.000
IHS Markit	133	+	0.000	+	0.000	1.021	0.101°	0.000	133	+	0.099	+	0.020	0.548	0.000°	0.000
Itau Unibanco	128	+	0.009	+	0.000	0.970	0.000	0.000	123	o	0.512	o	0.754	0.643	0.000	0.000
LCA Consultores	172	+	0.000	+	0.000	1.050	0.046*	0.000	172	+	0.022	+	0.000	0.682	0.000	0.000
M B Associados	149	+	0.069	+	0.000	1.017	0.021*	0.000	148	o	0.166	+	0.002	0.734	0.000	0.000
MCM Consultores	208	+	0.015	+	0.000	1.013	0.185*	0.000	206	o	0.212	+	0.000	0.710	0.000	0.000
Morgan Stanley	175	+	0.000	+	0.000	0.951	0.016°	0.000	175	o	0.460	+	0.035	0.647	0.000	0.000
Rosenberg Consultoria	199	+	0.071	+	0.000	1.001	0.067*	0.000	199	o	0.386	+	0.000	0.795	0.000	0.000
Santander Brazil	75	+	0.077	+	0.000	0.936	0.000	0.000	74	o	0.422	+	0.042	0.431	0.000	0.000
SILCON/C.R. Contador & Ass.	217	+	0.000	+	0.000	0.959	0.000	0.000	217	o	0.158	+	0.000	0.601	0.000	0.000
Tendências Consultoria Inte.	176	+	0.007	+	0.000	1.016	0.003*	0.000	156	+	0.061	+	0.000	0.636	0.000	0.000
Consensus (Mean)	225	+	0.001	+	0.000	0.979	0.009	0.000	225	+	0.041	+	0.000	0.665	0.000	0.000

= number of observations; TOTA coeff. = TOTA coefficient; Res. = result; o = no significant result; - = significantly worse than a naïve or random walk forecast; + = significantly better than a naïve or random walk forecast; p val. = p value; DWT = Durbin-Watson test; ° = heteroscedasticity could not be proven, so the p value was determined with simple standard errors; * = p values which have changed due to estimation with robust standard errors.

Table 4: Chilean monetary policy rate

	Forecast horizon 4 months								Forecast horizon 13 months							
		Diebold-Mariano test		Sign accuracy test		TOTA coeff.	Unbiasedness test			Diebold-Mariano test		Sign accuracy test		TOTA coeff.	Unbiasedness test	
Institution	#	Res	p value	Res	p value		F test p val.	DWT p val.	#	Res	p value	Res	p value		F test p val.	DWT p val.
Banchile Inversiones	99	o	0.125	+	0.000	0.964	0.495°	0.000	97	+	0.067	+	0.000	0.074	0.116°	0.000
Banco BICE	204	o	0.104	+	0.000	0.842	0.020°	0.000	202	o	0.126	+	0.000	0.163	0.000°	0.000
Banco de Chile	145	o	0.103	+	0.000	0.882	0.347°	0.000	140	o	0.221	+	0.000	0.260	0.000°	0.000
Banco Security	175	o	0.158	+	0.000	1.133	0.002°	0.000	175	o	0.101	+	0.000	0.534	0.000°	0.000
BTG Pactual (Celfin Capital)	146	+	0.053	+	0.000	0.855	0.018°	0.000	146	+	0.061	+	0.000	0.227	0.000	0.000
Cámara de Comercio de San.	163	+	0.057	+	0.000	0.836	0.151°	0.000	163	+	0.086	+	0.000	0.123	0.000	0.000
Corp Research	149	o	0.146	+	0.000	1.059	0.019°	0.000	149	+	0.088	+	0.000	0.417	0.004°	0.000
Dresdner Kleinwort	63	o	0.349	+	0.024	0.672	0.001*	0.000	47	o	0.728	o	0.086	0.142	0.001°	0.000
Econsult	72	+	0.052	+	0.000	0.883	0.049°	0.000	72	+	0.033	+	0.001	0.071	0.000°	0.000
Fontaine y Paúl Consultores	94	o	0.213	+	0.000	0.817	0.032°	0.000	94	o	0.527	+	0.003	0.036	0.000°	0.000
Gemines	215	+	0.059	+	0.000	0.958	0.000°	0.000	216	o	0.510	+	0.000	0.235	0.000°	0.000
HSBC	73	+	0.025	+	0.000	1.038	0.338°	0.000	67	o	0.213	+	0.000	0.726	0.002°	0.000
IHS Markit	104	+	0.094	+	0.000	0.979	0.942°	0.000	104	o	0.356	+	0.000	0.595	0.000	0.000
Larrain Vial	189	o	0.146	+	0.000	1.092	0.001°	0.000	189	o	0.152	+	0.000	0.358	0.000°	0.000
Libertad y Desarrollo	198	+	0.091	+	0.000	0.871	0.048°	0.000	198	o	0.114	+	0.000	0.189	0.000°	0.000
Pontifica Universidad Catolica	151	o	0.228	+	0.000	0.825	0.000°	0.000	151	o	0.435	+	0.000	0.136	0.000°	0.000
Santander Chile	168	o	0.157	+	0.000	0.892	0.081°	0.000	168	o	0.173	+	0.000	0.308	0.002°	0.000
Scotiabank (BBVA)	163	o	0.104	+	0.000	0.940	0.000°	0.000	164	o	0.385	+	0.001	0.553	0.000	0.000
UBS	65	+	0.018	+	0.001	0.818	0.000°	0.000	62	o	0.593	o	0.124	0.414	0.000°	0.000
Universidad Andrés Bello	63	+	0.003	+	0.000	0.857	0.199°	0.000	63	o	0.801	+	0.000	0.367	0.000°	0.000
Universidad de Chile	169	o	0.111	+	0.000	0.805	0.000°	0.000	169	o	0.448	+	0.000	0.119	0.000°	0.000
Consensus (Mean)	225	+	0.097	+	0.000	0.902	0.016°	0.000	225	o	0.167	+	0.000	0.238	0.000°	0.000

= Number of observations; TOTA coeff. = TOTA coefficient; Res = result; o = no significant result; - = significantly worse than a naïve or random walk forecast; + = significantly better than a naïve or random walk forecast; p val. = p value; DWT = Durbin-Watson test; ° = heteroscedasticity could not be proven, so the p value was determined with simple standard errors; * = p values which have changed due to estimation with robust standard errors.

Table 5: Mexican 28 days closing rate CETES

	Forecast horizon 4 months								Forecast horizon 13 months							
		Diebold-Mariano test		Sign accuracy test		TOTA coeff.	Unbiasedness test			Diebold-Mariano test		Sign accuracy test		TOTA coeff.	Unbiasedness test	
Institution	#	Res	p value	Res	p value		F test p val.	DWT p val.	#	Res	p value	Res	p value		F test p val.	DWT p val.
American Chamber Mex	208	+	0.007	+	0.000	0.836	0.000	0.000	208	o	0.367	+	0.006	0.521	0.000	0.000
Banamex	141	o	0.145	+	0.000	0.783	0.002	0.000	141	o	0.394	o	0.906	0.466	0.033*	0.000
BBVA	129	o	0.526	+	0.010	0.798	0.071*	0.022	129	+	0.045	+	0.000	0.338	0.200*	0.000
Bulltick	77	o	0.338	+	0.034	0.966	0.659*	0.000	78	o	0.661	+	0.010	0.368	0.000*	0.000
CAIE-ITAM	225	+	0.004	+	0.000	0.840	0.000*	0.055	225	+	0.071	+	0.006	0.461	0.000	0.000
CEESP	194	o	0.701	+	0.000	0.778	0.000	0.000	194	+	0.012	+	0.006	0.404	0.000	0.000
Consultores Econ	220	o	0.432	+	0.000	0.844	0.000	0.000	220	o	0.869	+	0.000	0.590	0.000	0.000
Deutsche Bank Rsrch	97	+	0.000	+	0.004	0.672	0.400*	0.212	97	+	0.053	o	0.977	0.340	0.056*	0.004
ESANE Consultores	77	o	0.818	o	0.128	0.293	0.020°	0.007	77	o	0.905	o	0.148	0.036	0.031*	0.000
Grupo Bursametrica	217	o	0.707	+	0.000	0.851	0.000	0.000	214	o	0.400	+	0.027	0.542	0.000	0.000
HSBC	157	+	0.016	+	0.000	0.982	0.200°	0.000	151	o	0.505	o	0.374	0.655	0.000*	0.000
IHS Markit	144	+	0.032	+	0.000	0.862	0.809°	0.057	144	+	0.061	+	0.006	0.454	0.000	0.000
ING	122	o	0.289	+	0.000	0.792	0.797°	0.000	122	o	0.883	o	0.795	0.445	0.154°	0.000
Invex Grupo Financiero	94	+	0.091	+	0.000	0.943	0.030°	0.000	81	o	0.109	o	0.306	0.647	0.024°	0.000
Jonathan Heath & Assoc	81	+	0.004	+	0.000	0.684	0.000	0.007	81	+	0.073	+	0.007	0.163	0.000	0.000
JP Morgan Chase Mex	107	o	0.363	+	0.000	0.926	0.253*	0.000	107	+	0.010	+	0.000	0.695	0.001*	0.000
Morgan Stanley	187	o	0.930	+	0.000	0.847	0.030*	0.000	187	+	0.023	+	0.000	0.652	0.272*	0.000
Oxford Economics	71	o	0.312	+	0.035	0.979	0.000°	0.000	72	o	0.293	o	0.636	0.730	0.000	0.000
Santander Mexico	156	o	0.111	+	0.000	0.879	0.000	0.000	155	o	0.179	o	0.986	0.615	0.000	0.000
Scotiabank	169	o	0.332	+	0.000	0.873	0.000	0.000	169	o	0.603	+	0.000	0.610	0.000	0.000
UBS	76	o	0.692	+	0.044	0.349	0.046*	0.043	74	o	0.620	o	0.441	0.135	0.037*	0.000
Ve Por Mas (Kleinwort)	90	o	0.224	+	0.037	0.688	0.000°	0.000	72	+	0.044	+	0.018	0.333	0.000*	0.001
Vector Casa de Bolsa	181	+	0.046	+	0.000	0.875	0.013*	0.021	180	o	0.106	+	0.015	0.544	0.002*	0.000
Consensus (Mean)	225	+	0.032	+	0.000	0.845	0.000	0.000	225	+	0.026	+	0.025	0.499	0.000	0.000

= number of observations; TOTA coeff. = TOTA coefficient; Res = result; o = no significant result; - = significantly worse than a naïve or random walk forecast; + = significantly better than a naïve or random walk forecast; p val. = p value; DWT = Durbin-Watson test; ° = heteroscedasticity could not be proven, so the p value was determined with simple standard errors; * = p values which have changed due to estimation with robust standard errors.

Table 6: Venezuelan 30 days deposit rate

	Forecast horizon 4 months								Forecast horizon 13 months							
		Diebold-Mariano test		Sign accuracy test		TOTA coeff.	Unbiasedness test			Diebold-Mariano test		Sign accuracy test		TOTA coeff.	Unbiasedness test	
Institution	#	Res	P value	Res	P value		F test P val.	DWT P val.	#	Res	P value	Res	P value		F test P val.	DWT P val.
Azpurua (AGPV)	192	o	0.135	o	0.189	0.489	0.001*	0.167	189	o	0.301	o	0.713	0.005	0.026*	0.000
Banco Mercantil	150	o	0.839	+	0.008	0.780	0.000°	0.000	103	o	0.321	-	0.047	0.097	0.000	0.000
Banesco	144	o	0.906^	+	0.000	0.577	0.784*	0.000	142	o	0.883^	+	0.001	0.002	0.014*	0.000
BBVA	88	o	0.311	+	0.002	0.554	0.069*	0.000	87	o	0.541	o	0.299	0.008	0.004*	0.000
Coyuntura - Maxim Ross As.	213	+	0.087	+	0.024	0.702	0.000	0.089	154	o	0.131	o	0.572	0.105	0.000	0.000
Datanalisis	115	-	0.089	o	0.905	0.588	0.000	0.021	107	o	0.158	o	0.086	0.043	0.000	0.000
Deutsche Bank Research	60	o	0.575	o	0.691	0.619	0.833*	0.001	60	o	0.517	o	0.151	0.571	0.000°	0.007
Ecoanalitica	117	o	0.655	+	0.003	0.867	0.000	0.017	117	o	0.396	+	0.007	0.099	0.000	0.000
Universidad Católica (UCAB)	90	o	0.758^	+	0.032	0.545	0.215*	0.000	90	o	0.397	o	0.227	0.015	0.710*	0.000
MPG Consultores	60	o	0.327	o	0.512	1,046	0.000*	0.767	1	NA	NA	NA	NA	NA	NA	NA
Multiplicas	87	o	0.198	o	0.699	0.889	0.000	0.002	59	o	0.657	o	0.851	0.090	0.000°	0.000
Oxford Economics	69	o	0.154	o	0.484	0.694	0.000*	0.000	69	o	0.303	o	0.559	0.137	0.017*	0.000
Santander Venezuela	65	+	0.051	+	0.003	0.577	0.372*	0.007	62	o	0.192	o	0.432	0.072	0.117*	0.000
VenEconomia	141	-	0.000	+	0.002	0.494	0.000	0.000	141	-	0.092	o	0.068	0.004	0.000	0.000
Consensus (Mean)	225	o	0.250	+	0.009	0.540	0.000	0.805	225	o	0.840	+	0.005	0.012	0.000	0.000

= Number of observations; TOTA coeff. = TOTA coefficient; Res = result; NA = not available; o = no significant result; - = significantly worse than a naïve or random walk forecast; + = significantly better than a naïve or random walk forecast; P val. = P value; DWT = Durbin-Watson test; ° = heteroscedasticity could not be proven, so the P value was determined with simple standard errors; * = P values which have changed due to estimation with robust standard errors; ^ = calculated with the Bartlett kernel.

Only one out of 29 forecast time series for the 30 days deposit rate in Venezuela (3.4%) exhibits no TOTA. It is only this one forecast time series which reflects the future direction of interest rates more strongly than the present trend. All 29 forecast time series (100%) turn out to be biased. This means that the forecasting errors are of a systematic nature and cannot be viewed as purely coincidental.

Overall, it can be stated that relatively frequently the efforts made to correctly forecast interest rates in Latin America in the period 2001–2019 were successful (Table 7). Just under a third of all forecast time series (31.7%) lead to significantly better forecasts than if a naïve forecast had been used, while slightly more than three-quarters of forecast time series (77.6%) predict the future direction of interest rates (rising or falling) significantly more precisely than a random walk forecast. This is in line with previous evidence that making reasonable interest rate forecasts tends to be easier in emerging markets.

Table 7: Success rates of interest rate forecasts

Country, subject of the forecast	Forecast horizon	Success rate Diebold-Mariano test in %	Success rate sign accuracy test in %	Success rate TOTA coefficient in %	Success rate test for unbiased-ness in %
Argentina, 30 days deposit rate	4 M	14.3%	81.0%	0.0%	4.8%
	13 M	9.5%	52.4%	0.0%	4.8%
Brazil, financing overnight rate (SELIC)	4 M	78.3%	100.0%	39.1%	0.0%
	13 M	30.4%	82.6%	0.0%	0.0%
Chile, monetary policy rate	4 M	45.5%	100.0%	18.2%	0.0%
	13 M	22.7%	90.9%	0.0%	0.0%
Mexico, 28 days closing rate (CETES)	4 M	37.5%	95.8%	0.0%	4.2%
	13 M	41.7%	62.5%	0.0%	0.0%
Venezuela, 30 days deposit rate	4 M	15.4%	60.0%	6.7%	0.0%
	13 M	0.0%	23.1%	0.0%	0.0%
Ø weighted		31.7%	77.6%	6.7%	1.9%

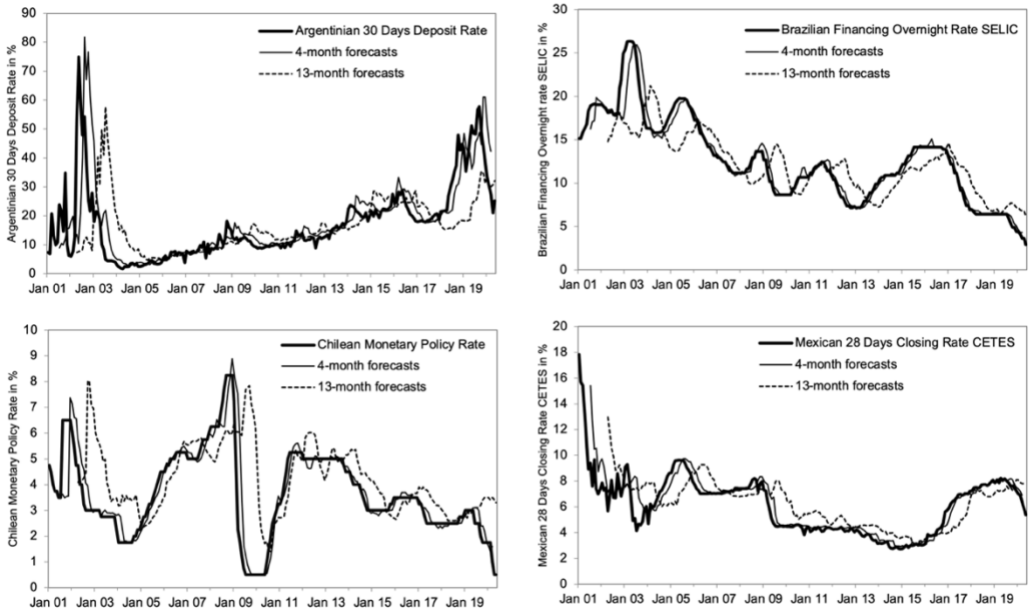
4 M = 4 months, 13 M = 13 months

These successes coincide with previous findings on the reliability of interest rate forecasts (see Filiz et al., 2019): earlier studies on various interest rates throughout the world also show that forecasts tend to be more successful when they are made for short and very short maturities. By contrast, forecasts for interest rates at the long end of the yield curve (such as returns on ten-year government bonds) are for the most part far less successful.

Furthermore, in numerous preceding studies, it can be noted that interest rate forecasters normally allow themselves to be strongly influenced by the current interest rate trend. If the current level of interest rates falls (rises), forecasts are frequently also revised downwards (upwards). This phenomenon, known as TOTA, also characterizes

the vast majority (93.3%) of the forecast time series from Latin America which we analyzed (see Figure 1).

Figure 1: Argentinian, Brazilian, Chilean, and Mexican interest rates



If the forecast horizon (four or 13 months here) is longer than the frequency of the forecasts (monthly in this case), TOTAs frequently lead to the forecasting errors (residuals) not being distributed randomly. Forecast time series of this kind thus also frequently fail the unbiasedness test (cf. Spiwoks et al., 2010). This is the situation in Latin America too. The majority of forecast time series (98.1%) prove to be biased.

Miah et al. (2016) had found that the consensus forecasts with a horizon of 12 months were biased for Argentina, Brazil, Mexico, and Venezuela, whereas the consensus forecasts with a horizon of three months were only found to be biased for Argentina and Venezuela. Likewise, in our study, all consensus forecasts with a horizon of 13 months as well as forecasts for Argentina and Venezuela with a horizon of four months turn out to be biased (Tables 2–6). In contrast to the mentioned study, our results imply that the consensus forecasts with a horizon of four months for Brazil and Mexico are biased, too, though the evidence is not as significant in Brazil as it is in the other countries we examined. Moreover, we reveal that the consensus forecast with a horizon of four months for Chile is the only one for which the null hypothesis of unbiasedness cannot be rejected at a 99% significance level. Chile has not been in the scope of previous studies.

In addition to previous studies, we show that on the individual institution level, the accuracy of the forecasts varies significantly. In Brazil, Chile, and Mexico, between 23 and 42% of the institutions manage to pass the Diebold-Mariano test even with their forecasts for a horizon of 13 months, which can be considered a remarkable success. What is even more exceptional is that two forecasters in Argentina and Mexico pass the unbiasedness test. This implies that the reliability of forecasts differs considerably between the institutions, and that one should carefully assess whose forecasts to follow.

All in all, our results suggest different chances of success for financial activity in the countries examined. Commercial institutions that use interest rate forecasts as indicators for economic decisions should take these findings into account in order to increase their likelihood of success on the Latin American market. As we have argued, reliance on insufficient interest rate forecasts can lead to banks not being able to refinance their activities, ultimately putting the existence of their business into risk. Our study reveals that especially on the Argentinian and Venezuelan market, this threat should be considered carefully due to the poor quality of predictions.

Likewise, private investors should take our findings on the accuracy of interest rate forecasts at the individual institution level into account when investing in these institutions. In general, providing accurate interest forecasts indicates that an institution has a good understanding of the market environment in which it operates and may therefore be more likely to be successful in the business of maturity transformation.

6 Summary

Since 2001, *Latin American Consensus Forecasts* has published monthly forecasts on interest rate trends at the short end of the yield curve in Argentina, Brazil, Chile, Mexico, and Venezuela. We examine the forecast time series from 2001 to 2019 with the aid of the Diebold-Mariano test, the sign accuracy test, the TOTA coefficient and the unbiasedness test. While doing so, we not only consider the time series of the consensus forecasts but all of the time series of forecasting institutions which issued at least 59 forecasts in the period of observation. Overall, we assess 209 forecast time series with a total of 28,451 individual forecasts.

The forecasts for interest rate trends in Brazil, Chile and Mexico in particular can be viewed as highly successful. The interest rate forecasts for Argentina and Venezuela, on the other hand, are much less accurate. This can possibly be traced back to the sovereign debt defaults (2001 and 2014) in Argentina and to increasing levels of political destabilization since 2013 in the case of Venezuela.

Just under a third of all forecast time series (31.7%) lead to significantly better forecasts than if a naïve forecast had been used, while somewhat more than three-quarters of forecast time series (77.6%) predict the future direction of interest rates (rising or falling) significantly more precisely than a random walk forecast.

However, this study also reveals that the majority of forecast time series (93.3%) exhibit TOTA. These forecast time series thus reflect present interest rate trends rather than future ones. In addition, the majority of the forecast time series (98.1%) are biased.

A further aspect is that forecasts with a forecast horizon of four months are usually far more reliable than those with a forecast horizon of 13 months analyzed in this study. This largely corresponds to the findings of numerous previous studies on interest rate forecasts throughout the world.

References

Andres, P., & Spiwoks, M. (1999), Forecast Quality Matrix – A Methodological Survey of Judging Forecast Quality of Capital Market Forecasts, *Journal of Economics and Statistics*, 219(5-6), 513-542.

Baghestani, H., Arzaghi, M., & Kaya, I. (2015), On the accuracy of Blue Chip forecasts of interest rates and country risk premiums, *Applied Economics*, 47(2), 113-122.

Baghestani, H., & Danila, L. (2014), Interest Rate and Exchange Rate Forecasting in the Czech Republic: Do Analysts Know Better than a Random Walk?, *Finance a Uver: Czech Journal of Economics & Finance*, 64(4), 282-295.

Baghestani, H., & Marchon, C. (2012), An evaluation of private forecasts of interest rate targets in Brazil, *Economics Letters*, 115(3), 352-355.

Beechey, M., & Österholm, P. (2014), Policy interest-rate expectations in Sweden: a forecast evaluation, *Applied Economics Letters*, 21(14), 984-991.

Chortareas, G., Jitmaneeroj, B., & Wood, A. (2012), Forecast rationality and monetary policy frameworks: evidence from UK interest rate forecasts, *Journal of International Financial Markets, Institutions & Money*, 22(1), 209-231.

Davies, A., & Lahiri, K. (1995), A new framework for analyzing survey forecasts using three-dimensional panel data, *Journal of Econometrics,* 68(1), 205-227.

Diebold, F. X., & Lopez, J. A. (1996), Forecast Evaluation and Combination, in Maddala, G. S. and Rao, C. R. (Eds.), *Handbook of Statistics*, Amsterdam, 241-268.

Diebold, F. X., & Mariano, R. S. (1995), Comparing Predictive Accuracy, *Journal of Business and Economic Statistics*, 13, 253-263.

Dua, P. (1988), Multiperiod forecasts of interest rates, *Journal of Business & Economic Statistics*, 6(3), 381-384.

Fassas, A., Papadamou, S., & Kenourgios, D. (2021), Evaluating survey-based forecasts of interest rates and macroeconomic variables, *Journal of Economic Studies*, 49(1), 140-158.

Filiz, I., Nahmer, T., Spiwoks, M., & Bizer, K. (2019), The accuracy of interest rate forecasts in the Asia-Pacific region: opportunities for portfolio management, *Applied Economics*, 51(59), 6309-6332.

Friedman, B. M. (1980), Survey evidence on the 'rationality' of interest rate expectations, *Journal of Monetary Economics*, 6(4), 453-465.

Gosnell, T. F., & Kolb, R. W. (1997), Accuracy of international interest rate forecasts, *Financial Review*, 32(3), 431-448.

Granger, C. W. J., & Newbold, P. (1973), Some comments on the evaluation of economic forecasts, *Applied Economics*, 5(1), 35-47.

Gubaydullina, Z., Hein, O., & Spiwoks, M. (2011), The Status Quo Bias of Bond Market Analysts, *Journal of Applied Finance & Banking*, 1(1), 31-51.

Henriksson, R. D., & Merton, R. C. (1981), On Market Timing and Investment Performance, Statistical Procedures for Evaluating Forecasting Skills, *Journal of Business*, 54, 513-533.

Jongen, R., Verschoora, W. F. C., & Wolff, C. C. P. (2011), Time-variation in term premia: International survey-based evidence, *Journal of International Money and Finance*, 30(4), 605-622.

Joutz, F., & Stekler, H. O. (2000), An evaluation of the predictions of the Federal Reserve, *International Journal of Forecasting*, 16(1), 17-38.

Knüppel, M., & Schultefrankenfeld, G. (2013), The empirical (ir)relevance of the interest rate assumption for central bank forecasts, Working Paper No. 11/2013, Deutsche Bundesbank.

Kunze, F., & Gruppe, M. (2014), Performance of Survey Forecasts by Professional Analysts: Did the European Debt Crisis Make it Harder or Perhaps Even Easier?, *Social Sciences*, 3(1), 128-139.

Kunze, F., Wegener, C., Bizer, K., & Spiwoks, M. (2017), Forecasting European interest rates in times of financial crisis – What insights do we get from international survey forecasts?, *Journal of International Financial Markets, Institutions and Money*, 48, 192-205.

Merton, R. C. (1981), On Market Timing and Investment Performance, An Equilibrium Theory of Value for Market Forecasts, *Journal of Business*, 54, 363-406.

Miah, F., Khalifa, A. A., & Hammoudeh, S. (2016), Further evidence on the rationality of interest rate expectations: A comprehensive study of developed and emerging economies, *Economic Modelling*, 54, 574-590.

Mincer, J., & Zarnowitz, V. (1969), The Evaluation of Economic Forecasts, in: Mincer, J. (Ed.), Economic Forecasts and Expectation, Columbia Press, New York, 3-46.

Mose, J. S. (2005), Expert Forecasts of Bond Yields and Exchange Rates, *Danmarks Nationalbank Monetary Review*, Copenhagen, 91-95.

Newey, W. K., & West, K. D. (1987), A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica*, 55(3), 703-708.

Pesando, J. E. (1979), On the random walk characteristics of short-and long-term interest rates in an efficient market, *Journal of Money, Credit and Banking*, 11(4), 457-466.

Spiwoks, M., Bedke, N., & Hein, O. (2010), Topically Orientated Trend Adjustment and Autocorrelation of the Residuals - An Empirical Investigation of the Forecasting Behavior of Bond Market Analysts in Germany, *Journal of Money, Investment and Banking*, 14, 16-35.

Spiwoks, M., Gubaydullina, Z., & Hein, O. (2015), Trapped in the Here and Now – New Insights into Financial Market Analyst Behavior, *Journal of Applied Finance & Banking*, 5(1), 35-56.

Tabak, B. M., & Feitosa, M. A. (2008), How Informative are Interest Rate Survey-based Forecasts?, *Brazilian Administrative Review*, 5(4), 304-318.

Appendix A: Supplements to the unbiasedness test

Table A-1: Argentinian 30 days deposit rate

Institution	4 months forecast horizon			13 months forecast horizon		
	p value Breusch-Pagan test for heteroscedasticity	p value F test Robust estimation	p value F test Normal estimation	p value Breusch-Pagan test for heteroscedasticity	p value F test Robust estimation	p value F test Normal estimation
Abeceb.com	0.000	0.000	0.000	0.052	0.019	0.005
Análisis de Coyuntura (ACM)	0.000	0.000	0.000	0.000	0.000	0.017
ALPHA	0.000	0.472	0.081	0.000	0.195	0.002
Banco Credicoop	0.000	0.000	0.000	0.000	0.000	0.000
Banco Galicia	0.000	0.000	0.002	0.000	0.000	0.000
BBVA	0.000	0.000	0.000	0.000	0.052	0.004
Datarisk	0.000	0.000	0.000	0.002	0.000	0.000
Deutsche Bank Research	0.116	0.276	0.038	0.179	0.001	0.004
Eco Go Consultores	0.000	0.000	0.000	0.000	0.000	0.000
Ecolatina	0.000	0.000	0.000	0.000	0.000	0.000
Econometrica	0.000	0.018	0.000	0.000	0.036	0.000
Econviews	0.000	0.000	0.003	0.000	0.000	0.000
Espert & Asociados	0.000	0.000	0.000	0.321	0.000	0.000
FIEL	0.000	0.000	0.000	0.000	0.000	0.000
IHS Markit	0.150	0.000	0.000	0.000	0.000	0.000
M A Broda & Asociados	0.000	0.001	0.000	0.876	0.000	0.000
Macroview S.A.	0.000	0.122	0.289	0.275	0.295	0.526
Orlando Ferreres & Asoc	0.000	0.000	0.000	0.147	0.000	0.000
Oxford Economics	0.000	0.000	0.000	0.010	0.000	0.000
Santander Investment	0.000	0.000	0.000	0.000	0.000	0.001
Consensus (mean)	0.000	0.000	0.000	0.000	0.000	0.000

Table A-2: Brazilian financing overnight rate SELIC

Institution	4 months forecast horizon			13 months forecast horizon		
	p value Breusch-Pagan test for heteroscedasticity	p value F test Robust estimation	p value F test Normal estimation	p value Breusch-Pagan test for heteroscedasticity	p value F test Robust estimation	p value F test Normal estimation
Banco Fator	0.014	0.000	0.000	0.003	0.000	0.000
Banco Votorantim	0.000	0.084	0.165	0.002	0.000	0.000
BofA - Merrill Lynch	0.075	0.000	0.010	0.059	0.000	0.000
Barclays	0.002	0.001	0.000	0.000	0.000	0.000
BBVA	0.182	0.465	0.502	0.737	0.000	0.000
Capital Economics	0.967	0.019	0.207	0.882	0.384	0.412
Datalynk	0.091	0.065	0.093	0.004	0.000	0.000
Deutsche Bank	0.069	0.009	0.022	0.001	0.000	0.000
Dresdner Kleinwort	0.204	0.001	0.000	0.744	0.000	0.000
Eaton	0.932	0.198	0.259	0.409	0.090	0.084
HSBC (Lloyds TSB Brazil)	0.004	0.259	0.120	0.931	0.000	0.000
IDEAglobal	0.002	0.000	0.005	0.005	0.000	0.000
IHS Markit	0.108	0.066	0.101	0.164	0.000	0.000
Itau Unibanco	0.001	0.000	0.000	0.000	0.000	0.000
LCA Consultores	0.065	0.046	0.115	0.006	0.000	0.000
M B Associados	0.000	0.021	0.055	0.001	0.000	0.000
MCM Consultores	0.000	0.185	0.238	0.000	0.000	0.000
Morgan Stanley	0.211	0.006	0.016	0.002	0.000	0.000
Rosenberg Consultoria	0.000	0.067	0.142	0.000	0.000	0.000
Santander Brazil	0.017	0.000	0.000	0.023	0.000	0.000
SILCON/C.R. Contador & Ass.	0.039	0.000	0.000	0.006	0.000	0.000
Tendências Consultoria Inte.	0.000	0.003	0.004	0.000	0.000	0.000
Consensus (mean)	0.032	0.009	0.009	0.010	0.000	0.000

Table A-3: Chilean monetary policy rate

Institution	4 months forecast horizon			13 months forecast horizon		
	p value Breusch-Pagan test for heteroscedasticity	p value F test Robust estimation	p value F test Normal estimation	p value Breusch-Pagan test for heteroscedasticity	p value F test Robust estimation	p value F test Normal estimation
Banchile Inversiones	0.858	0.561	0.495	0.390	0.132	0.116
Banco BICE	0.807	0.040	0.020	0.281	0.000	0.000
Banco de Chile	0.571	0.216	0.347	0.512	0.000	0.000
Banco Security	0.371	0.014	0.002	0.639	0.000	0.000
BTG Pactual (Celfin Capital)	0.219	0.082	0.018	0.010	0.000	0.000
Cámara de Comercio de San.	0.670	0.248	0.151	0.092	0.000	0.000
Corp Research	0.364	0.010	0.019	0.732	0.003	0.004
Dresdner Kleinwort	0.020	0.001	0.009	0.356	0.000	0.001
Econsult	0.221	0.043	0.049	0.122	0.000	0.000
Fontaine y Paúl Consultores	0.328	0.068	0.032	0.830	0.000	0.000
Gemines	0.624	0.000	0.000	0.686	0.000	0.000
HSBC	0.412	0.248	0.338	0.944	0.002	0.002
IHS Markit	0.720	0.952	0.942	0.087	0.000	0.000
Larrain Vial	0.892	0.002	0.001	0.797	0.000	0.000
Libertad y Desarrollo	0.582	0.073	0.048	0.463	0.000	0.000
Pontifica Universidad Catolica	0.556	0.001	0.000	0.640	0.000	0.000
Santander Chile	0.759	0.115	0.081	0.186	0.006	0.002
Scotiabank (BBVA)	0.915	0.000	0.000	0.035	0.000	0.000
UBS	0.171	0.000	0.000	0.432	0.000	0.000
Universidad Andrés Bello	0.266	0.081	0.199	0.561	0.000	0.000
Universidad de Chile	0.240	0.000	0.000	0.926	0.000	0.000
Consensus (mean)	0.968	0.022	0.016	0.601	0.000	0.000

Table A-4: Mexican 28 days closing rate CETES

Institution	4 months forecast horizon			13 months forecast horizon		
	p value Breusch-Pagan test for heteroscedasticity	p value F test Robust estimation	p value F test normal estimation	p value Breusch-Pagan test for heteroscedasticity	p value F test Robust estimation	p value F test Normal estimation
American Chamber Mex	0.000	0.000	0.000	0.000	0.000	0.000
Banamex	0.005	0.002	0.002	0.026	0.033	0.072
BBVA	0.014	0.071	0.002	0.003	0.200	0.178
Bulltick	0.006	0.659	0.390	0.001	0.000	0.002
CAIE-ITAM	0.000	0.000	0.009	0.000	0.000	0.000
CEESP	0.028	0.000	0.000	0.000	0.000	0.000
Consultores Econ	0.003	0.000	0.000	0.000	0.000	0.000
Deutsche Bank Rsrch	0.001	0.400	0.438	0.046	0.056	0.074
ESANE Consultores	0.493	0.026	0.020	0.000	0.031	0.000
Grupo Bursametrica	0.000	0.000	0.000	0.000	0.000	0.000
HSBC	0.445	0.238	0.200	0.002	0.000	0.001
IHS Markit	0.219	0.785	0.809	0.050	0.000	0.000
ING	0.246	0.785	0.797	0.279	0.283	0.154
Invex Grupo Financiero	0.782	0.014	0.030	0.450	0.002	0.024
Jonathan Heath & Assoc	0.003	0.000	0.000	0.000	0.000	0.000
JP Morgan Chase Mex	0.000	0.253	0.037	0.016	0.001	0.021
Morgan Stanley	0.001	0.030	0.004	0.000	0.272	0.264
Oxford Economics	0.117	0.000	0.000	0.027	0.000	0.000
Santander Mexico	0.000	0.000	0.000	0.000	0.000	0.000
Scotiabank	0.000	0.000	0.000	0.000	0.000	0.000
UBS	0.026	0.046	0.008	0.000	0.037	0.001
Ve Por Mas (Kleinwort)	0.103	0.000	0.000	0.002	0.000	0.002
Vector Casa de Bolsa	0.000	0.013	0.096	0.000	0.002	0.024
Consensus (mean)	0.001	0.000	0.000	0.000	0.000	0.000

Table A-5: Venezuelan 30 days deposit rate

Institution	4 months forecast horizon			13 months forecast horizon		
	p value Breusch-Pagan test for heteroscedasticity	p value F test Robust estimation	p value F test Normal estimation	p value Breusch-Pagan test for heteroscedasticity	p value F test Robust estimation	p value F test Normal estimation
Azpurua (AGPV)	0.000	0.001	0.000	0.000	0.026	0.000
Banco Mercantil	0.304	0.001	0.000	0.000	0.000	0.000
Banesco	0.000	0.784	0.639	0.001	0.014	0.024
BBVA	0.000	0.069	0.000	0.001	0.004	0.000
Coyuntura - Maxim Ross As.	0.000	0.000	0.000	0.000	0.000	0.000
Datanalisis	0.000	0.000	0.000	0.000	0.000	0.000
Deutsche Bank Research	0.002	0.833	0.557	0.959	0.000	0.000
Ecoanalitica	0.000	0.000	0.000	0.000	0.000	0.000
Universidad Católica (UCAB)	0.002	0.215	0.407	0.000	0.710	0.399
MPG Consultores	0.000	0.000	0.002	NA	NA	NA
Multiplicas	0.028	0.000	0.000	0.242	0.000	0.000
Oxford Economics	0.000	0.000	0.373	0.001	0.017	0.000
Santander Venezuela	0.004	0.372	0.109	0.004	0.117	0.045
VenEconomia	0.000	0.000	0.000	0.000	0.000	0.000
Consensus (mean)	0.000	0.000	0.000	0.000	0.000	0.000

NA = not available.

Chapter VIII

Sticky Stock Market Analysts

Co-authored by Ibrahim Filiz, Marco Lorenz, and Markus Spiwoks
Contribution Jan René Judek: 45%

Published:
Journal of Risk and Financial Management, *Vol. 14*, *Issue 12*, *593*, *1-27*. *(Dec 2021)*
https://doi.org/10.3390/jrfm14120593

Also published as:

Discussion Paper in Sonderforschungsgruppe Institutionenanalyse (sofia), No. 21-3, Darmstadt, February 2021.

Working Paper in Wolfsburg Working Papers (WWP), Ostfalia University of Applied Sciences, No. 21-03, Wolfsburg, April 2021.

Abstract

Technological progress in recent years has made new methods available for making forecasts in a variety of areas. We examine the success of ex-ante stock market forecasts of three major stock market indices, i.e., the German Stock Market Index (DAX), the Dow Jones Industrial Index (DJI), and the Euro Stoxx 50 (SX5E). We test whether the forecasts prove true when they reach their effective dates and are therefore suitable for active investment strategies. We revive the thoughts of the American sociologist William Fielding Ogburn, who argues that forecasters consistently underestimate the variability of the future. In addition, we draw on some contemporary measures of forecast quality (prediction-realization diagram, test of unbiasedness, and Diebold–Mariano test). We reveal that (a) unusual events are underrepresented in the forecasts, (b) the dispersion of the forecasts lags behind that of the actual events, (c) the slope of the regression lines in

the prediction-realization diagram is < 1, (d) the forecasts are highly biased, and (e) the quality of the forecasts is not significantly better than that of naïve forecasts. The overall behavior of the forecasters can be described as "sticky" because their forecasts adhere too strongly to long-term trends in the indices and are thus characterized by conservatism.

Keywords

Stock market forecasting; Forecasting bias; variability of reality; conservatism of predictors.

JEL Classification

D83; D84; D91; G17; G41.

1 Introduction

Capital market forecasts often show a closer connection to the capital market development of the present than to the capital market development of the future. This phenomenon is known as topically orientated trend adjustment (Andres & Spiwoks, 1999). It occurs equally in share price forecasts, interest rate forecasts, exchange rate forecasts, and commodity price forecasts (see, e.g., Filiz et al., 2019; Kunze et al., 2018; Spiwoks et al., 2015; Spiwoks & Hein, 2007). A tendency to underestimate the variability of reality could be an important cause (Spiwoks et al., 2015).

The American sociologist William Fielding Ogburn discovers almost 90 years ago that forecasters systematically underestimate the actual variability of reality (Ogburn, 1934). He provides a concrete research approach to identify such behavior. Presumably because Ogburn deals with the prognosis of sporting events and not with the prognosis of economic events, he has so far not been noticed by economic research.

During an empirical analysis of the forecasting behavior of experts and lay people, Ogburn (1934) concludes that the variability of reality is consistently underestimated. He traces this back to a tendency which he calls the "conservatism of the predictors". In detail, he is referring to:

1. Unusual events (e.g., a sudden drop in an otherwise rising trendline) are forecasted more seldom than they occur in reality, whereas normal events (e.g., a recently rising trendline continuing to rise) are over-represented in forecasts.

2. The standard deviation of the forecasts is lower than the standard deviation of the actual events.

3. The extent of the forecasted changes lags behind the scale of the actual changes.

Active investment strategies have been popular since the emergence of modern stock markets (Maxwell & van Vuuren, 2019; Lofthouse, 1996; Friend & Vickers, 1965; Cowles, 1933). In order to successfully design active investment strategies such as market timing, stock picking, or index picking, forecasts of future stock market developments are indispensable. New forecasting methods are constantly being discussed: econometric models (Goyal et al., 2021; Chen & Vincent, 2016; Welch & Goyal, 2008), artificial neural networks (Rajab & Sharma, 2019; Atsalakis & Valavanis, 2009), artificial intelligence (Mallikarjuna & Rao, 2019), capital market simulations with multi-agent models (Yang et al., 2020; Krichene & El-Aroui, 2018; Arthur et al., 1997), modelling based on the expectations of capital market agents (Atmaz et al., 2021; Greenwood & Shleifer, 2014), and neuro-psycho-economics approaches (Ortiz-Teran et al., 2019; Kandasamy et al., 2016; Werner et al., 2009). However, testing these approaches using ex-post forecasts in an out-of-sample data domain repeatedly leads to apparent forecasting successes that then may not materialize in real ex-ante settings (Kazak & Pohlmeier, 2019). When the variability of reality is systematically underestimated, this can contribute towards very costly errors in the field of stock market forecasts. Under certain circumstances, basing

active investment strategies on inappropriate stock market forecasts can lead to serious losses and even bankruptcy, when expected returns do not occur. Due to the necessity of reliable forecasts for a successful active investment strategy, stock market forecasting is a dynamic field of research.

The reliability of stock market forecasts is rarely examined. There are many studies on pre-tax profit forecasts (Ramnath et al., 2008), but research on the success of actual ex-ante forecasts in stock prices, stock market indices, or stock market returns are still a rarity. So far, it has not been in the focus of research whether stock market forecasts are characterized by a systematic underestimation of the variability of reality as found by Ogburn (1934). This research gap is even more surprising because the necessary investigation tools have long been available in the form of Theil's prediction-realization diagram and the test for unbiasedness. We raise the question of how successful experts were in forecasting major stock indices (DAX, Dow Jones Industrial Index, Euro Stoxx 50) in the period from 1992 to 2020. We use Ogburn's (1934) examination instruments. But we also go beyond this and use current standard procedures such as the comparison to the naïve forecasts (Diebold–Mariano test) and the unbiasedness test.

The forecasts turn out to be quite unreliable. Indeed, forecasters underestimate the variability of reality. This offers interesting starting points for improving the forecasting process.

2 Literature Review

2.1 Technological Progress in Stock Market Forecasting

There is a rich literature on the appliance of advanced econometric methodology in the forecasting process in order to identify meaningful predictors for future events. Guo (2006) uses ordinary and dynamic least squares regressions to analyze whether four different variables can be used as predictors for stock returns. The study concludes that the consumption-wealth ratio can indeed be used for statistically significant forecasts. Chen and Vincent (2016) also use different econometric models applied to full-sample approaches and out-of-sample approaches in order to analyze the informational value of different variables for the development of the Standard and Poor's 500 index (S&P 500) for the period 1964 to 2011. They conclude that the market momentum and the investor sentiment can indeed serve as potential predictors for bear markets. In a similar study, Neely et al. (2014) find that adding technical variables to the commonly used macroeconomic predictors can significantly improve the quality of forecasts for the equity risk premium.

Welch and Goyal (2008) examine the informative value of 13 frequently used variables such as dividend yields or inflation. In contrast to the researchers mentioned above, they find that none of the 13 variables can be used to predict the S&P 500 index returns from 1926 to 2004 neither in-sample nor out-of-sample. Quite importantly, they also find that none of the information available at the time of a potential investment decision would

have helped to gain an idea of future developments. A couple of years later, the same authors extend their research to 29 additional variables that have been brought up in the discussion in the meantime. In spite of the advances in research methods, they still diagnose a poor usefulness in predicting the equity premium in-sample and out-of-sample (Goyal et al., 2021).

Bahrami et al. (2018) add to the research by finding that even though most variables themselves do not lead to significant forecasts, combining forecasts from individual predictive models significantly improves the quality of stock return forecasts for ten advanced emerging markets across the globe.

Whereas most studies cited above apply OLS regression models, Nyberg (2013) examines the suitability of dynamic binary time series models for predicting the S&P 500 index between 1957 and 2010. The author concludes that both in-sample and out-of-sample, dynamic binary time series models are able to successfully forecast bull and bear markets.

A very dynamic research area is capital market simulation with multi-agent models. Heterogeneous agents interact with one another on an artificial stock market. Their demand for shares and their supply of shares are brought together in a stock exchange, so that the development of the share prices results from the actions of the individual agents. These in turn observe the development of the share price and adjust their further behavior to the development of the share price. In this way, the special dynamics of interactions on stock markets can be modeled and examined more closely. The artificial stock markets are validated using the stylized facts (e.g., fat tails, gain-loss asymmetry, volatility clustering, volume-volatility correlation). The price patterns of artificial stock markets should correspond to the price patterns of real stock markets.

The first highlight of this research area is the Santa Fe Artificial Stock Market (Arthur et al., 1997). The Frankfurt Artificial Stock Market (Hein et al., 2012) also takes into account a realistic stock exchange mechanism, different communication structures between the agents, and different investment philosophies of the agents. Recently, for example, information asymmetries (Krichene & El-Aroui, 2018), memory length and confidence level (Bertella et al., 2014), risk preference (Chen & Huang, 2008), tick size systems (Yang et al., 2020), and different types of stocks (Ponta & Cincotti, 2018) have been taken into account in artificial stock markets. Artificial stock markets have the significant advantage that extreme events (crashes) can be observed more frequently and can be better analyzed than on real stock markets. The decisive disadvantage of the artificial stock markets is that the models are still too abstract to lead to very concrete share price forecasts.

Another very dynamic research area uses survey data to examine the expectations of capital market players more closely (e.g., Atmaz et al., 2021; Cassella & Gulen, 2019; Cassella & Gulen, 2018; Greenwood & Shleifer, 2014). In some approaches, different types of investors (lay people vs. professionals or contrarians vs. extrapolators) are taken into account. The different expectations of these investor groups are then used to develop models for describing or forecasting share price developments. These approaches appear

particularly promising because the special importance of the expectations for capital market events is emphasized. In addition, real capital market data are linked with survey data on the expectations of capital market players in a very differentiated manner. In contrast to the approaches of capital market simulation based on multi-agent models, these research approaches remain close to the observable reality of price formation on the stock markets.

In recent years, there have also been promising results regarding neuro-fuzzy systems used for stock price forecasting. For example, Atsalakis and Valavanis (2009) create a neuro-fuzzy system that outperforms a traditional "buy and hold"-strategy regarding the Athens and the New York Stock Exchange. Even in a direct comparison to econometric methods, Rajab and Sharma (2019) show that neuro-fuzzy approaches to forecasting the Bombay Stock Exchange, CNX Nifty, and S&P 500 can significantly outperform multiple regression analysis models or generalized autoregressive conditional heteroscedasticity models.

On the other hand, Mallikarjuna and Rao (2019) find that traditional linear and non-linear models are more accurate at forecasting daily stock market returns of selected indices from developed, emerging, and frontier markets for the period 2000 to 2018 than newly emerged artificial intelligence and frequency domain models. However, neither of the four models nor hybrid approaches provide satisfying results across the markets in their study.

In the field of neuro-psycho-economic approaches, Kandasamy et al. (2016) show that interoception, i.e., the perception of physiological signals from within the body, seems to play a role in the success of professional financial traders. Werner et al. (2009) also show that people with good cardiac perception perform better when choosing between profit and loss options.

In the context of ex-post forecasts in the out-of-sample area, these approaches sometimes show enormous potential. However, many of these approaches have yet to prove their suitability for actual ex-ante forecasts. Their informative value for ex-ante forecasts might be limited due to, for example, differences in estimation risk and low statistical power (Kazak & Pohlmeier, 2019).

2.2 Ex-Ante Stock Market Forecasts

The actual success of stock market forecasts is thus best checked against real ex-ante forecasts. In the area of interest rate forecasts, the evaluation of continuously published forecasts has a long tradition (Filiz et al., 2021; Fassas et al., 2021; Filiz et al., 2019; Kunze et al., 2017; Miah et al., 2016; Pierdzioch, 2015; Baghestani et al., 2015; Oliver & Pasaogullari, 2015; Spiwoks et al., 2015). In the area of stock market forecasting, however, there are only a small number of studies that check continuously published stock market forecasts for their reliability (see the synoptic overview in Table 1).

Lakonishok (1980) analyzes forecasts for the S&P 425 index in the period from 1947 to 1974. He concludes that the reliability of the forecasts does not go recognizably beyond that of naïve forecasts. In this context, a naïve forecast is defined as the assumption that the prevailing value for the variable being forecast at the time the forecast is made will also prevail in the future. In addition, the forecasts are biased and systematically underestimate the returns of the S&P 425. Dimson and Marsh (1984) analyze the forecasted returns of 206 selected British shares in the period from 1980 to 1981. The authors conclude that the forecasts are successful and can lead to systematic excess returns. Fraser and MacDonald (1993) examine forecasts for the development of the French CAC 40 index in the period from 1984 to 1987. This reveals that the forecasts are less reliable than naïve forecasts. Furthermore, it is evident that the forecasts tend to be oriented towards the present rather than the future.

Table 1: Synoptic overview of studies on ex-ante stock market forecasts

Study	Subject of the Forecast	Methods	Time Scale	Result
Lakonishok (1980)	S&P 425	Unbiasedness test with Theil–Sen estimator, Theil's U, turning point errors	1947–1974	–
Dimson and Marsh (1984)	Selected British shares	Comparison of forecast and actual return via *t*-test, Unbiasedness test	1980–1981	+
Fraser and MacDonald (1993)	CAC 40	Unbiasedness test, root mean squared error	1984–1987	–
Spiwoks (2004)	Dow Jones Industrial Index, DAX, FT-SE 100, CAC 40, MIBtel, and the Nikkei 225	Analysis of turning point errors, Theil's U, TOTA coefficient	1994–2004	–
Benke (2006)	DAX	Comparison of absolute frequencies regarding forecasting errors, direction of error, and comparison to naïve forecasts without statistical test	1992–2005	–
Spiwoks and Hein (2007)	Dow Jones Industrial Index, DAX, FT-SE 100, CAC 40, MIBtel, and the Nikkei 225	Root mean squared relative error, mean absolute relative error	1994–2004	–
Bacchetta et al. (2009)	Dow Jones Industrial Index, and Nikkei 225	Log Regression	1998–2005	+
Fujiwara et al. (2013)	TOPIX	Augmented Dickey–Fuller test, ADF-Fisher chi-square test	1998–2010	–

+ = Overall, the forecasts are assessed as good; – = overall, the forecasts are assessed as being flawed.

Spiwoks (2004) and Spiwoks and Hein (2007) consider forecasts for six international share indices (the Dow Jones Industrial Index, the DAX, the FT-SE 100, the CAC 40, MIBtel, and the Nikkei 225) issued in the period from 1994 to 2004. The results are very similar. Almost without exception, the forecast time series exhibit greater forecasting errors than the respective naïve forecast. In addition, they exhibit topically orientated

trend adjustment (Andres & Spiwoks, 1999). In other words, they reflect the present situation more than anything else, and hardly provide any insights into future trends.

Benke (2006) examines DAX forecasts for the period from 1992 to 2005. He establishes that the forecasters consistently underestimate the extent of the actual changes. Bacchetta et al. (2009) analyze forecasts for the Dow Jones Industrial Index and the Nikkei 225 in the period from 1998 to 2005. The authors conclude that the forecasts are suitable for achieving systematic excess returns. Fujiwara et al. (2013) observe TOPIX forecasts in the years from 1998 to 2010. They argue that the forecasters are too strongly orientated towards their previous forecasts and systematically underestimate the actual trends of the TOPIX.

As we want to consider the abilities of professional stock market analysts, experimental studies in which the subjects are asked to make stock market forecasts themselves (e.g., Theissen, 2007; De Bondt, 1993) are not considered here.

2.3. Hypotheses

Capital market forecasts often describe the present rather than the future. Spiwoks et al. (2015) cite the systematic underestimation of the variability of reality as a possible reason for the phenomenon of topically oriented trend adjustments in capital market forecasts. The American sociologist William Fielding Ogburn (1934) is the first to address the systematic underestimation of the variability of reality in predicting future events. He presumes that (1) unusual events (e.g., a sudden drop in an otherwise rising trendline) are forecasted too seldom, that (2) the standard deviation of the forecasts is lower than the standard deviation of the actual events, and that (3) the forecasted changes lag behind the actual changes.

We check whether the forecasts for the German Stock Market Index (DAX), the Dow Jones Industrial Index (DJI) and the Euro Stoxx 50 (SX5E) also show these three properties. In formulating the hypotheses, we assume that the observations made by Ogburn (1934) who investigated forecasts of sporting events also apply to stock market forecasts.

Unlike the DAX, the DJI and the SX5E are price indices. Nevertheless, their long-term development is considered to be non-stationary. Over the long term, a rising trend can be recognized in all three stock indices. To this extent, it is simple to define unusual and normal events. A normal event is an increase in the share price index. An unusual event is a decrease in the share price index. Hypotheses 1 and 2 are therefore:

Hypothesis 1: *Falls in stock market indices are forecasted more seldom than they occur in reality.*

Hypothesis 2: *The standard deviation of the forecasted changes of the stock market indices is lower than the standard deviation of the actual changes in the indices.*

Should the systematical underestimation of the variability of reality be true in our data basis, investors would be exposed to a high risk, as relatively large changes in trends, also

negative ones, would not be reflected adequately in the forecasts. The best way to test this assumption is to compute a prediction-realization diagram (Theil, 1958) that compares the forecasted relative share price changes to the actual relative share price changes (as described in the Methods section). If the forecast changes are smaller than the actual changes, this leads to a regression line with a slope of <1 in the prediction-realization diagram. Hypothesis 3 therefore reads:

Hypothesis 3: *The slope of the regression lines in the prediction-realization diagram is lower than one (slope < 1).*

If the predicted changes lag behind the actual changes and it is thus true that the forecasters are guided by conservatism, the forecasts are not unbiased. This can be verified best by means of the test of unbiasedness using the Mincer-Zarnowitz regression (as described in the Methods section). The use of the unbiasedness test is of particular interest here because it can be used to determine whether the underestimation of the changes in the prognosis object can be viewed as statistically significant. Hypothesis 4 is therefore:

Hypothesis 4: *The forecasts prove to be biased.*

An assessment of capital market forecasts is incomplete if the forecasts are not compared to the naïve forecasts. In view of the results of previous studies (Spiwoks & Hein, 2007; Spiwoks, 2004; Fraser & MacDonald, 1993; Lakonishok, 1980), we expect that the quality of the forecasts will not be significantly better than that of naïve forecasts. If this is the case, investors should by no means consider the forecasts, as the naïve forecast is readily available at any time. Hypothesis 5 is therefore:

Hypothesis 5: *The quality of the forecasts is not significantly higher than that of naïve forecasts.*

3 Data Basis

We evaluate DAX forecasts which were published between 1992 and 2020 in the *Handelsblatt* newspaper (HB). The forecasts have a forecast horizon of one year. In addition, we evaluate forecasts for the DAX and the Euro Stoxx 50 which were published in the period from 2002 to 2020 in the *Frankfurter Allgemeine Zeitung* (FAZ). We also analyze forecasts for the Dow Jones Industrial Index which were published between 2004 and 2020 in the FAZ. The time scales differ as we have taken into account all stock price forecasts since the beginning of their publication in order to get more meaningful results. These forecasts have forecast horizons of six and twelve months (Table 2). We provide the dataset used in our study as a supplementary in an Excel format. The dataset comprises all analyzed forecasts published annually in the *Frankfurter Allgemeine Zeitung* and *Handelsblatt* between 1992 and 2020.

In Table 2, we also provide descriptive statistics and show both the minimum and maximum predicted percentage index level changes as well as the median and mean

value of the predicted percentage index level changes for the examined data. The descriptive statistics on forecast index level changes in Table 2 are shown in percentages to give a clearer picture of the data. The institutes did not forecast percentage index level changes, but rather the respective index levels. For example, M.M. Warburg & Co. predicted the DAX index level at the end of the year 2009 at 3600 points. At the time the forecast was issued, the DAX had an index level of 4810.20 points. Thus, the institute forecast the largest price decline of 25.16%. The WGZ-Bank forecast the maximum percentage increase in the index level of the DAX in 2003. While the DAX had an index level of 2892.63 points at the time the forecast was made, the bank forecast a percentage increase of 72.85% to 5000 points at the end of the year. On average, the institutes forecast an index level increase of the DAX of 8.76% (median 8.08%) in the period considered from 1992 to 2020 (see Table A-1 in Appendix A for more detailed descriptive statistics on our data basis). In Figure 1, we provide an overview of the 12-month forecasts examined by showing the mean values of the forecasts, the associated actual index values, and the naïve forecasts.

Table 2: Data basis and summary statistics

Source	Subject	Period	N	Min (in %)	Max (in %)	Median (in %)	Mean (in %)	N	Min (in %)	Max (in %)	Median (in %)	Mean (in %)
			Forecast horizon 6 months					Forecast horizon 12 months				
HB	DAX	1992-2020	NA	NA	NA	NA	NA	964	-25.16	72.85	8.08	8.76
FAZ	DAX	2002-2020	282	-33.47	18.68	3.38	2.34	402	-25.16	45.20	8.14	8.94
FAZ	DJI	2004-2020	203	-21.45	23.06	1.62	1.39	259	-20.24	42.43	6.07	5.95
FAZ	SX5E	2002-2020	270	-34.63	22.57	3.24	2.32	381	-20.33	36.87	7.88	8.03
Σ			755					2,006				

HB = Handelsblatt; FAZ = Frankfurter Allgemeine Zeitung; DAX = German Stock Market Index; DJI = Dow Jones Industrial Index; SX5E = Euro Stoxx 50; N = number of forecasts issued; Min = minimum; Max = maximum; NA = not available.

The forecasts are from private German banks such as Fürst Fugger Privatbank or Bethmann Bank, from German state banks such as Helaba or Bayerische Landesbank, from major German banks such as Deutsche Bank or Commerzbank, and from international banks like Goldman Sachs, J.P. Morgan, or BNP Paribas. For a detailed overview of which institutes published forecasts in which newspaper, see Appendices B and C.

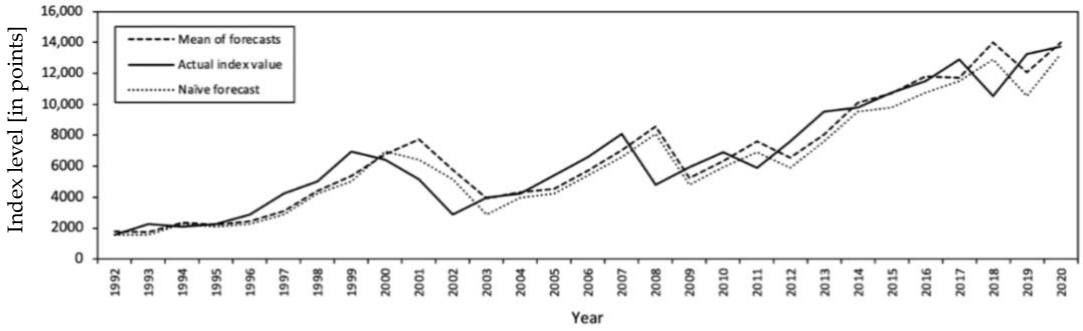
The methods applied by the individual institutions in order to obtain their forecasts are not disclosed. The forecasts are collected by HB and FAZ through annual quantitative surveys. For example, at the end of 2019, the newspapers collected and published forecasts that were drawn up for the middle and the end of 2020.

To the best of our knowledge, an analysis of the quality of actual ex-ante forecasts for the Euro Stoxx 50 has not yet been the subject of the literature (Table 1). Ex-ante forecasts

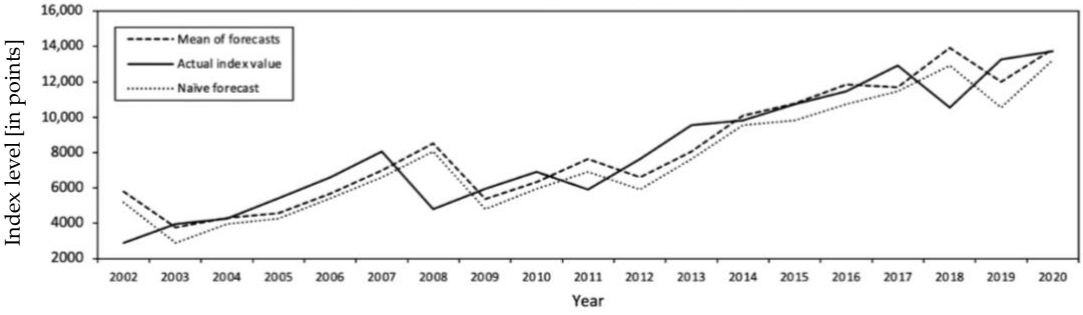
of the Dow Jones Industrial Index and the DAX have also not been considered since 2005. Since then, technological progress has led to the emergence of numerous new forecasting tools and methods, which are discussed in our literature section. Overall, our data basis consists of 2,761 forecasts covering a period of time of up to 29 years per time series. We are therefore convinced that an analysis of this data basis is a useful addition to the existing literature on stock market forecasts.

Figure 1: Means of 12M forecasts, actual index values, and naïve forecasts of the DAX, DJI, and SX5E

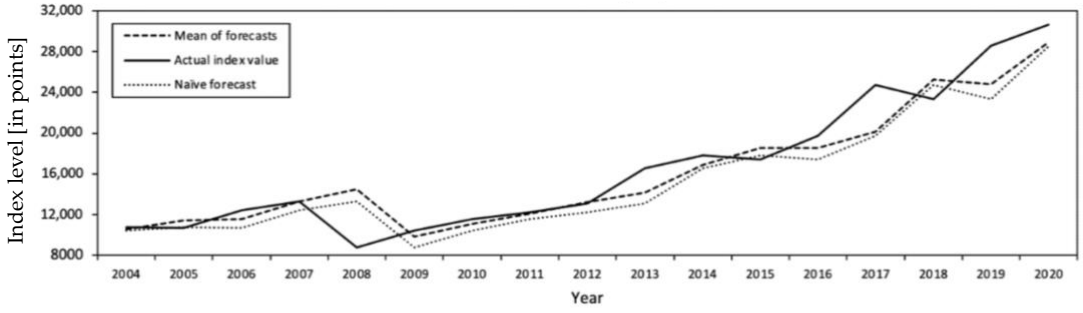
12M forecasts of the DAX (HB)



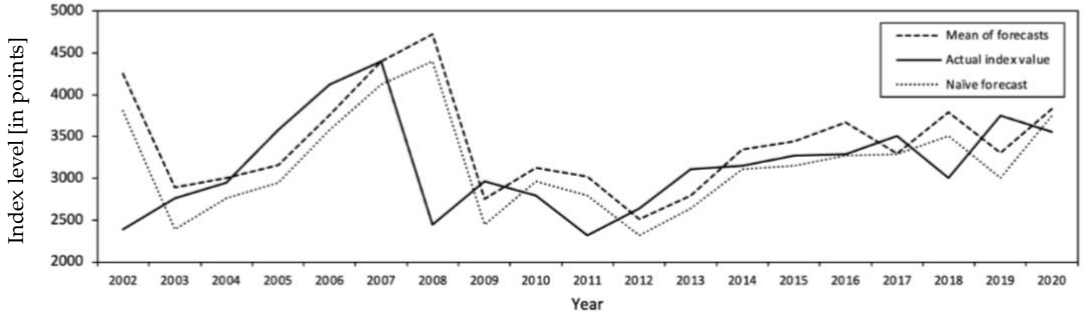
12M forecasts of the DAX (FAZ)



12M forecasts of the DJI (FAZ)



12M forecasts of the SX5E (FAZ)



4 Methods

Fundamentally, we follow Ogburn's assessment of forecasting: Ogburn (1934) assumes that forecasters suffer from conservatism. Therefore, we examine whether (1) unusual events are forecast too infrequently, (2) the standard deviation of the forecasts is lower than the standard deviation of the actual events, and (3) forecast changes lag behind actual changes. We consider these three aspects in the forecasts as a whole, but also individually for all forecasters who issue forecasts for at least ten years. In addition, we also go beyond Ogburn's methodology and include some contemporary additions to address the assessment of forecast quality from today's perspective. As statistical tools to measure the quality of the survey-based forecasts we use Theil's prediction-realization diagram (Theil, 1958), the test for the unbiasedness of the forecasts, and the Diebold–Mariano test for a comparison to the respective naïve forecast.

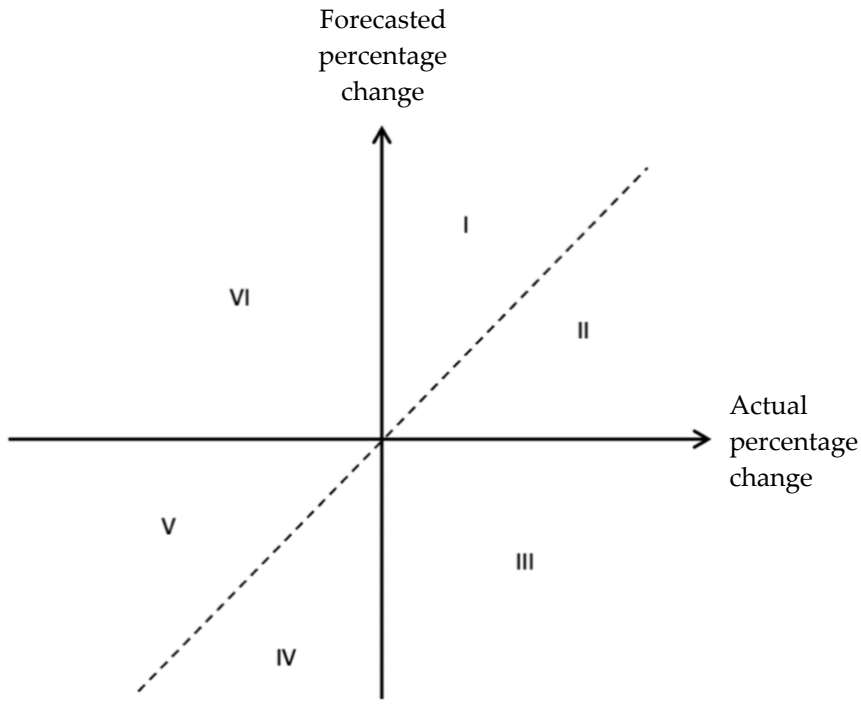
We draw on the prediction-realization diagram for a qualitative assessment of forecasting errors. For this purpose, we first calculate the forecasted relative changes (ρPF) and the realized relative actual stock price changes (ρPA). A_t shows the actual event at the time for which the forecast is applied and A_{t-h} shows the actual event at the time when the forecast was made.

$$\rho PF = \frac{P_t - A_{t-h}}{A_{t-h}} \text{ and } \rho PA = \frac{A_t - A_{t-h}}{A_{t-h}}$$

P	= forecast of the actual event
A	= actual event
t	= time
h	= forecast horizon

The forecasted percentage changes and the actual percentage changes are plotted and compared in the prediction-realization diagram (Figure 2). The dashed diagonal line in the prediction-realization diagram reflects the area in which the forecasted percentage changes and the actual realized percentage changes coincide (perfect forecasts). A good forecast time series is therefore characterized by the fact that the values are close to the diagonal. Using an OLS regression, we examine whether the slope of the regression line resulting from the forecasts considered is equal to one. When the variability of actual events is systematically underestimated, the slope of the regression lines in the prediction-realization diagram should be lower than one. A flat course of the regression lines (slope < 1) indicates an underestimation of the actual changes.

Figure 2: Prediction-realization diagram following Theil (1958)



I The percentage increase of the stock market index is overestimated.
II The percentage increase of the stock market index is underestimated.
III The stock market index rises, although a fall is forecasted.
IV The percentage decrease of the stock market index is overestimated.
V The percentage decrease of the stock market index is underestimated.
VI The stock market index falls, although a rise is forecasted.

For all forecasters who have been taking part in forecasting surveys for at least ten years, we determine the slope of the regression lines individually. All of the other forecasts are evaluated within the framework of the total number of forecasts analyzed and within the framework of the consensus forecasts.

Furthermore, we perform the unbiasedness test using the Mincer-Zarnowitz regression (Mincer & Zarnowitz, 1969) to examine whether forecasting errors are systematic. The Mincer-Zarnowitz regression takes the following form:

$$A_t = \alpha + \beta P_t + u_t$$

A_t = event that actually occurred in time t (dependent variable)
α = constant
β = coefficient of the respective forecasts
P_t = forecast of the actual event in time t
u_t = error term in time t

Based on this equation, forecasts are considered unbiased if α is not significantly different to 0, and β is not significantly different to 1. Likewise, the error term u_t may not be autocorrelated. Forecasts are considered unbiased when, with a low probability of error, the joint hypothesis of $\alpha = 0$ and $\beta = 1$ does not have to be rejected. This is checked by using the Wald test (Wald, 1943). A further condition is the absence of autocorrelation in the values of the error term u_t, which is examined with the Wooldridge test (Wooldridge, 2002). If, according to these criteria, a forecast time series is unbiased, Granger and Newbold (1974) argue that this by no means signifies that the forecasts are perfect. They merely do not exhibit any systematic errors.

Finally, we compare the forecasts with the naïve forecast. A forecaster who has obtained a notable insight into the future trend of the subject matter should at least be able to make more accurate forecasts than if one were to always assume that nothing at all will change (naïve forecast).

Simple measurements of forecast quality (such as the mean absolute squared error or the mean squared error) enable us to make a comparison with a naïve forecast. However, these simple approaches do not permit an assessment of statistical significance. This deficit is remedied by using the Diebold–Mariano test (Diebold & Mariano, 1995). To do so, we calculate the mean squared error for the time series of the expert prognoses and for the time series of the naïve forecasts. The test statistics of the Diebold–Mariano test are defined as follows:

$$DM = \frac{\frac{1}{T}\sum\left(V(P_{t1}) - V(P_{t2})\right)}{\sqrt{\hat{y}\ d/T}}$$

T	= number of observations
V	= loss function
P_1	= naïve forecast
P_2	= expert forecast
$\sqrt{\hat{y}\ d/T}$	= joint spread of the two loss functions

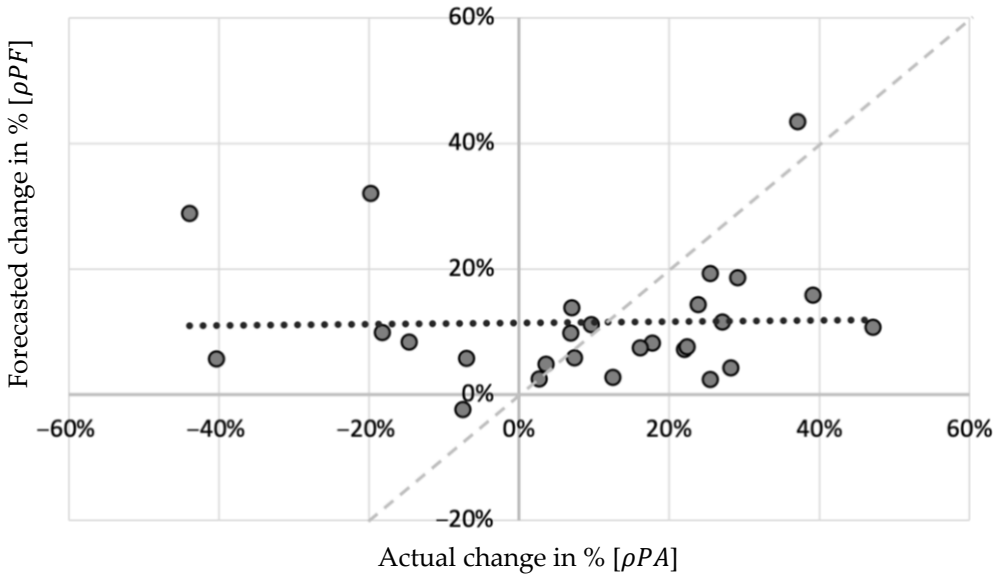
The null hypothesis tested in this way is that the naïve forecast (P_1) and the expert forecast (P_2) have the same accuracy. Neither one of the two alternatives thus provides clearly better results. The numerator is the mean deviation between the loss function V of the two forecasting approaches to be compared. Normally a squared loss function is assumed. In other words, the squared errors of the two forecast approaches are compared (P_1 and P_2). The denominator is the joint spread of the two loss functions. This is estimated on the basis of the long-term autocovariances of the loss function. In the case of large samples, this test value is asymptotically normally distributed.

As the methods and variables used by the forecasters in our data basis are not disclosed, we focus on the overall quality of the forecasts in terms of accuracy and unbiasedness. An assessment of the informative value of different forecasting approaches is not in the scope in this study.

5 Results

To provide a more detailed insight into our results, we first show the individual forecast quality of two selected German private banks. The graphic representation of the DAX forecasts of the German private bank Berenberg in a prediction-realization diagram indicates that conservative forecasting is at work here (Figure 3).

Figure 3: Prediction-realization diagram of the DAX forecasts of Berenberg



Dotted line = regression line; dashed line = perfect forecasts according to the prediction-realization diagram.

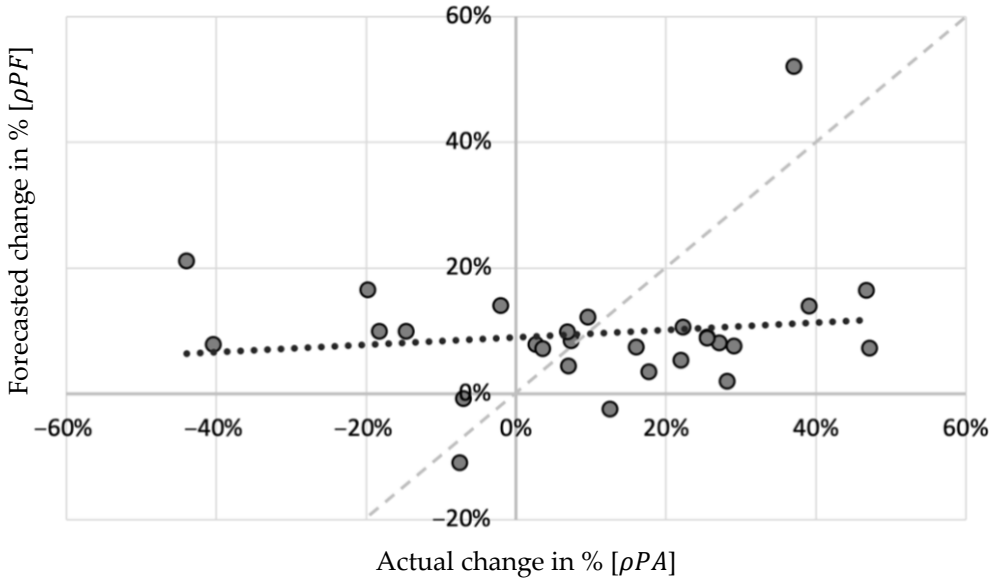
Berenberg issued a total of 27 DAX forecasts in the observation period (1992–2020). It is recognizable straight away that only one fall in the DAX is forecasted (3rd quadrant), but that the DAX actually does fall in seven out of the 27 years (3rd and 4th quadrant). This means that unusual events (falls in the DAX) are under-represented in the forecasts.

In addition, it can be seen that the dispersion of the actual events (scattering along the ρPA axis) is greater than the dispersion of the forecasts (scattering along the ρPF axis). The standard deviation of the actual events is 22.76%. The standard deviation of the forecasts, however, is only 9.98% (Table 3). The slope in the dotted regression line in the prediction-realization diagram of 0.011 is thus nowhere near the threshold value 1 (dashed diagonal line) (Table 3). The variability of the actual events is dramatically underestimated.

As another example, we consider the prediction-realization diagram of DAX forecasts made by the Franco-German private bank Oddo BHF (Figure 4).

This reveals a picture which is very similar to that of the prediction-realization diagram for Berenberg. In the period 1992–2020, at the end of each year Oddo BHF forecasts the DAX for the coming year. This occurs a total of 28 times. A fall in the DAX is forecasted on three occasions. In reality, however, the DAX falls in eight of the 28 years. This means that unusual events (falls in the DAX) are under-represented in the forecasts.

Figure 4: Prediction-realization diagram of the DAX forecasts of Oddo BHF



Dotted line = regression line; dashed line = perfect forecasts according to the prediction-realization diagram.

In addition, it can be seen that the dispersion of the actual events (scattering along the ρPA axis) is greater than the dispersion of the forecasts (scattering along the ρPF axis). The standard deviation of the actual events is 23.39%. The standard deviation of the forecasts, however, is only 10.41% (Table 3). The slope of 0.059 in the dotted regression line in the prediction-realization diagram is thus nowhere near the threshold value 1 (dashed diagonal line) (Table 3). The variability of the actual events is dramatically underestimated.

Table 3 depicts the main results of the DAX forecasts from the *Handelsblatt* newspaper. All of the forecasters who have taken part in the forecasting surveys of the *Handelsblatt* for at least ten years are analyzed individually. All of the forecasters who issue less than 10 forecasts in the period from 1992 to 2020 are not analyzed individually but are taken into account as part of the overall analysis of all forecasts and within the framework of the consensus forecasts (final lines in Table 3).

The seventh column of Table 3 indicates whether fewer falls in the DAX are forecasted than actually occur. As the DAX is a performance index and exhibits a rising trend over the long term, all falls in the index are interpreted as ´unusual events´. According to

Ogburn (1934), conservative forecasting leads to ´normal events´ (here: an increase in the DAX) being over-represented in the forecasts, while 'unusual events' (here: a decrease in the DAX) are under-represented in the forecasts. This is the case in 33 of the 38 forecasters who are analyzed individually here: a proportion of 86.8%. Unusual events are also under-represented in the consensus forecasts and when the total number of the forecasts is considered as a whole. The detailed data is given in Table 3, where one can see how often a falling DAX was forecast, and how often the DAX really falls. One can also note how often an upward trend was forecast for the DAX, and how often the DAX really rises (Table 3).

The picture is clearer in the case of the standard deviations. According to Ogburn (1934), conservative forecasting leads to standard deviations of the forecasts which are lower than the standard deviations of the actual events. The tenth column of Table 3 considers whether this applies to the DAX forecasts and reveals that this is the case in all 38 of the 38 forecasters analyzed. Also, with regard to the consensus forecasts and when all 964 forecasts are considered, the standard deviation of the forecasts lags behind the standard deviation of the actual events (Table 3).

Ogburn (1934) states that conservative forecasting leads to an underestimation of the variability of reality. In the prediction-realization diagram, this should lead to a slope in the regression lines which is lower than one. The last column of Table 3 illustrates this aspect. It can be seen that in 38 out of 38 cases, the slope in the regression lines is lower than one. The fact that the slopes are usually clearly below the threshold value of one is also revealed in the detailed data on the intercepts and the slopes in the regression lines (Table 3).

The German quality newspaper the *Frankfurter Allgemeine Zeitung* (FAZ) only started a regular survey of forecasts in 2002. As a result, the share price falls in the years 2000 and 2001 no longer have an effect. It is interesting to see whether this leads to significantly different results in the forecasts. In addition, the *Frankfurter Allgemeine Zeitung* not only surveys annual forecasts, but also six-month forecasts. It is quite possible that the characteristics of the forecasts with differing forecast horizons vary. Once again, all of the forecasters who have taken part in the forecasting surveys of the FAZ at least ten times are analyzed individually (Table 4).

Table 3: Main results of the DAX forecasts from 1992 to 2020 from the *Handelsblatt*

Institution	Forecasts issued	Forecast		Actual		Normal events over-represented in the forecasts	Standard deviation		SD of the forecasts < SD of the actual events	Regression line		Slope of the regression lines < 1
		DAX falls	DAX rises	DAX falls	DAX rises		Forecast	Actual		Intercept	Slope	
Bank Julius Bär	23	2	21	8	15	Yes	0.062	0.248	Yes	0.088	-0.023	Yes
Bank of America	11	0	11	2	9	Yes	0.066	0.207	Yes	0.117	-0.001	Yes
Bankhaus Lampe	25	1	24	6	19	Yes	0.081	0.234	Yes	0.089	0.097	Yes
Bayerische LB	26	1	25	6	20	Yes	0.067	0.230	Yes	0.080	-0.006	Yes
Berenberg	27	1	26	7	20	Yes	0.100	0.228	Yes	0.114	0.011	Yes
Bethmann Bank	12	2	10	5	7	Yes	0.095	0.284	Yes	0.101	-0.109	Yes
BNP Paribas	18	3	15	4	14	Yes	0.061	0.223	Yes	0.056	0.140	Yes
Commerzbank	28	2	26	7	21	Yes	0.089	0.234	Yes	0.120	-0.064	Yes
Credit Suisse	13	2	11	5	8	Yes	0.072	0.290	Yes	0.106	0.059	Yes
Dekabank	19	1	18	4	15	Yes	0.101	0.227	Yes	0.090	0.154	Yes
Deutsche Bank	25	2	23	7	18	Yes	0.070	0.237	Yes	0.091	-0.043	Yes
Dresdner Bank	15	0	15	5	10	Yes	0.084	0.276	Yes	0.080	0.099	Yes
DZ Bank	29	7	22	8	21	Yes	0.107	0.231	Yes	0.073	0.088	Yes
Haspa	13	0	13	3	10	Yes	0.047	0.202	Yes	0.080	0.045	Yes
Hauck & Aufh.	26	5	21	6	20	Yes	0.101	0.235	Yes	0.072	-0.040	Yes
Helaba	28	8	20	7	21	No	0.108	0.234	Yes	0.053	0.092	Yes
HSBC Trinkaus	22	3	19	7	15	Yes	0.085	0.256	Yes	0.080	-0.022	Yes
J.P. Morgan	22	4	18	6	16	Yes	0.100	0.244	Yes	0.084	0.038	Yes
LBB	18	3	15	6	12	Yes	0.140	0.233	Yes	0.088	0.027	Yes
LBBW	21	1	20	6	15	Yes	0.107	0.226	Yes	0.090	0.093	Yes
Lehman Brothers	12	5	7	4	8	No	0.098	0.259	Yes	0.040	0.062	Yes
M.M. Warburg	29	3	26	8	21	Yes	0.091	0.231	Yes	0.076	-0.016	Yes
Morgan Stanley	14	6	8	4	10	No	0.123	0.285	Yes	0.030	0.136	Yes
National-Bank	15	3	12	3	12	No	0.086	0.202	Yes	0.082	0.028	Yes
NATIXIS	17	1	16	3	14	Yes	0.065	0.231	Yes	0.077	0.057	Yes
NordLB	12	2	10	2	10	No	0.038	0.153	Yes	0.041	-0.089	Yes
Oddo BHF	28	3	25	8	20	Yes	0.104	0.234	Yes	0.090	0.059	Yes
Pictet & Cie.	13	3	10	5	8	Yes	0.114	0.279	Yes	0.092	-0.074	Yes
Postbank	11	0	11	3	8	Yes	0.069	0.225	Yes	0.098	0.087	Yes
Sal. Oppenheim	21	2	19	5	16	Yes	0.093	0.248	Yes	0.067	0.111	Yes
Santander	24	1	23	7	17	Yes	0.093	0.239	Yes	0.116	0.101	Yes
Société Générale	20	4	16	5	15	Yes	0.096	0.228	Yes	0.065	0.043	Yes
SYZ & Co.	10	0	10	2	8	Yes	0.058	0.235	Yes	0.144	-0.042	Yes
UBS	14	3	11	4	10	Yes	0.120	0.242	Yes	0.112	0.007	Yes
Unicredit HVB	28	3	25	8	20	Yes	0.079	0.233	Yes	0.083	0.043	Yes
VP Bank	11	1	10	2	9	Yes	0.042	0.155	Yes	0.084	0.034	Yes
WestLB	21	3	18	7	14	Yes	0.106	0.260	Yes	0.081	0.124	Yes
WGZ Bank	16	1	15	5	11	Yes	0.172	0.211	Yes	0.110	0.301	Yes
Consensus	29	1	28	8	21	Yes	0.065	0.231	Yes	0.085	0.037	Yes
All forecasts	964	117	847	264	700	Yes	0.091	0.230	Yes	0.084	0.034	Yes

DAX = German Stock Market Index; SD = Standard deviation.

The results are in fact somewhat less clear than those for the DAX forecasts from the *Handelsblatt*. In 24 out of 33 cases (72.7%), normal events (increase in the DAX) are over-represented in the forecasts (seventh column in Table 4). Unusual events are also under-represented in the consensus forecasts and when all 282 six-month and all 402 twelve-month forecasts are considered as a whole.

The result of the standard deviations is quite clear: In 31 out of 33 cases (93.9%), the forecasts lag behind the actual events (tenth column in Table 4). This finding also applies to the consensus forecasts as well as when all 282 six-month and all 402 twelve-month forecasts are considered as a whole.

The fact that the forecasters persistently underestimate the variability of reality is revealed most clearly in the slope of the regression lines in the prediction-realization diagram (last column in Table 4). In 33 out of 33 cases, the slope is below one. This result also applies to the consensus forecasts as well as when all 282 six-month and all 402 twelve-month forecasts are considered as a whole.

The forecasts of the Dow Jones Industrial Index yield only slightly different results. Once again, all of the forecasters who have taken part in the forecasting survey at least ten times are analyzed individually (Table 5).

The Dow Jones Industrial Index is a price index, but it exhibits a long-term rising trend, nevertheless. To this extent, one can also presume here that a rise in the index can be considered a normal event, and that a fall in the index represents an unusual event. In ten out of 16 cases (62.5%), normal events (increase of the Dow Jones Industrial Index) are over-represented in the forecasts (seventh column in Table 5). Unusual events are also under-represented in the consensus forecasts and when all 203 six-month and all 259 twelve-month forecasts are considered as a whole.

The result for the standard deviations is more marked. In 14 out of 16 cases (87.5%), the fluctuations in the forecasts lag behind those of the actual events (tenth column in Table 5). This finding also applies to the consensus forecasts as well as when all 203 six-month and all 259 twelve-month forecasts are considered as a whole.

The fact that the forecasters persistently underestimate the variability of reality is revealed most clearly in the slope of the regression lines in the prediction-realization diagram (last column in Table 5). In 16 out of 16 cases, the slope is below one. This is also the same for the consensus forecasts as well as when all 203 six-month and all 259 twelve-month forecasts are viewed as a whole.

The picture drawn by the forecasts of the Euro Stoxx 50 is even more distinct (Table 6). Here again, all of the forecasters who have taken part in the forecasting survey at least ten times are analyzed individually. All of the other forecasts form part of the consensus forecasts and are also evaluated as part of the total number of forecasts.

Table 4: Main results of the DAX forecasts from 2002 to 2020 from the FAZ

Institution	Forecasts issued	Forecast		Actual		Normal events over-represented in the forecasts	Standard deviation		SD of the forecasts < SD of the actual events	Regression line		Slope of the regression lines < 1
		DAX falls	DAX rises	DAX falls	DAX rises		Forecast	Actual		Intercept	Slope	
Forecast horizon 6 months												
Bayern LB	10	5	5	3	7	No	0.047	0.094	Yes	0.028	-0.286	Yes
Deka Bank	16	3	13	5	11	Yes	0.061	0.096	Yes	0.040	-0.002	Yes
DZ Bank	16	6	10	5	11	No	0.065	0.096	Yes	0.009	0.032	Yes
Helaba	14	6	8	5	9	No	0.075	0.102	Yes	0.025	-0.375	Yes
HSH Nordbank	10	7	3	4	6	No	0.095	0.098	Yes	-0.030	-0.039	Yes
HVB-Unicredit B.	16	4	12	6	10	Yes	0.063	0.104	Yes	0.035	-0.035	Yes
LBBW	17	3	14	6	11	Yes	0.048	0.102	Yes	0.019	0.090	Yes
M.M. Warburg	17	3	14	6	11	Yes	0.122	0.102	No	0.030	-0.039	Yes
Oddo BHF	10	1	9	4	6	Yes	0.041	0.121	Yes	0.049	-0.058	Yes
Postbank	13	6	7	4	9	No	0.071	0.104	Yes	0.008	-0.087	Yes
Santander A. Mgmt.	13	1	12	3	10	Yes	0.029	0.099	Yes	0.033	0.073	Yes
Société Générale	10	6	4	3	7	No	0.087	0.072	No	-0.023	-0.431	Yes
Consensus	17	2	15	6	11	Yes	0.028	0.102	Yes	0.024	-0.077	Yes
All forecasts	282	83	199	103	179	Yes	0.072	0.095	Yes	0.024	-0.076	Yes
Forecast horizon 12 months												
Allianz SE	11	0	11	2	9	Yes	0.044	0.155	Yes	0.072	0.018	Yes
Bayern LB	11	0	11	2	9	Yes	0.036	0.159	Yes	0.069	0.011	Yes
BNP Paribas	12	1	11	3	9	Yes	0.055	0.210	Yes	0.066	0.110	Yes
Commerzbank	18	0	18	4	14	Yes	0.081	0.233	Yes	0.119	0.032	Yes
Deka Bank	18	1	17	3	15	Yes	0.104	0.195	Yes	0.082	0.200	Yes
Deutsche Bank	10	0	10	2	8	Yes	0.047	0.212	Yes	0.104	-0.017	Yes
DWS	13	0	13	3	10	Yes	0.027	0.202	Yes	0.076	0.038	Yes
DZ Bank	18	2	16	4	14	Yes	0.066	0.222	Yes	0.072	0.063	Yes
Helaba	15	6	9	3	12	No	0.121	0.196	Yes	0.025	0.249	Yes
HSBC Tk.&Bh.	13	2	11	3	10	Yes	0.066	0.262	Yes	0.065	-0.102	Yes
HSH Nordbank	11	2	9	3	8	Yes	0.080	0.213	Yes	0.055	0.192	Yes
HVB-Unicredit B.	18	1	17	4	14	Yes	0.078	0.228	Yes	0.077	0.077	Yes
J.P. Morgan	12	1	11	3	9	Yes	0.064	0.233	Yes	0.095	0.140	Yes
LBBW	19	0	19	4	15	Yes	0.097	0.227	Yes	0.091	0.093	Yes
M.M. Warburg	19	1	18	4	15	Yes	0.097	0.227	Yes	0.078	-0.018	Yes
Oddo BHF	17	1	16	4	13	Yes	0.045	0.225	Yes	0.093	-0.092	Yes
Postbank	14	0	14	3	11	Yes	0.070	0.208	Yes	0.096	0.048	Yes
Santander A. Mgmt.	16	0	16	3	13	Yes	0.052	0.195	Yes	0.107	0.048	Yes
Société Générale	11	4	7	2	9	No	0.088	0.155	Yes	0.067	-0.347	Yes
UBS	10	1	9	1	9	No	0.118	0.151	Yes	0.136	0.027	Yes
WestLB	11	2	9	3	8	Yes	0.128	0.282	Yes	0.075	0.204	Yes
Consensus	19	0	19	4	15	Yes	0.061	0.227	Yes	0.087	0.064	Yes
All forecasts	402	31	371	88	314	Yes	0.083	0.215	Yes	0.085	0.054	Yes

DAX = German Stock Market Index; FAZ = Frankfurter Allgemeine Zeitung; SD = Standard deviation.

Table 5: Main results of the forecasts of the DJI from 2004 to 2020 from the FAZ

Institution	Forecasts issued	Forecast		Actual		Normal events over-represented in the forecasts	Standard deviation		SD of the forecasts < SD of the actual events	Regression line		Slope of the regression lines < 1
		DJI falls	DJI rises	DJI falls	DJI rises		Forecast	Actual		Intercept	Slope	
Forecast horizon 6 months												
Deka Bank	15	5	10	8	7	Yes	0.070	0.066	No	0.018	0.171	Yes
Helaba	14	6	8	6	8	No	0.081	0.077	No	0.019	-0.406	Yes
LBBW	16	7	9	8	8	Yes	0.052	0.073	Yes	0.010	0.116	Yes
M.M. Warburg	15	3	12	7	8	Yes	0.061	0.075	Yes	0.034	0.233	Yes
Postbank	12	6	6	5	7	No	0.053	0.079	Yes	0.003	0.035	Yes
Santander A. Mgmt.	13	1	12	6	7	Yes	0.019	0.081	Yes	0.026	-0.095	Yes
Consensus	16	4	12	8	8	Yes	0.019	0.073	Yes	0.014	0.036	Yes
All forecasts	203	67	136	106	97	Yes	0.061	0.070	Yes	0.014	0.040	Yes
Forecast horizon 12 months												
BNP Paribas	10	0	10	3	7	Yes	0.040	0.183	Yes	0.072	-0.059	Yes
Commerzbank	10	0	10	3	7	Yes	0.052	0.169	Yes	0.081	0.120	Yes
Deka Bank	16	6	10	4	12	No	0.099	0.137	Yes	0.051	0.002	Yes
Helaba	15	7	8	3	12	No	0.107	0.149	Yes	0.008	0.193	Yes
HSH Nordbank	11	5	6	3	8	No	0.067	0.163	Yes	0.022	-0.032	Yes
LBBW	17	4	13	4	13	No	0.058	0.142	Yes	0.053	-0.042	Yes
M.M. Warburg	17	1	16	4	13	Yes	0.071	0.142	Yes	0.063	-0.107	Yes
Oddo BHF	15	0	15	3	12	Yes	0.022	0.147	Yes	0.058	0.054	Yes
Postbank	13	0	13	3	10	Yes	0.063	0.160	Yes	0.084	0.012	Yes
Santander A. Mgmt.	16	0	16	4	12	Yes	0.051	0.146	Yes	0.070	0.093	Yes
Consensus	17	0	17	4	13	Yes	0.033	0.142	Yes	0.055	0.006	Yes
All forecasts	259	33	226	65	194	Yes	0.066	0.140	Yes	0.057	0.029	Yes

DJI = Dow Jones Industrial Index; FAZ = Frankfurter Allgemeine Zeitung; SD = Standard deviation.

Conservatism among forecasters can lead to them forecasting unusual events too rarely. The Euro Stoxx 50 is a price index, but in spite of this it exhibits a long-term upward trend. To this extent, one can also presume here that a rise in the index can be considered a normal event, and that a fall in the index represents an unusual event. In the predictions of 24 of the 26 forecasters analyzed individually (92.3%), unusual events are under-represented (seventh column in Table 6). The consensus forecasts and the overall total of all 270 six-month forecasts and all 381 twelve-month forecasts also show that unusual events are forecast more seldom than they occur in reality.

The standard deviations provide a very clear picture. The standard deviations of the forecasts lag behind the standard deviations of the actual results in 26 out of 26 cases (tenth column in Table 6). This also applies to the consensus forecasts and the overall total of 270 forecasts with a forecast horizon of six months and all 381 forecasts with a forecast horizon of twelve months.

Table 6: The main results for the Euro Stoxx 50 forecasts from 2002 to 2020 from the FAZ

Institution	Forecasts issued	Forecast		Actual		Normal events over-represented in the forecasts	Standard deviation		SD of the forecasts < SD of the actual events	Regression line		Slope of the regression lines < 1
		SX5E falls	SX5E rises	SX5E falls	SX5E rises		Forecast	Actual		Intercept	Slope	
Forecast horizon 6 months												
Bayern LB	10	4	6	5	5	Yes	0.043	0.078	Yes	0.011	-0.244	Yes
Deka Bank	16	3	13	8	8	Yes	0.063	0.093	Yes	0.049	0.022	Yes
DZ Bank	16	3	13	8	8	Yes	0.064	0.093	Yes	0.030	0.186	Yes
Helaba	14	6	8	8	6	Yes	0.079	0.095	Yes	0.019	-0.406	Yes
HSH Nordbank	10	6	4	6	4	No	0.085	0.099	Yes	-0.030	-0.214	Yes
HVB-Unicredit B.	16	3	13	8	8	Yes	0.070	0.101	Yes	0.023	-0.085	Yes
LBBW	17	6	11	9	8	Yes	0.053	0.098	Yes	0.028	0.088	Yes
M.M. Warburg	16	2	14	8	8	Yes	0.073	0.101	Yes	0.055	-0.014	Yes
Oddo BHF	10	2	8	5	5	Yes	0.042	0.116	Yes	0.033	-0.009	Yes
Postbank	13	6	7	7	6	Yes	0.060	0.097	Yes	0.004	-0.100	Yes
Santander A. Mgmt.	13	2	11	6	7	Yes	0.033	0.099	Yes	0.030	0.110	Yes
Consensus	17	5	12	9	8	Yes	0.030	0.098	Yes	0.023	-0.018	Yes
All forecasts	270	82	188	144	126	Yes	0.073	0.094	Yes	0.023	-0.007	Yes
Forecast horizon 12 months												
Allianz SE	11	0	11	4	7	Yes	0.042	0.130	Yes	0.071	-0.035	Yes
Bayern LB	11	0	11	3	8	Yes	0.039	0.127	Yes	0.058	-0.044	Yes
BNP Paribas	11	1	10	3	8	Yes	0.044	0.194	Yes	0.076	-0.069	Yes
Commerzbank	18	1	17	5	13	Yes	0.064	0.195	Yes	0.080	0.017	Yes
Deka Bank	18	1	17	5	13	Yes	0.093	0.170	Yes	0.094	0.107	Yes
DWS	12	0	12	5	7	Yes	0.043	0.175	Yes	0.078	-0.019	Yes
DZ Bank	18	1	17	6	12	Yes	0.075	0.193	Yes	0.090	0.096	Yes
Helaba	15	5	10	5	10	No	0.117	0.177	Yes	0.048	0.292	Yes
HSBC Tk.&Bh.	14	3	11	4	10	Yes	0.082	0.209	Yes	0.065	-0.141	Yes
HSH Nordbank	11	1	10	4	7	Yes	0.071	0.195	Yes	0.076	0.119	Yes
HVB-Unicredit B.	18	0	18	6	12	Yes	0.064	0.193	Yes	0.070	0.050	Yes
LBBW	19	1	18	6	13	Yes	0.078	0.190	Yes	0.088	0.003	Yes
M.M. Warburg	19	1	18	6	13	Yes	0.083	0.190	Yes	0.074	-0.073	Yes
Oddo BHF	17	1	16	6	11	Yes	0.047	0.192	Yes	0.072	-0.074	Yes
Postbank	14	0	14	4	10	Yes	0.054	0.190	Yes	0.086	0.032	Yes
Santander A. Mgmt.	16	0	16	5	11	Yes	0.053	0.178	Yes	0.095	0.078	Yes
WestLB	11	1	10	4	7	Yes	0.088	0.231	Yes	0.073	0.127	Yes
Consensus	19	0	19	6	13	Yes	0.044	0.190	Yes	0.083	0.020	Yes
All forecasts	381	29	352	123	258	Yes	0.073	0.179	Yes	0.080	0.017	Yes

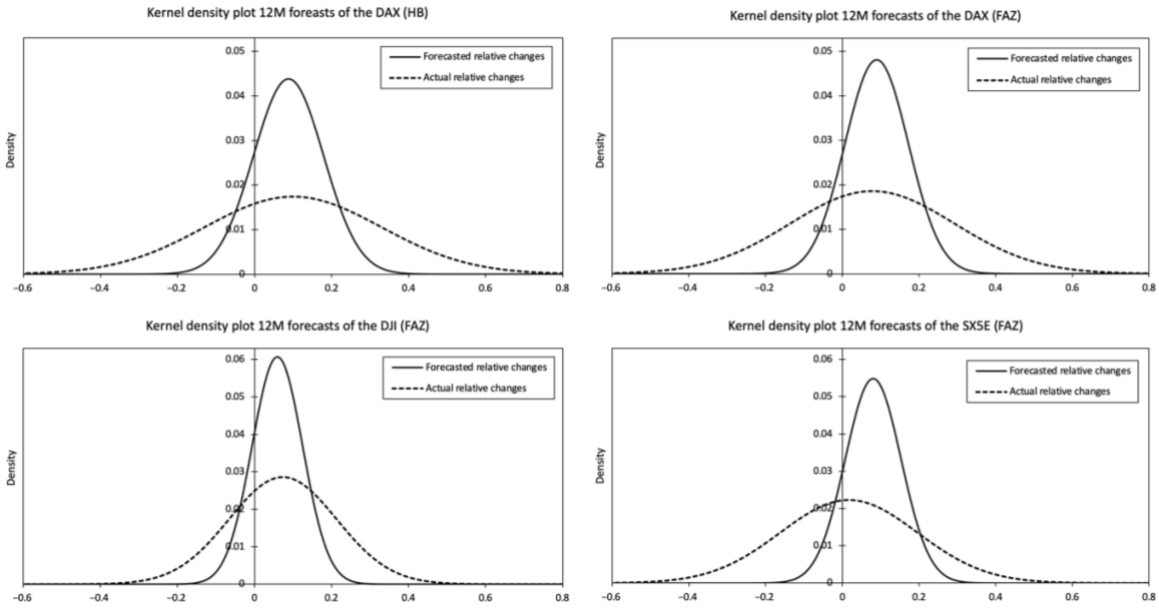
SX5E = Euro Stoxx 50; FAZ = Frankfurter Allgemeine Zeitung; SD = Standard deviation.

Finally, it can be seen that the slope in the regression lines in the prediction-realization diagrams is significantly below one in 26 out of 26 cases. The forecasters are thus obviously underestimating the variability of reality (last column in Table 6). These

findings are also confirmed when the consensus forecasts and the overall total number of forecasts are considered.

Without exception, it can be observed that the forecasters underestimate the variability of reality. This fact can be clearly seen when looking at the kernel density plots of the forecast relative changes in share prices and the actual relative changes in share prices (Figure 5). The spread of the forecasts is much smaller than the spread of the actual events.

Figure 5: Kernel density plots of the forecast and actual changes of stock-market indices



HB = Handelsblatt; FAZ = Frankfurter Allgemeine Zeitung; DAX = German Stock Market Index; DJI = Dow Jones Industrial Index; SX5E = Euro Stoxx 50; 12M = Forecast horizon of 12 months.

This can be also seen in the fact that the slope in the regression lines in the prediction-realization diagram always remains below the threshold value of one. This leads us to the assessment that this aspect in particular deserves special attention. The unbiasedness test takes the slope of the regression line in the prediction-realization diagram into account as an essential element. Forecasts are viewed as unbiased when the slope in the regression line does not diverge significantly from one, the intercept of the regression line does not deviate significantly from zero, and the residuals are randomly distributed. The decisive advantage of this approach lies in the opportunity to go beyond purely descriptive statistics and to examine the statistical significance of the results.

In all seven cases, it can be seen that given an error probability of ≤ 1% either the slope of the regression line in the prediction-realization diagram is ≠ 1 and/or the intercept is ≠ 0. In addition, the residuals are obviously not randomly distributed in six of the seven cases. The forecasts are clearly not unbiased (Table 7).

Table 7: Unbiasedness test

Stock market index	Source	Forecast horizon	Number of observations	Slope	Intercept	F test p-value	Wooldridge test p-value
DAX	HB	12M	964	0.034	0.084	0.000	0.000
DAX	FAZ	6M	282	-0.075	0.024	0.000	0.000
DAX	FAZ	12M	402	0.054	0.085	0.000	0.006
DJI	FAZ	6M	203	0.040	0.014	0.010	0.098
DJI	FAZ	12M	259	0.029	0.057	0.000	0.623
SX5E	FAZ	6M	270	-0.007	0.023	0.000	0.091
SX5E	FAZ	12M	381	0.017	0.080	0.000	0.042

DAX = German Stock Market Index; DJI = Dow Jones Industrial Index; SX5E = Euro Stoxx 50; HB = Handelsblatt; FAZ = Frankfurter Allgemeine Zeitung; 12M = 12 months; 6M = 6 months.

Finally, with the aid of the Diebold–Mariano test we examine whether the quality of the forecasts is significantly superior—from a statistical perspective—to that of naïve forecasts (Table 8). The result is that the forecasts of the Euro Stoxx 50 are significantly poorer than the corresponding naïve forecasts, and the quality of the forecasts for the DAX and the Dow Jones Industrial Index does not go significantly beyond that of naïve forecasts.

Table 8: Comparison of the forecasts with the naïve forecast

Stock market index	Source	Forecast horizon	Diebold-Mariano test	
			Result	p-value
DAX	HB	12M	o	0.8143
DAX	FAZ	6M	o	0.1221
DAX	FAZ	12M	o	0.7429
DJI	FAZ	6M	o	0.7053
DJI	FAZ	12M	o	0.3491
SX5E	FAZ	6M	-	0.0000
SX5E	FAZ	12M	-	0.0540

o = no significant result, - = significantly poorer than the naïve forecasts, + = significantly better than the naïve forecast, DAX = German Stock Market Index; DJI = Dow Jones Industrial Index; SX5E = Euro Stoxx 50; HB = Handelsblatt; FAZ = Frankfurter Allgemeine Zeitung; 12M = 12 months; 6M = 6 months.

In Table 9 the results of the hypothesis testing are summarized. In Hypotheses 1–3, the result which was determined for "all forecasts" in a forecasting area is used. In the case of the DAX forecasts from the *Handelsblatt* survey, for example, that is the 964 forecasts which are noted in the final line of Table 3. For Hypothesis 4, the results of the unbiasedness test (Table 7) are taken into account, and for Hypothesis 5 the results of the Diebold-Mariano test (Table 8).

Table 9: The results of hypothesis testing

Stock market index	Source	Forecast horizon	Hypothesis 1	Hypothesis 2	Hypothesis 3	Hypothesis 4	Hypothesis 5
DAX	HB	12M	+	+	+	+	+
DAX	FAZ	6M	+	+	+	+	+
DAX	FAZ	12M	+	+	+	+	+
DJI	FAZ	6M	+	+	+	+	+
DJI	FAZ	12M	+	+	+	+	+
SX5E	FAZ	6M	+	+	+	+	+
SX5E	FAZ	12M	+	+	+	+	+

+ = null hypothesis rejected; - = null hypothesis not rejected; DAX = German Stock Market Index; DJI = Dow Jones Industrial Index; SX5E = Euro Stoxx 50; HB = Handelsblatt; FAZ = Frankfurter Allgemeine Zeitung; 12M = 12 months; 6M = 6 months.

In the case of Hypothesis 1 there is a uniform pattern for all areas of forecasting and all forecast horizons. Normal events (index rises) are over-represented in the forecasts. Unusual events (index falls) are under-represented in the forecasts. Null Hypothesis 1 has to be rejected in all seven cases.

In the case of Hypothesis 2 there are no differences between the subjects of the forecasts and the forecast horizons. In all seven cases, Null Hypothesis 2 has to be rejected. The dispersion of the forecasts (measured against the standard deviation) thus lags behind the dispersion of the actual events.

A uniform picture is also shown with regard to Hypothesis 3. In all seven forecasting areas the slope of the regression line in the prediction-realization diagrams is clearly below one. Null Hypothesis 3 has to be rejected in all seven cases. This means that the rates of change of the stock-market indices are significantly underestimated.

In the case of Hypothesis 4 there are also no relevant differences regarding the subjects of the forecasts or the forecast horizons. In all seven areas, the forecasts prove to be biased. These results are highly significant. In all seven cases, Null Hypothesis 4 has to be rejected.

In Hypothesis 5 there is also a concurring result for all seven forecast groups. Null Hypothesis 5 has to be discarded. The precision of the forecasts does not go significantly beyond that of naïve forecasts.

The findings of Ogburn (1934) are thus fully confirmed in the stock market forecasts which we analyzed. It can certainly be stated that these stock-market analysts systematically underestimate the variability of reality and that the success rate of their forecasts does not extend beyond that of naïve forecasts. Their behavior can be described as "sticky" because their forecasts adhere too strongly to long-term trends in the indices to provide meaningful information about current events.

This study expands on existing research as it is the first of its kind to analyze ex-ante forecasts for the SX5E. The picture obtained is similar to that of the stock indices examined previously. The forecasts are mostly biased and not significantly better than naïve forecasts. About 15 years ago, ex-ante forecasts for the DAX and the DJI were last examined (Table 1). In the meantime, technological progress has led to the emergence of numerous promising new forecasting methods, as discussed in our literature review. However, our results indicate that this has not, at least so far, contributed to a significant increase in the quality of the forecast.

Our findings allow different conclusions to be drawn with regard to the efficient market hypothesis (Fama, 1970). On the one hand, the Diebold–Mariano test shows that the forecast quality is poor. This is compatible with the efficient market hypothesis, since no excess returns can be achieved on the basis of the forecasts. On the other hand, the efficient market hypothesis assumes that economic subjects are fully informed. The permanent underestimation of the variability of reality that the prediction-realization diagram reveals should therefore not occur. The acting subjects do not seem to take notice of the discrepancy between their own actions and reality, since no correction of the behavior is made in the subsequent forecasts.

The forecasters systematically underestimate the variability of reality. Against the background of Mandelbrot's fractal theory, it seems reasonable to conclude that forecasters—as long as they think in terms of "trending" and "mean reversion"—systematically underestimate the Hurst exponent (Mandelbrot, 2004) of stock market developments.

Overall, the forecast quality for all three indices is not sufficient to enable an active investment strategy on the basis of the forecasts that is likely to be successful. Moreover, since unusual events (e.g., a sudden drop in an otherwise rising trendline) are seldom successfully forecasted, an active investment strategy based on the forecasts harbors risks that can cause severe financial damage to investors. Thus, we advise private and professional investors to consider a passive investment strategy instead when deciding how to invest their assets.

The path which has to be followed to obtain better stock market forecasts thus becomes clear: analysts have to be more courageous. They need to react to new trends with more flexibility. They have to leave their comfort zone more frequently and stand by assessments which are not necessarily approved of by the majority of their peers. That alone will presumably not suffice to generate reliable stock market forecasts: they will also need to work hard on the quality of their approaches to forecasting. To this end, a variety of interesting approaches are already discussed in the literature, e.g., economic forecasts based on newspaper texts or news from online media and attention to news events (Milas et al., 2021; Kalamara et al., 2020; Ben-Rephael et al., 2017). If analysts want to significantly improve the reliability of their forecasts, there is no alternative but to change their overly cautious, highly conservative, and thus inflexible attitudes.

Finally, our study also has some limitations. First of all, it should be mentioned that we are looking at forecasts for entire stock indices. Even if the forecasters do not manage to successfully predict the development of a stock index, this does not mean that the entire stock market is per se unsuitable for an active investment strategy. It is still conceivable that stocks of individual companies in the index can be predicted successfully. In this case, an active investment strategy based on the forecasts for individual stocks could be very promising. Second, forecasting future events with a six- to twelve-month horizon is a major hurdle. As the forecast quality tends to increase as the horizon decreases (Dua, 1988), it is conceivable that, for example, monthly forecasts for the same indices would lead to significantly better results. Last but not least, we analyze the entire time series from beginning to end for each forecaster. Even though this leads to a large sample size, which enables a clearer picture of the forecast quality overall, differences in the forecast quality over time may remain undetected. This could be the case in particular for the forecasts published in *Handelsblatt*, which extend over a period of 29 years.

Our results provide initial indications that patterns discovered almost 90 years ago that massively deteriorate forecast quality can still be found in stock market forecasts today. We therefore encourage future research efforts to examine whether our results prevail in additional datasets. Furthermore, we believe that deeper analysis of the rationale for conservative forecasting and an assessment of its financial impact on investors are promising areas of research that would deepen our understanding of ex-ante stock market forecasts.

6 Summary

We examine forecasts for the German Stock Market Index (DAX), the Dow Jones Industrial Index (DJI), and the Euro Stoxx 50 (SX5E) which were published in the period 1992 to 2020 in the German business newspaper *Handelsblatt* (HB) and the quality broadsheet the *Frankfurter Allgemeine Zeitung* (FAZ). These forecasts have a horizon of six and twelve months. The forecasts are from German and international banks such as Deutsche Bank, Goldman Sachs, J.P. Morgan, or BNP Paribas.

We take up the thoughts of Ogburn (1934), who, on the basis of a small empirical survey, became convinced that forecasters consistently underestimate the variability of the future, and that their forecasting is of a conservative nature. However, we also go beyond this and use some contemporary measures (prediction-realization diagram, test of unbiasedness, Diebold–Mariano test) to test ex-ante forecasts for their success at the time of validity.

Conservative forecasting behavior leads to unusual events being under-represented in forecasts, to the dispersion of the forecasts (as measured by their standard deviation) lagging behind the dispersion of the actual events, and to the extent of the forecasted changes being smaller than the actual changes. The latter aspect is reflected in a flat course of the regression line in the prediction-realization diagram (slope < 1) and thus also leads to failure in the unbiasedness test.

We analyze a total of 2,761 forecasts which are divided up into seven groups according to the subject of the forecast (DAX, DJI, SX5E), the forecast horizon (6 and 12 months), and the source (FAZ, HB). The findings are that in all seven groups (a) unusual events are under-represented in the forecasts, (b) the dispersion of the forecasts lags behind that of actual events, (c) the slope in the regression lines in the prediction-realization diagram is < 1, (d) the forecasts are biased to a highly significant degree, and (e) that the quality of the forecasts is not significantly better than that of naïve forecasts.

It is more than surprising how closely these stock market forecasts for the years 1992 to 2020 correspond to the characteristics which Ogburn described back in the 1930s. The stock market analysts prove to be too conservative, inflexible, and cautious. If they want to improve the reliability of their forecasts, they should change their conservative and inflexible forecasting behavior and consider promising new approaches and technologies in their forecasting process. For private and professional investors, building active investment strategies based on the insufficient stock market forecasts examined can involve enormous financial risks and is therefore not recommended.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/jrfm14120593/s1.

Author Contributions: Data curation, I.F., J.R.J., M.L. and M.S.; Formal analysis, I.F., J.R.J., M.L. and M.S.; Software, I.F.; Visualization, I.F., J.R.J. and M.L.; Writing—original draft, J.R.J., M.L. and M. S.; Writing—review & editing, I.F., J.R.J., M.L. and M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Our data basis can be accessed at Supplementary Materials.

Acknowledgments: The authors thank the editor and the anonymous reviewers for their constructive comments and useful suggestions, which were very helpful to enhance the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

Andres, P., & Spiwoks, M. (1999), Forecast Quality Matrix – A Methodological Survey of Judging Forecast Quality of Capital Market Forecasts, *Journal of Economics and Statistics*, 219(5-6), 513-542.

Arthur, W., Holland, J., LeBaron, B., Palmer, R., & Taylor, P. (1997), Asset pricing under endogenous expectations in an artificial stock market, in: Arthur, W., Durlauf, S., Lane, D. (eds.), The economy as an evolving complex system II, Addison-Wesley, Reading.

Atmaz, A., Cassella, S., Gulen, H., & Ruan, F. (2021), Contrarians, Extrapolators, and Stock Market Momentum and Reversal, Available at SSRN 3722540.

Atsalakis, G. S., & Valavanis, K. P. (2009), Forecasting stock market short-term trends using a neuro-fuzzy based methodology, *Expert systems with Applications*, 36(7), 10696-10707.

Bacchetta, P., Mertens, E., & van Wincoop, E. (2009), Predictability in financial markets: What do survey expectations tell us?, *Journal of International Money and Finance*, 28(3), 406-426.

Baghestani, H., Arzaghi, M., & Kaya, I. (2015), On the Accuracy of Blue Chip Forecasts of Interest Rates and Country Risk Premiums, *Applied Economics*, 47(2), 113-122.

Bahrami, A., Shamsuddin, A., & Uylangco, K. (2018), Out-of-sample stock return predictability in emerging markets, *Accounting & Finance*, 58(3), 727-750.

Ben-Rephael, A., Da, Z., & Israelsen, R. D. (2017), It Depends on Where You Search: Institutional Investor Attention and Underreaction to News, *Review of Financial Studies*, 30(9), 3009-3047.

Benke, H. (2006), Was leisten Kapitalmarktprognosen?, Die Sicht eines Stiftungsmanagers, *Zeitschrift für das gesamte Kreditwesen*, 59(17), 902-906.

Bertella, M. A., Pires, F. R., Feng, L., & Stanley, H. L. (2014), Confidence and the Stock Market: An Agent-Based Approach, *Plos One*, 9(1), 1-9.

Cassella, S., & Gulen, H. (2019), Belief-based Equity Market Sentiment, *Available at SSRN 3123083*.

Cassella, S., & Gulen, H. (2018), Extrapolation Bias and the Predictability of Stock Returns by Price-Scaled Variables, *The Review of Financial Studies,* 31(11), 4345-4397.

Chen, S. H., & Huang, Y. C. (2008), Risk preference, forecasting accuracy and survival dynamics: Simulations based on a multi-asset agent-based artificial stock market, *Journal of Economic Behavior & Organization*, 67(3-4), 702-717.

Chen, Y. T., & Vincent, K. (2016), The Role of Momentum, Sentiment, and Economic Fundamentals in Forecasting Bear Stock Market, *Journal of Forecasting*, 35(6), 504-527.

Cowles, A. (1933), Can stock market forecasters forecast?, *Econometrica: Journal of the Econometric Society*, 1(3), 309-324.

De Bondt, W. P. (1993), Betting on trends: Intuitive forecasts of financial risk and return, *International Journal of Forecasting*, 9(3), 355-371.

Diebold, F. X., & Mariano, R. S. (1995), Comparing Predictive Accuracy, *Journal of Business and Economic Statistics*, 13, 253-263.

Dimson, E., & Marsh, P. (1984), An analysis of brokers' and analysts' unpublished forecasts of UK stock returns, *The Journal of Finance*, 39(5), 1257-1292.

Dua, P. (1988), Multiperiod forecasts of interest rates, *Journal of Business & Economic Statistics*, 6(3), 381-384.

Fama, E. (1970), Efficient Capital Markets: A Review of Theory and Empirical Work, *The Journal of Finance*, 25(2), 383-417.

Fassas, A., Papadamou, S., & Kenourgios, D. (2021), Evaluating survey-based forecasts of interest rates and macroeconomic variables, *Journal of Economic Studies,* 49(1), 140-158.

Filiz, I., Judek, J. R., Lorenz, M., & Spiwoks, M. (2021), Interest rate forecasts in Latin America, *Journal of Economic Studies*, ahead-of-print. DOI: 10.1108/JES-02-2021-0108

Filiz, I., Nahmer, T., Spiwoks, M., & Bizer, K. (2019), The accuracy of interest rate forecasts in the Asia-Pacific region: opportunities for portfolio management. *Applied Economics*, 51, 6309-6332.

Fraser, P., & McDonald, R. (1993), The efficiency of CAC stock price forecasts: a survey based perspective, *Revue économique*, 44(5), 991-1000.

Friend, I., & Vickers, D. (1965), Portfolio selection and investment performance. *The Journal of Finance*, 20(3), 391-415.

Fujiwara, I., Ichiue, H., Nakazono, Y., & Shigemi, Y. (2013), Financial markets forecasts revisited: Are they rational, stubborn or jumpy?, *Economics Letters*, 118(3), 526-530.

Goyal, A., Welch, I., & Zafirov, A. (2021), A Comprehensive Look at the Empirical Performance of Equity Premium Prediction II, *Available at SSRN 3929119*.

Granger, C. W., & Newbold, P. (1974), Spurious regressions in econometrics, *Journal of Econometrics*, 2(2), 111-120.

Greenwood, R., & Shleifer, A. (2014), Expectations of Returns and Expected Returns, *The Review of Financial Studies*, 27(3), 714-746.

Guo, H. (2006), On the out-of-sample predictability of stock market returns, *The Journal of Business*, 79(2), 645-670.

Hein, O., Schwind, M., & Spiwoks, M. (2012), Network Centrality and Stock Market Volatility: The Impact of Communication Topologies on Prices, *Journal of Finance and Investment Analysis*, 1(1), 199-232.

Kalamara, E., Turrell, A. E., Redl, C. E., Kapetanios, G., & Kapadia, S. (2020), Making Text Count: Economic Forecasting Using Newspaper Text, *Bank of England Research Paper Series,* 865, 1-49.

Kandasamy, N., Garfinkel, S. N., Page, L., Hardy, B., Critchley, H. D., Gurnell, M., & Coates, J. M. (2016), Interoceptive Ability Predicts Survival on a London Trading Floor, *Scientific Reports*, 6, 32986.

Kazak, E., & Pohlmeier, W. (2019), Testing out-of-sample portfolio performance, *International Journal of Forecasting*, 35(2), 540-554.

Krichene, H., & El-Aroui, M.-A. (2018), Artificial stock markets with different maturity levels: simulation of information asymmetry and herd behavior using agent-based and network models, *Journal of Economic Interaction and Coordination*, 13(3), 511-535.

Kunze, F., Spiwoks, M., Bizer, K., & Windels, T. (2018), The usefulness of oil price forecasts – evidence from survey prediction, *Managerial and Decision Economics*, 39(4), 427-446.

Kunze, F., Wegener, C., Bizer, K., & Spiwoks, M. (2017), Forecasting European interest rates in times of financial crisis – What insights do we get from international survey forecasts?, *Journal of International Financial Markets, Institutions and Money*, 48, 192-205.

Lakonishok, J. (1980), Stock market return expectations: Some general properties, *The Journal of Finance*, 35(4), 921-931.

Lofthouse, S. (1996), Why Active Investment Management is Popular: The Psychology of Extraordinary Beliefs, *Journal of Interdisciplinary Economics*, 7(1), 41-61.

Mallikarjuna, M., & Rao, R. P. (2019), Evaluation of forecasting methods from selected stock market returns, *Financial Innovation*, 5, 1-16.

Mandelbrot, B. (2004), The (Mis)Behavior of Markets – A Fractal View of Risk, Ruin and Reward, Basic Books, 186-195.

Maxwell, M., & van Vuuren, G. (2019), Active investment strategies under tracking error constraints, *International Advances in Economic Research*, 25(3), 309-322.

Miah, F., Khalifa, A. A., & Hammoudeh, S. (2016), Further evidence on the rationality of interest rate expectations: A comprehensive study of developed and emerging economies, *Economic Modelling*, 54, 574-590.

Milas, C., Panagiotidis, T., & Dergiades, T. (2021), Does It Matter Where You Search? Twitter versus Traditional News Media, *Journal of Money, Credit and Banking*, 153(7), 1757-1795.

Mincer, J., & Zarnowitz, V. (1969), The Evaluation of Economic Forecasts, in: Mincer, J. (Ed.), Economic Forecasts and Expectation, Columbia University Press, New York, 3-46.

Neely, C. J., Rapach, D. E., Tu, J., & Zhou, G. (2014), Forecasting the equity risk premium: the role of technical indicators, *Management Science*, 60(7), 1772-1791.

Nyberg, H. (2013), Predicting bear and bull stock markets with dynamic binary time series models, *Journal of Banking & Finance*, *37*(9), 3351-3363.

Ogburn, W. F. (1934), Studies in Prediction and the Distortion of Reality, *Social Forces*, 13, 224-229.

Oliver, N., & Pasaogullari, M. (2015), Interest Rate Forecasts in Conventional and Unconventional Monetary Policy Periods, *Economic commentary*, 5, 1-4.

Ortiz-Teran, E., Ortiz, T., Turrero, A., & Lopez-Pascual, J. (2019), Neural implications of investment banking experience in decision-making under risk and ambiguity, *Journal of Neuroscience, Psychology, and Economics*, 12(1), 34-44.

Pierdzioch, C. (2015), A note on the directional accuracy of interest-rate forecasts, *Applied Economics Letters*, 22(13), 1073-1077.

Ponta, L., & Cincotti, S. (2018), Traders' Networks of Interactions and Structural Properties of Financial Markets: An Agent-Based Approach, *Hindawi Complexity*, 2018(4), 1-9.

Rajab, S., & Sharma, V. (2019), An interpretable neuro-fuzzy approach to stock price forecasting, *Soft Computing*, 23(3), 921-936.

Ramnath, S., Rock, S., & Shane, P. (2008), The Financial Analyst Forecasting Literature: A Taxonomy with Suggestions for Further Research, *International Journal of Forecasting*, 24(1), 34-75.

Spiwoks, M. (2004), The Usefulness of ZEW Stock Market Forecasts for Active Portfolio Management Strategies, *Journal of Economics and Statistics*, 224(5), 557-578.

Spiwoks, M., Gubaydullina, Z., & Hein, O. (2015), Trapped in the Here and Now - New Insights into Financial Market Analyst Behavior, *Journal of Applied Finance and Banking*, 5(1), 29-50.

Spiwoks, M., & Hein, O. (2007), Die Währungs-, Anleihen- und Aktienmarktprognosen des Zentrums für Europäische Wirtschaftsforschung, *AStA Wirtschafts- und Sozialstatistisches Archiv*, 1(1), 43-52.

Theil, H. (1958), Economic Forecasts and Policy, North Holland Publishing Company, Amsterdam.

Theissen, E. (2007), An analysis of private investors' stock market return forecasts, *Applied Financial Economics*, 17(1), 35-43.

Wald, A. (1943), Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large, *Transactions of the American Mathematical Society*, 54(3), 426-482.

Welch, I., & Goyal, A. (2008), A comprehensive look at the empirical performance of equity premium prediction, *The Review of Financial Studies*, 21(4), 1455-1508.

Werner, N. S., Jung, K., Duschek, S., & Schandry, R. (2009), Enhanced cardiac perception is associated with benefits in decision-making, *Psychophysiology*, 46(6), 1123-1129.

Wooldridge, J. M. (2002), Econometric Analysis of Cross Section and Panel Data, Cambridge (MA), MIT Press.

Yang, X., Zhang, J., & Ye, Q. (2020), Tick size and market quality: Simulations based on agent-based artificial stock markets, *Intelligent Systems in Accounting, Finance and Management*, 27(3), 125-141.

Appendix A: Detailed Summary Statistics on Data Basis

Table A-1: Detailed Summary Statistics on DAX, DJI and SX5E forecasts of our data basis

Source	Subject	Year	N	Min (pts.)	Max (pts.)	Median (pts.)	Mean (pts.)	Actual (pts.)	N	Min (pts.)	Max (pts.)	Median (pts.)	Mean (pts.)	Actual (pts.)
			Forecast horizon 6 months						**Forecast horizon 12 months**					
HB	DAX	1992	NA	NA	NA	NA	NA	NA	21	1,600	1,900	1,780	1,764	1,545.05
		1993	NA	NA	NA	NA	NA	NA	25	1,550	1,900	1,750	1,726	2,266.68
		1994	NA	NA	NA	NA	NA	NA	28	1,840	2,500	2,400	2,339	2,106.58
		1995	NA	NA	NA	NA	NA	NA	33	1,950	2,500	2,200	2,225	2,253.88
		1996	NA	NA	NA	NA	NA	NA	28	2,250	2,700	2,450	2,449	2,888.69
		1997	NA	NA	NA	NA	NA	NA	34	2,600	3,800	3,100	3,095	4,249.69
		1998	NA	NA	NA	NA	NA	NA	33	4,000	4,800	4,413	4,413	5,002.39
		1999	NA	NA	NA	NA	NA	NA	34	4,580	6,000	5,400	5,390	6,958.14
		2000	NA	NA	NA	NA	NA	NA	37	6,200	7,620	6,790	6,771	6,433.61
		2001	NA	NA	NA	NA	NA	NA	33	6,100	9,000	7,800	7,722	5,160.10
		2002	NA	NA	NA	NA	NA	NA	38	5,100	6,650	5,750	5,779	2,892.63
		2003	NA	NA	NA	NA	NA	NA	33	3,300	5,000	3,915	3,921	3,965.16
		2004	NA	NA	NA	NA	NA	NA	34	3,500	5,000	4,300	4,318	4,256.08
		2005	NA	NA	NA	NA	NA	NA	33	4,100	5,000	4,600	4,558	5,408.26
		2006	NA	NA	NA	NA	NA	NA	38	5,000	6,100	5,800	5,717	6,596.92
		2007	NA	NA	NA	NA	NA	NA	37	6,000	7,500	7,078	7,027	8,067.32
		2008	NA	NA	NA	NA	NA	NA	35	7,700	9,250	8,500	8,566	4,810.20
		2009	NA	NA	NA	NA	NA	NA	31	3,600	6,500	5,250	5,230	5,957.43
		2010	NA	NA	NA	NA	NA	NA	38	4,500	7,500	6,345	6,339	6,914.19
		2011	NA	NA	NA	NA	NA	NA	39	6,200	8,300	7,600	7,605	5,898.35
		2012	NA	NA	NA	NA	NA	NA	37	5,500	7,600	6,573	6,573	7,612.39
		2013	NA	NA	NA	NA	NA	NA	35	6,900	8,890	8,029	8,024	9,552.16
		2014	NA	NA	NA	NA	NA	NA	33	8,900	11,000	10,200	10,123	9,805.55
		2015	NA	NA	NA	NA	NA	NA	36	9,500	11,800	10,753	10,706	10,743.01
		2016	NA	NA	NA	NA	NA	NA	36	9,250	13,000	11,850	11,793	11,481.06
		2017	NA	NA	NA	NA	NA	NA	30	11,000	12,300	11,800	11,724	12,917.64
		2018	NA	NA	NA	NA	NA	NA	33	12,300	15,000	14,000	14,009	10,558.96
		2019	NA	NA	NA	NA	NA	NA	31	10,000	13,400	12,000	12,053	13,249.01
		2020	NA	NA	NA	NA	NA	NA	31	12,500	15,000	14,000	13,999	13,718.78
			Forecast horizon 6 months						**Forecast horizon 12 months**					
FAZ	DAX	2002	14	4,900	6,000	5,650	5,554	4,382.56	19	5,100	6,650	5,750	5,808	2,892.63
		2003	NA	NA	NA	NA	NA	3,220.58	17	3,000	4,200	3,800	3,780	3,965.16
		2004	14	3,600	4,500	4,200	4,184	4,052.73	15	3,833	4,700	4,300	4,299	4,256.08
		2005	15	3,900	4,600	4,400	4,330	4,586.28	21	4,100	4,750	4,570	4,560	5,408.26
		2006	17	5,000	5,950	5,700	5,616	5,683.31	20	5,100	6,100	5,725	5,689	6,596.92
		2007	14	6,200	7,100	6,612	6,623	8,007.32	20	6,000	7,400	7,000	6,988	8,067.32
		2008	14	7,250	8,700	8,066	8,081	6,418.32	18	7,700	9,200	8,500	8,503	4,810.20
		2009	17	3,200	5,700	4,900	4,725	4,808.64	17	3,600	6,500	5,400	5,353	5,957.43
		2010	19	4,800	6,800	6,000	5,875	5,965.52	22	5,300	7,100	6,375	6,333	6,914.19
		2011	19	6,300	8,000	7,300	7,289	7,376.24	26	6,200	8,300	7,600	7,618	5,898.35
		2012	14	4,800	7,000	6,105	6,009	6,416.28	22	5,500	7,600	6,594	6,588	7,612.39
		2013	14	7,000	8,200	7,659	7,618	7,959.22	20	7,250	8,890	8,035	8,069	9,552.16
		2014	16	8,500	10,200	9,660	9,620	9,833.07	23	8,900	11,000	10,150	10,092	9,805.55
		2015	18	8,700	11,000	10,300	10,035	10,944.97	23	9,500	11,500	10,900	10,773	10,743.01
		2016	17	10,200	12,250	11,400	11,388	9,680.09	23	10,800	12,600	11,900	11,859	11,481.06
		2017	19	10,600	12,400	11,500	11,494	12,325.12	24	10,400	12,300	11,800	11,713	12,917.64
		2018	19	12,500	15,000	13,700	13,658	12,306.00	25	12,300	14,500	14,000	13,938	10,558.96
		2019	NA	NA	NA	NA	NA	12,398.80	24	10,000	13,400	12,000	11,986	13,249.01
		2020	22	12,000	14,500	13,625	13,460	12,310.93	23	12,500	14,500	14,000	13,833	13,718.78

HB = Handelsblatt; FAZ = Frankfurter Allgemeine Zeitung; DAX = German Stock Market Index; N = Number of forecasts issued; Min = Minimum; Max = Maximum; pts. = points; NA = not available.

Continued Appendix A, Table A-1:

Source	Subject	Year	N	Min (pts.)	Max (pts.)	Median (pts.)	Mean (pts.)	Actual (pts.)	N	Min (pts.)	Max (pts.)	Median (pts.)	Mean (pts.)	Actual (pts.)
			Forecast horizon 6 months						*Forecast horizon 12 months*					
FAZ	DJI	2004	10	9,800	11,000	10,422	10,444	10,435.48	10	10,000	11,200	10,500	10,544	10,783.01
		2005	10	10,800	11,200	11,010	11,020	10,274.97	14	11,000	12,000	11,420	11,440	10,717.50
		2006	14	10,000	11,800	11,223	11,196	11,150.22	15	10,300	12,500	11,500	11,575	12,463.15
		2007	12	12,200	14,000	12,800	12,805	13,408.62	14	11,440	14,000	13,400	13,276	13,264.82
		2008	13	12,500	14,500	13,729	13,729	11,350.01	16	13,500	15,300	14,500	14,513	8,776.39
		2009	14	6,900	10,800	9,000	9,000	8,447.00	16	7,000	12,500	9,940	9,880	10,428.05
		2010	16	8,900	12,100	10,600	10,433	9,774.02	18	10,000	12,100	11,050	11,118	11,577.51
		2011	14	10,500	13,900	11,904	11,808	12,414.34	16	10,200	13,500	12,064	12,127	12,217.56
		2012	9	10,800	13,500	12,363	12,363	12,880.09	13	12,375	15,000	13,200	13,240	13,104.14
		2013	8	12,100	14,000	13,487	13,381	14,909.60	11	13,000	15,300	14,150	14,150	16,576.66
		2014	12	14,500	16,800	16,500	16,364	16,826.60	14	15,700	17,700	17,000	16,908	17,823.07
		2015	14	14,000	18,800	18,000	17,586	17,619.51	17	16,000	19,400	18,547	18,547	17,425.03
		2016	12	17,500	19,000	18,123	18,245	17,929.99	15	17,000	19,500	18,700	18,568	19,762.60
		2017	16	18,700	21,900	19,949	19,897	21,349.63	17	18,200	21,200	20,103	20,103	24,719.22
		2018	14	22,000	27,200	24,825	24,735	24,271.41	18	22,000	28,500	25,208	25,215	23,327.46
		2019	NA	NA	NA	NA	NA	26,599.96	18	24,000	28,000	26,250	24,782	28,538.44
		2020	15	27,250	29,200	28,500	28,404	25,812.88	17	27,100	30,400	28,909	28,909	30,606.48
			Forecast horizon 6 months						*Forecast horizon 12 months*					
FAZ	SX5E	2002	14	3,600	4,300	4,062	4,023	3,133.39	17	3,710	4,600	4,300	4,251	2,386.41
		2003	NA	NA	NA	NA	NA	2,419.51	15	2,300	3,200	2,900	2,890	2,760.66
		2004	13	2,500	3,300	2,879	2,879	2,811.08	14	2,750	3,300	3,004	3,008	2,951.01
		2005	15	2,800	3,200	3,050	3,030	3,181.54	19	3,000	3,350	3,200	3,160	3,578.93
		2006	17	3,350	3,800	3,700	3,671	3,648.92	18	3,450	3,950	3,777	3,754	4,119.94
		2007	14	4,000	4,750	4,208	4,215	4,489.77	20	3,700	4,600	4,400	4,394	4,399.72
		2008	14	4,200	4,900	4,508	4,515	3,352.81	18	4,400	5,100	4,700	4,726	2,447.62
		2009	15	1,600	3,000	2,500	2,469	2,401.69	17	1,950	3,350	2,756	2,756	2,964.96
		2010	17	2,400	3,300	2,910	2,896	2,573.32	20	2,600	3,700	3,100	3,124	2,792.82
		2011	17	2,400	3,400	2,950	2,905	2,848.53	22	2,500	3,350	3,009	3,018	2,316.55
		2012	14	1,700	2,600	2,300	2,279	2,264.72	22	2,050	2,850	2,505	2,510	2,635.93
		2013	15	2,162	2,800	2,626	2,626	2,602.59	20	2,590	3,050	2,799	2,797	3,109.00
		2014	15	2,750	3,400	3,250	3,208	3,228.25	23	3,000	3,600	3,400	3,344	3,146.43
		2015	17	2,800	3,550	3,300	3,245	3,424.30	22	3,200	3,720	3,444	3,438	3,267.52
		2016	16	3,145	3,750	3,550	3,543	2,864.74	22	3,425	3,800	3,683	3,665	3,290.52
		2017	18	3,000	3,500	3,271	3,261	3,441.88	23	3,100	3,500	3,300	3,295	3,503.96
		2018	18	3,450	4,050	3,748	3,746	3,395.60	23	3,400	4,000	3,800	3,793	3,001.42
		2019	NA	NA	NA	NA	NA	3,473.69	23	2,800	3,700	3,300	3,305	3,745.16
		2020	21	3,400	4,000	3,713	3,713	3,234.07	23	3,500	4,050	3,850	3,833	3,552.64

FAZ = Frankfurter Allgemeine Zeitung; DJI = Dow Jones Industrial Index; SX5E = Euro Stoxx 50; N = Number of forecasts issued; Min = Minimum; Max = Maximum; pts. = points; NA = not available.

Appendix B: Forecasters in the *Handelsblatt* newspaper

1. ABN Amro
2. Adca-Bank
3. B. Metzler Seel. Sohn & Co.
4. Baader Bank
5. Baden-Württembergische Bank
6. Bank in Liechtenstein
7. Bank Julius Bär
8. Bank of America
9. Bank Sarasin
10. Bankhaus Ellwanger & Geiger
11. Bankhaus Lampe
12. Bankhaus Metzler
13. Banque Nationale de Paris
14. Barclays
15. Bayerische Landesbank
16. Bayerische Vereinsbank
17. Berenberg
18. Bethmann Bank
19. BNP Paribas
20. Cheuvreux
21. Citi
22. Commerzbank
23. Crédit Lyonnais
24. Credit Suisse
25. Daiwa Europe (Deutschland)
26. Dekabank
27. Deutsche Bank
28. Donner & Reuschel
29. Dresdner Bank
30. DZ Bank
31. Fürst Fugger Privatbank
32. Fürstl. Castell'sche Bank
33. Goldman Sachs
34. Gontard & Metallbank
35. GZ-Bank
36. Haspa
37. Hauck & Aufhäuser
38. Helaba
39. HSBC Trinkaus
40. HSH Nordbank
41. IKB
42. IMI Bank
43. J. Safra Sarasin
44. J.P. Morgan
45. Kepler Equities
46. Kleinwort Benson Research
47. LB Rheinland-Pfalz
48. LBB Landesbank Berlin
49. LBBW
50. Lehman Brothers
51. LGT Bank in Liechtenstein
52. M.M. Warburg & Co.
53. Macquarie
54. Merck Finck & Co.
55. Merrill Lynch
56. Morgan Stanley
57. National-Bank
58. NATIXIS
59. NIBC
60. Nomura
61. NordLB
62. Oddo BHF
63. Pictet & Cie.
64. Postbank
65. Royal Bank of Scotland
66. S.G. Warburg
67. Sal. Oppenheim
68. Santander
69. Saxo Bank
70. SBC Warburg
71. Schröder Bank
72. Schröder Münchmeyer Hengst
73. Schroder Salomon Smith Barney
74. Schweizerischer Bankverein
75. SGZ-Bank
76. Société Générale
77. SYZ & Co.
78. Targobank
79. UBS
80. Unicredit HypoVereinsbank
81. Union Bancaire Priveé
82. Union Bank of Switzerland
83. Vereins- und Westbank
84. Vontobel
85. VP Bank
86. Weberbank
87. WestLB
88. WGZ Bank

Appendix C: Forecasters in the *Frankfurter Allgemeine Zeitung*

1. Adig
2. Allianz SE
3. Bankgesellschaft Berlin
4. Bankhaus Lampe
5. Barclays Capital
6. Bayern LB
7. Berenberg
8. BNP Paribas
9. Citigroup
10. Commerzbank
11. CSFB
12. Deka Bank
13. Deutsche Bank
14. Deutsche Bank/Postbank
15. DIT
16. Dresdner Bank
17. DWS
18. DZ Bank
19. Erste Group
20. Goldman Sachs
21. Helaba
22. HSBC Trinkaus & Burkhardt
23. HSH Nordbank
24. HVB-Unicredit Bank
25. IKB
26. ING Deutschland
27. J.P. Morgan
28. Julius Bär
29. Landesbank Berlin
30. Landesbank Rheinland-Pfalz
31. LBBW
32. M.M. Warburg
33. Macquarie
34. Merck Finck Invest
35. Merrill Lynch
36. Morgan Stanley
37. Nomura
38. Nord LB
39. Oddo BHF
40. Postbank
41. Raiffeisen Bank International
42. Sal. Oppenheim
43. Santander Asset Management
44. Société Générale
45. UBS
46. Union Bancaire Privée
47. Union Investment
48. Vereins- und Westbank
49. Weberbank
50. WestLB
51. WGZ Bank

Declaration of contribution to each essay of the cumulative dissertation

I contributed to the seven essays of the cumulative dissertation as follows:

1. **Reducing Algorithm Aversion through Experience**
   co-authored by Ibrahim Filiz, Marco Lorenz, and Markus Spiwoks

   45% - Conceptualization, Resources, Methodology, Software, Data curation, Formal analysis, Investigation, Validation, Writing (Original draft), Visualization.

2. **The Extent of Algorithm Aversion in Decision-making Situations with Varying Gravity**
   co-authored by Ibrahim Filiz, Marco Lorenz, and Markus Spiwoks

   45% - Conceptualization, Resources, Software, Data curation, Formal analysis, Validation, Writing (Original draft), Writing (Review & Editing), Visualization.

3. **Comparing Different Kinds of Influence on an Algorithm in Its Forecasting Process and Their Impact on Algorithm Aversion**
   co-authored by Zulia Gubaydullina, Marco Lorenz, and Markus Spiwoks

   45% - Conceptualization, Resources, Software, Data curation, Formal analysis, Validation, Writing (Original draft), Writing (Review & Editing), Visualization.

4. **Algorithm Aversion as an Obstacle in the Establishment of Robo Advisors**
   co-authored by Ibrahim Filiz, Marco Lorenz, and Markus Spiwoks

   45% - Resources, Methodology, Software, Data curation, Formal analysis, Validation, Writing (Original draft), Writing (Review & Editing), Visualization.

5. **Willingness to Use Algorithms Varies with Social Information on Weak vs. Strong Adoption: An Experimental Study on Algorithm Aversion**

   100% - Conceptualization, Resources, Methodology, Software, Data curation, Formal analysis, Investigation, Validation, Writing (Original draft), Visualization.

6. **Interest Rate Forecasts in Latin America**
   co-authored by Ibrahim Filiz, Marco Lorenz, and Markus Spiwoks

   45% - Conceptualization, Resources, Methodology, Data collection, Data preparation, Formal analysis, Writing (Original draft), Writing (Review & Editing), Visualization.

7. **Sticky Stock Market Analysts**
   co-authored by Ibrahim Filiz, Marco Lorenz, and Markus Spiwoks

   45% - Conceptualization, Resources, Methodology, Data collection, Data preparation, Formal analysis, Writing (Original draft), Writing (Review & Editing), Visualization.