

Lehmann, Cedric A.; Haubitz, Christiane B.; Fügener, Andreas; Thonemann, Ulrich W.

Article — Published Version

The risk of algorithm transparency: How algorithm complexity drives the effects on the use of advice

Production and Operations Management

Provided in Cooperation with:

John Wiley & Sons

Suggested Citation: Lehmann, Cedric A.; Haubitz, Christiane B.; Fügener, Andreas; Thonemann, Ulrich W. (2022) : The risk of algorithm transparency: How algorithm complexity drives the effects on the use of advice, Production and Operations Management, ISSN 1937-5956, Wiley, Hoboken, NJ, Vol. 31, Iss. 9, pp. 3419-3434, <https://doi.org/10.1111/poms.13770>

This Version is available at:

<https://hdl.handle.net/10419/264995>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



<http://creativecommons.org/licenses/by-nc-nd/4.0/>

ORIGINAL ARTICLE

The risk of algorithm transparency: How algorithm complexity drives the effects on the use of advice

Cedric A. Lehmann  | Christiane B. Haubitz  | Andreas Fügener  |
Ulrich W. Thonemann 

Department of Supply Chain Management and Management Science, University of Cologne, Cologne, Germany

Correspondence

Ulrich W. Thonemann, Department of Supply Chain Management and Management Science, University of Cologne, 50923 Cologne, Germany.
Email: ulrich.thonemann@uni-koeln.de

Handling Editor: Elena Katok

Funding information

Deutsche Forschungsgemeinschaft, Grant/Award Number: EXC 2126/1 - 390838866

Abstract

Although algorithmic decision support is omnipresent in many managerial tasks, a lack of algorithm transparency is often stated as a barrier to successful human–machine collaboration. In this paper, we analyze the effects of algorithm transparency on the use of advice from algorithms with different degrees of complexity. We conduct a set of laboratory experiments in which participants receive identical advice from algorithms with different levels of transparency and complexity. Our results indicate that not the algorithm itself, but the individually perceived appropriateness of algorithmic complexity moderates the effects of transparency on the use of advice. We summarize this effect as a plateau curve: While perceiving an algorithm as too simple severely harms the use of its advice, the perception of an algorithm as being too complex has no significant effect. Our insights suggest that managers do not have to be concerned about revealing algorithms that are perceived to be appropriately complex or too complex to decision-makers, even if the decision-makers do not fully comprehend them. However, providing transparency on algorithms that are perceived to be simpler than appropriate could disappoint people's expectations and thereby reduce the use of their advice.

KEYWORDS

algorithm transparency, algorithm complexity, decision-making, decision support, use of advice

1 | INTRODUCTION

In recent decades, advances in data availability and computational power have increased the use of algorithms for day-to-day decision-making. People increasingly use complex “black-box” algorithms that reveal little about their underlying principles and often only provide a final recommendation. In various decision-making domains, complex algorithms, such as machine learning models, have been successfully employed to improve the quality of decisions (Brynjolfsson & McAfee, 2014; LeCun et al., 2015).

However, there is also a belief that the black-box nature of such algorithms results in a low acceptance of their advice (Burton et al., 2020), an issue that can be addressed by

explaining the underlying algorithmic principles (Glikson & Woolley, 2020). This observation motivated a project involving the service division of a large equipment manufacturer for whom we developed a spare parts inventory optimization tool. The company decided to make the algorithmic principles transparent to users because they expected greater adherence to the recommendations of the tool with greater transparency. The algorithm is now used by planners in more than 80 country organizations. Feedback has indicated that planners appreciate transparency, but whether the positive perception of transparency actually increases the use of advice compared with a less transparent tool remains an open question.

To explore this issue, we asked 450 undergraduate students in an operations management class about their attitudes toward the use of algorithms in managerial decisions. The vast majority (94%) indicated that they would like to receive

Accepted by Elena Katok, after two revisions.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs License](https://creativecommons.org/licenses/by-nc-nd/4.0/), which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Production and Operations Management* published by Wiley Periodicals LLC on behalf of Production and Operations Management Society.

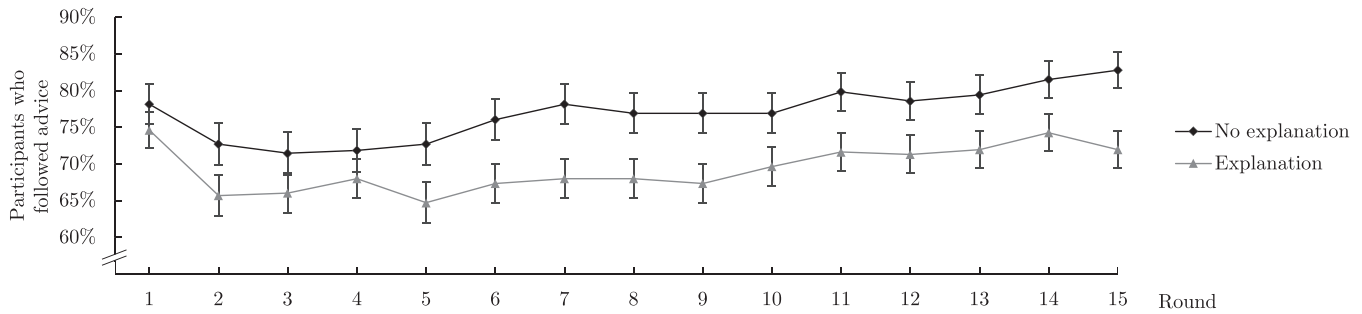


FIGURE 1 Share of participants who followed the algorithmic advice by round (randomized sequence of spare parts sets). Error bars indicate standard errors

algorithmic advice when facing a managerial task. Approximately as many (94%) wanted to be informed about the algorithmic procedures. Although the questions were hypothetical, the results indicate a preference for transparent algorithmic advice.

In the same class, we conducted an experiment related to the inventory management problem faced by the company that motivated our research. For 15 different sets of spare parts, participants had to decide on the parts to stock, considering part demand rates, part weights, and a total weight limit—a classical combinatorial optimization task known as the “knapsack-problem” (Dantzig, 1957). The problem can be solved by a greedy algorithm, which sorts the spare parts by their value-to-weight ratio and selects the part with the highest ratio as long as the capacity is not exceeded. Although this procedure does not necessarily optimally solve the problem, it usually produces good results (Diubin & Korbut, 2008). In our experiment, the solutions of the greedy algorithm were optimal. We divided the students into two groups. Both groups received the same algorithmic advice, but one group additionally received an explanation of the algorithmic principles, whereas the other did not. The group that received the explanation followed the algorithmic advice less often than the group that did not receive an explanation (Figure 1). We provide further details on the experiment in the [Supporting Information](#) of this paper.

The results are somewhat surprising. The participants reported that they would like to receive algorithmic advice and that they would appreciate transparency about the algorithm. However, when they were provided with both for the relatively simple algorithm that we used, they followed its advice less often with transparency than without. These observations raise questions regarding the value of transparency: Does greater transparency generally reduce the use of advice, or does the relationship depend on factors such as the complexity of an algorithm? In other words, is complexity a moderator for the effect of algorithm transparency on the use of advice? This issue has received little attention in the literature, and we address it in this paper.

We analyze the effects of the transparency of algorithms with different complexities on the use of advice in managerial decision tasks. We conduct a preregistered laboratory experiment in which we manipulate transparency by informing only

a subgroup of participants about the underlying principles of different advice-giving algorithms. To test whether the transparency of simple algorithms has a different effect than the transparency of more complex algorithms, we also vary the degree of algorithm complexity. We provide explanations of two different algorithms that vary in their underlying principles but generate identical advice. The results indicate that the effect of transparency on the use of algorithmic advice does not depend on the objective complexity of the algorithm but on the perceived appropriateness of its complexity. Our results can be summarized by a plateau curve: Providing transparency on algorithms that are perceived to be too complex for a task does not reduce the use of advice. However, providing transparency on algorithms that are perceived to be too simple for the task does reduce the use of advice. This plateau curve is supported by the results of a second experiment, in which we changed the complex algorithm and increased the task information available to the participants. Our results provide a better understanding of the benefits and risks of algorithm transparency and indicate when managers should make algorithms transparent and when not.

2 | LITERATURE REVIEW

Decision-makers appreciate receiving advice to improve the quality of decisions and to share the responsibility for possible negative consequences (Harvey & Fischer, 1997; Yaniv, 2004a, 2004b). Research on advice taking and decision-making is broad, and we refer the reader to Bonaccio and Dalal (2006) for a comprehensive review. In this context, we focus on drivers of the use of advice and specifically on the role of algorithm transparency. We then review advice taking in demand forecasting, a critical managerial task that we consider in our experiments.

2.1 | The use of algorithmic advice

The advice-taking literature suggests that receiving advice often improves decisions (Sniezek et al., 2004; Yaniv, 2004a). However, decision-makers tend not to follow advice as much as they would be in their interest (Bonaccio & Dalal, 2006)

and place insufficient weight on the advice (Dietvorst et al., 2015; Harvey & Fischer, 1997; Logg et al., 2019; Yaniv, 2004a, 2004b). In a cue-learning task, Harvey and Fischer (1997) observe a weight of only 20–30% on human advice, even if the advisor has more expertise than the decision-maker. In different estimation tasks, Logg et al. (2019) observe a weight of 34–52% on algorithmic advice.

Yu et al. (2019) suggest that people use algorithmic advice based on their perceived performance. Even without revealed performance measures, humans are capable of detecting the quality of algorithmic advice through outcomes and adapting their use accordingly (De Baets & Harvey, 2020). Dietvorst et al. (2015) observe that people lose confidence in algorithms over time when they see algorithms err. If similar advice is provided by a human, confidence in their advice does not fade as quickly. This so-called “algorithm aversion” is supported by Prahla and Van Swol (2017), and its effect even increases if tasks are perceived as subjective (Castelo et al., 2019).

Different beliefs exist about whether the black-box nature of algorithms is harmful for advice utilization. Some studies indicate that people are not opposed to taking advice from algorithms without having much information on their underlying principles. For example, Logg et al. (2019) show that people are generally willing to work with black-box algorithms. Dietvorst et al. (2015) find no contrary evidence because participants in one of their studies preferred to receive advice from a black-box algorithm over taking advice from a human when no direct feedback was given. However, Burton et al. (2020) argue that opaque algorithms can harm successful human–algorithm interactions. Glikson and Woolley (2020) suggest that this could be addressed by providing transparency on the algorithmic principles, which can increase the trust in algorithms.

Algorithm transparency describes the presence of information on an algorithm’s structure, procedure, decision rationale, or performance. This information must be presented in a way that users can grasp coherences, leading to a higher understanding (Seymour, 2018). Algorithm transparency can be increased by revealing the input data, explaining the underlying procedure, or tracking and displaying the performance of the algorithm. In this study, we focus on procedural transparency and manipulate whether the underlying principles of an algorithm are visible to the decision-maker. A high level of procedural transparency typically explains the procedures that transform data into advice (Bertino et al., 2019). Therefore, a high level of procedural transparency also reveals the complexity of the applied principles. Previous research indicates that greater algorithmic complexity might be associated with stronger trust in an algorithm (Glikson & Woolley, 2020).

Several studies analyze the effect of different explanations on the perceived understanding of algorithms. Cai et al. (2019) conduct studies in which they increase transparency through normative and comparative explanations. They find that normative explanations increase the understanding and capability perceptions of the system, whereas comparative

explanations reveal the algorithm’s boundaries and limitations. Lu et al. (2020) show that both verbal explanations and examples have a positive effect on perceived understanding and satisfaction. However, only the presentation of both general and case-specific examples increases the perceived intuitiveness, fairness, and impact of the algorithm.

The literature discusses the effects of algorithm transparency on the use of algorithmic advice. Previous research has focused on subjective tasks that involve personal taste. In an early study, Sinha and Swearingen (2002) conclude that people appreciate music recommender systems that they perceive as transparent and exhibit higher degrees of confidence in their advice. Similarly, Wang and Benbasat (2007) show that explaining how an e-commerce recommender system derived a product recommendation increased users’ trust in the technical competence of the recommender agent. Cramer et al. (2008) analyze the effects of the transparency of an art recommender system. While an explanation of why certain artwork is recommended increased the acceptance of the recommendations, it did not affect the general trust in the recommender system. Kizilcec (2016) examines the effects of the transparency of a grading algorithm. His results show that students with violated grade expectations had lower trust in the algorithm than students whose expectations were met. Furthermore, he finds that revealing excessive information can have a negative effect on trust in an algorithm. Springer and Whittaker (2018) analyze the effects of the transparency of an algorithm that predicts a user’s mood based on a short self-written text about a past emotional experience. Increased transparency led to reduced perceived accuracy, even when expectations were met. The study of Lehmann et al. (2020) produced similar results. They find that transparency through explanation increases understanding of the algorithmic procedures but reduces the perceived value of advice.

The results of these studies point in different directions. Although prior research indicates that transparency has a positive effect on trust (Wang & Benbasat, 2007), appreciation, and confidence (Sinha & Swearingen, 2002), subsequent studies reveal mixed (Cramer et al., 2008; Kizilcec, 2016) or even negative effects (Lehmann et al., 2020; Springer & Whittaker, 2018) on trust or perceived accuracy. Studies have been conducted with subjective tasks, such as product recommendations or grading. For objective tasks, the effects of algorithm transparency on the use of algorithmic advice have received little attention. Moreover, most studies have measured the perception of algorithmic advice with indicators such as trust, perceived accuracy, or algorithm appreciation without linking it to the actual use of advice. Analyzing an objective task for which advice is presented in the form of a point estimate would, however, allow for a more precise measurement of the use of advice. In addition, previous studies have not considered the role of algorithm complexity. This study addresses these gaps and analyzes the effect of algorithm transparency on the use of algorithmic advice for an objective managerial task, that is, demand forecasting.

2.2 | Advice taking in demand forecasting

Demand forecasting is a critical task for organizations and a commonly analyzed task in behavioral operations management (Kremer et al., 2011; Lawrence et al., 2006). Forecasting tasks have also been used to analyze the general human attitude toward algorithms (Dietvorst et al., 2015; Prahla & Van Swol, 2017; Logg et al., 2019).

Forecasting algorithms have been continuously improved (Fildes, 2006) and are increasingly used in practice (Fildes & Petropoulos, 2015). Nevertheless, most companies do not entirely rely on algorithmic forecasts but include human judgment in their forecasting routines (e.g., Fildes & Goodwin, 2007; Fildes & Petropoulos, 2015; Khosrowabadi et al., 2022; Perera et al., 2019). This approach is referred to as judgmental forecasting.

Research on judgmental forecasting analyzes how to efficiently combine human judgment and algorithms (Arvan et al., 2019; Perera et al., 2019; Webby & O'Connor, 1996). Although human forecasters can be inconsistent, prone to biases, and affected by wrong incentives (Fildes & Goodwin, 2007; Fildes et al., 2009; Franses & Legerstee, 2011; Perera et al., 2019), forecasting algorithms are based on a fixed set of rules that might not be adaptable to specific situations. The most frequently applied are judgmental adjustments of algorithmic advice (Arvan et al., 2019; Perera et al., 2019), in which statistically generated forecasts are manually adjusted by the forecaster, for example, to include intuition, additional information, and expertise (e.g., Fildes & Goodwin, 2007; Franses & Legerstee, 2011; Webby & O'Connor, 1996). Depending on the characteristics of the decision-maker, situation, quality of the algorithm, and information available, these adjustments might improve or impair forecast accuracy (e.g., Fildes et al., 2009; Franses & Legerstee, 2011; Fildes et al., 2019; Khosrowabadi et al., 2022).

De Baets and Harvey (2020) find that people rely more on good rather than on poor forecasting models but not as much as they should. This finding is in line with the results of Goodwin and Fildes (1999), who argue that statistical forecasts are often underutilized, despite their quality and appropriateness. This phenomenon has been attributed to general distrust toward forecasting advice (Goodwin et al., 2013) and algorithm aversion (Dietvorst et al., 2015). Although some indication exists that a “lack of transparency in [...] underlying processes can contribute to a forecaster’s distrust in systems” (Arvan et al., 2019), little is known about the effect of algorithm transparency on the use of algorithmic advice or the role of algorithm complexity in this relationship. We address these issues in this study.

3 | THEORY DEVELOPMENT

The effects of algorithm transparency on human decisions depend on the expectations of the decision-maker (Burton et al., 2020; Kizilcec, 2016; Springer & Whittaker, 2019). Logg et al. (2019) surveyed the expectations of human

decision-makers for the approaches used by algorithms (a survey of 226 respondents using thematic coding). A plurality, 42% of the respondents, believe that an algorithm is a set of mathematical equations, 26% regard it as a step-by-step procedure, and 14% expect an algorithm to be a single formula or logic. These results indicate that many humans expect algorithms to have numerous computational steps and several mathematical formulas, which we refer to as having a higher level of complexity. In other words, humans expect the algorithm to have some complexity, even if such complexity is not always necessary to provide good advice and outperform human judgment (Fischer & Harvey, 1999).

We define *simple* and *complex* algorithms as approaches that fall short of or go beyond the definitions expressed in Logg et al. (2019). A *simple algorithm* is a mathematical approach that can be easily understood by a typical user, such as a simple formula, and a *complex algorithm* is a sophisticated approach that is difficult for a typical user to understand, such as multiple mathematical computations. Regardless of its complexity, we only consider algorithms that fit to the problem and provide valuable advice. We aim to analyze how making such simple and complex algorithms transparent affects the use of their advice.

Question 1. How does increasing the transparency of simple and complex algorithms affect the use of their advice?

The survey of Logg et al. (2019) also observed that expectations toward algorithms can vary widely. Decision-makers with different expectations could have different perceptions of the capability of an algorithm to provide meaningful advice for a decision task. One decision-maker can perceive an algorithm’s complexity as appropriate for solving a task, and another decision-maker might perceive the same algorithm as too complex or too simple to solve the same task. Therefore, algorithm transparency may also affect the use of advice with respect to the decision-makers’ individual perceptions of the appropriateness of algorithmic complexity.

Question 2. How does increasing the transparency of algorithms that are perceived as too simple or too complex affect the use of their advice?

We expect that the use of advice when making an algorithm transparent rather depends on the individually perceived appropriateness of the algorithm’s complexity than on the algorithm itself. If an algorithm is perceived to be simpler than appropriate, it falls short of people’s expectations, leaving them disappointed and underwhelmed. This could result in a poor perception of its advice. Because poor perception of advice leads to a low use of advice (Yaniv, 2004b), we expect that increasing the transparency of algorithms that are perceived to be simpler than appropriate decreases the use of advice compared with nontransparent algorithms. In contrast, algorithms that are perceived to be more complex than appropriate tend to exceed the expectations of human decision-makers. There is no consistent theory about the

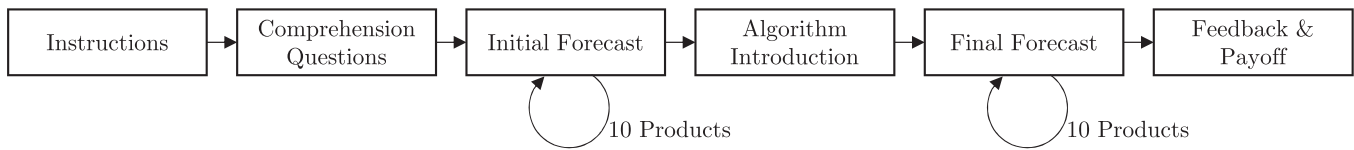


FIGURE 2 Sequence of events within the experiment

expected results: On the one hand, human decision-makers might feel overwhelmed, unable to understand the algorithms, and not willing to put trust in them but to rely on their own judgment (Gigerenzer & Gaissmaier, 2011). This could lead to a decrease in the weight on advice if an algorithm is made transparent and perceived to be more complex than appropriate. On the other hand, Parasuraman and Manzey (2010) and Dzindolet et al. (2003) report a positive attitude toward advanced technologies, and we presume that humans might appreciate high algorithmic complexity. Consequently, making an algorithm transparent that is perceived to be more complex than appropriate could also lead to a higher use of advice compared with a nontransparent algorithm.

To disentangle our two research questions, we develop two experiments. Both experiments share the same simple algorithm but differ in the description of the task, thereby manipulating the appropriateness of complexity and the complex treatment thereby increasing the robustness of the results.

4 | EXPERIMENT 1

We consider a demand forecasting task and design an experiment with a judge–advisor system structure. First, the decision-maker (the “judge”) makes an initial demand forecast. Second, he or she receives advice from an algorithm. Third, the judge can update the initial forecast and provide a final forecast. Judge–advisor systems are often used in the literature to analyze human behavior in advice taking (Gino, 2008; Logg et al., 2019; Yaniv, 2004b). Judgment is measured before and after receiving advice, and the relative shift in judgment indicates the use of advice (Bonaccio & Dalal, 2006). The laboratory experiment was preregistered with the Open Science Framework: osf.io/wdvhr.¹

4.1 | Experimental design

We designed an experiment in which participants had to forecast demand before and after they received a demand forecast from an algorithm. We used three treatments that differ in the transparency of the algorithmic procedures and the complexity of the algorithms. In all treatments, the participants received identical numerical advice, but the information provided and the algorithmic procedures differed.

Figure 2 depicts the sequence of events in the experiment. After reading the instructions and answering comprehension questions, participants made their initial forecasts for 10 dif-

ferent products. For each product, they observed the demand history of 10 periods and forecasted the demand in period 11 by entering an integer number in an input field. We simulated stationary demand data for each product with a mean value between 300 and 1200 and a coefficient of variation of 0.3. The demand history was displayed in a line chart (Figure 3), which is common in practice and perceived to be the best presentation style when the characteristics of the demand data are unknown (Harvey & Bolger, 1996). All participants received the same demand data.

After the participants had forecasted the demand for the 10 products, they were informed that an algorithm had also computed a forecast. Between the treatments, we varied the level of algorithm transparency and complexity. Participants could use the algorithmic advice for each product to update their initial forecast. After making their final forecasts, the participants were asked to indicate their agreement with the statement “I understood how the algorithm derived its recommendations” on a 7-point Likert-type scale (1 = *strongly disagree*; 4 = *neither agree nor disagree*; 7 = *strongly agree*). Furthermore, we asked the participants who received information about the algorithmic principles about the perceived appropriateness of algorithmic complexity for solving the forecasting task on a 7-point Likert-type scale (1 = *much less than appropriate*; 4 = *appropriate*; 7 = *much more than appropriate*).

At the end of the experiment, the participants observed the actual demand and their resulting forecast error. We monetarily incentivized high forecast accuracy using a payoff composition similar to that in Kremer et al. (2011). The payoff scheme consists of a fixed reward of \$0.50 and a bonus of up to \$0.10 per product depending on the final forecast accuracy. The accuracy was measured with the absolute percentage error that we bounded by 0 and 1:

$$\text{Absolute percentage error} = \frac{|\text{Final forecast} - \text{Actual demand}|}{\text{Actual demand}} \quad (1)$$

The bonus for each product was \$0.10 · (1 – absolute percentage error).

We measured the use of the algorithmic advice using the weight on advice (e.g., Harvey & Fischer, 1997; Yaniv, 2004a, 2004b):

$$\text{Weight on advice} = \frac{\text{Initial forecast} - \text{Final forecast}}{\text{Initial forecast} - \text{Algorithmic advice}} \quad (2)$$

The weight on advice computes the relative shift between the initial forecast and the final forecast with respect to the

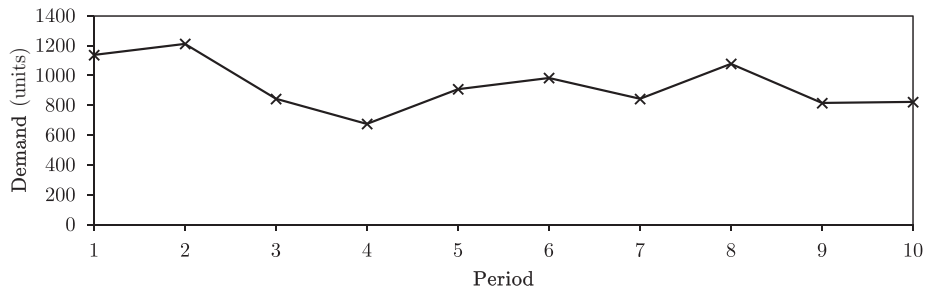


FIGURE 3 Demand history of a product with a mean of 1000 and a coefficient of variation of 0.3

algorithmic advice. A weight on advice of 0 implies that the participant did not change his or her initial forecast; a value of 1 implies that the initial forecast was replaced by the algorithmic advice. If the initial forecast equals the algorithmic advice, we set the weight on advice to 0. Values less than 0 or greater than 1 are winsorized to increase interpretability (e.g., Gino & Moore, 2007; Logg et al., 2019). In our analyses, we use the participant's average weight on advice over all 10 products.

We applied three treatments. Participants in all treatments received the same numerical advice, but the information provided and the algorithmic principles differed.

In the *nontransparent* treatment, participants were merely informed that an algorithm had computed a forecast, and they received no additional information on its underlying principles.

In the *transparent-simple* treatment, the algorithm determines the arithmetic mean demand of the 10 periods as the forecast for period 11. This algorithmic forecast is an unbiased and ex ante error minimizing estimation for stationary demand. In the experiment, participants were informed that the algorithm “calculates the forecast for a product by computing the average of the demand history of the last 10 periods.” We illustrated the computation of the average with the corresponding formula for an example product that was not used in the actual experiment.

In the *transparent-complex* treatment, the algorithm is a neural network that uses the demand history as input and provides the demand forecast for period 11 as output. It was trained with 100,000 structurally similar stationary demand histories. In the experiment, we explained the basic principles of an artificial neural network and provided the calculations necessary to comprehend and reproduce the recommended forecast.

While the underlying principles of the algorithms differ substantially, they generate the same algorithmic advice for the 10 different products. Detailed explanations for all treatments are provided in the [Supporting Information](#) of this paper.

4.2 | Experimental protocol

The experiment was programmed in oTree (Chen et al., 2016) and conducted on Amazon's Mechanical Turk (MTurk)

on May 12 and May 13, 2020. We sought to collect data from 500 participants to detect a medium-sized effect ($f = 0.15$) at a power of 0.85 and $\alpha = 0.05$ (analysis of variance [ANOVA] with three groups). A total of 1322 MTurk workers agreed to work on the experiment. A total of 176 workers did not proceed to the comprehension questions, and 626 failed to answer them correctly on the first attempt, which was a requirement to work on the experiment. Eleven workers did not finish the experiment after having passed the comprehension questions. This left us with 509 participants. At the beginning of the experiment, the participants were randomly assigned to one of the three treatments. A total of 171 participants were assigned to the nontransparent treatment, 171 participants were assigned to the transparent-simple treatment, and 167 participants were assigned to the transparent-complex treatment. On average, the participants took 10.4 min to conduct the experiment and earned \$1.36.

4.3 | Results

Table 1 summarizes the main results of our experiment. We first check whether the manipulation of the different treatments was successful. Next, we analyze the effects of transparency on the weight on advice of the simple and complex algorithms compared to the nontransparent algorithm. Then, we explore the perceived appropriateness of complexity as a moderator of the effect of algorithm transparency on the weight on advice. Unless stated otherwise, we conduct two-sided Mann–Whitney U tests with Holm–Bonferroni correction for multiple tests (Holm, 1979).

4.3.1 | Validation of treatment manipulation

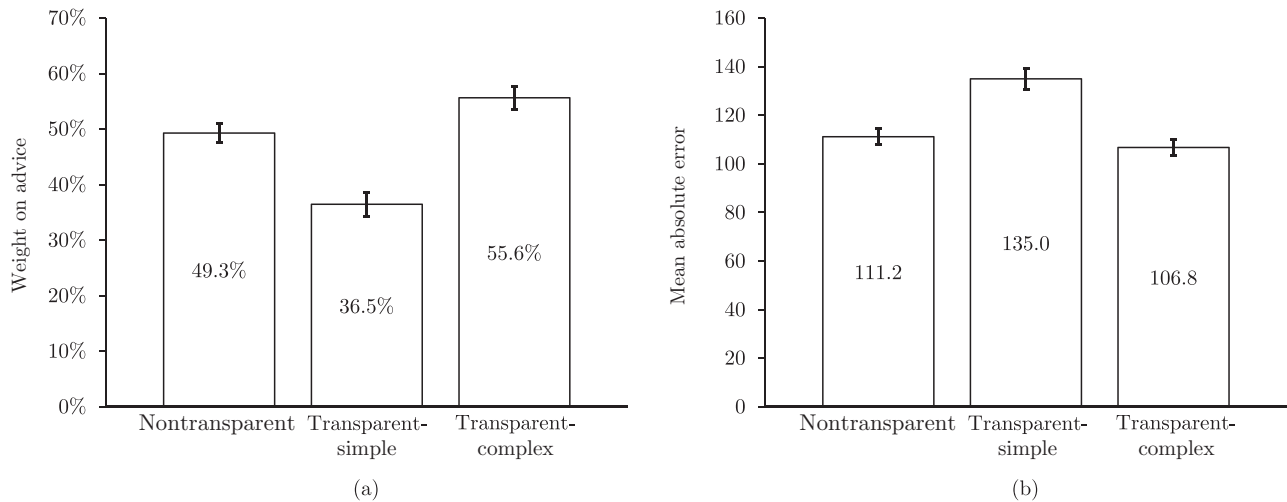
To check if the manipulation of the different treatments was successful, we analyze three questions: First, are the two algorithms intended to represent simple and complex algorithms also perceived to be “simple” or “complex,” respectively? Second, does providing transparency lead to a higher level of understanding of an algorithm? Third, is the simple algorithm better understood than the complex algorithm?

Table 1 shows that participants in the transparent-simple treatment rated the appropriateness of complexity at 3.02 on average, and participants in the transparent-complex

TABLE 1 Result overview: Means of the relevant measures in the different treatments (standard error)

Measure	Nontransparent	Transparent-simple	Transparent-complex
Appropriateness of complexity	-	3.02 (0.10)	4.55 (0.10)
Understanding	3.65 (0.13)	6.36 (0.07)	4.64 (0.12)
Weight on advice (%)	49.30 (1.72)	36.47 (2.13)	55.64 (2.07)
Performance (MAE)	111.24 (3.23)	135.00 (4.24)	106.79 (3.32)

Abbreviation: MAE, mean absolute error.

**FIGURE 4** Comparison of average (a) weight on advice and (b) MAE for the three treatments. Error bars indicate standard errors

treatment rated it at 4.55 on average. The ratings are different from one another ($p < 0.001$) and different from 4 ($p < 0.001$ for both tests), the value that corresponds to an appropriate level of complexity. This indicates that our simple and complex algorithms were also perceived as such by the participants.

The level of understanding of the algorithms' mechanisms was self-reported at the end of the experiment. Participants in the nontransparent treatment indicated a level of understanding of 3.65 on the 7-point Likert-type scale. The understanding in the transparent-simple treatment was 6.36 and was significantly higher than that in the nontransparent treatment ($p < 0.001$). The understanding in the transparent-complex treatment was 4.64 and significantly higher than that in the nontransparent treatment ($p < 0.001$) but significantly lower than that in the transparent-simple treatment ($p < 0.001$). We can conclude that transparency leads to a higher level of understanding and that the simple algorithm is better understood than the complex algorithm.

The results suggest that our manipulation was successful.

4.3.2 | Effects of transparency of simple and complex algorithms on the weight on advice

We now analyze the difference in the weight on advice across treatments (Figure 4a). In the nontransparent treatment, par-

ticipants had a weight on advice of 49%. Soll and Larrick (2009) and Soll et al. (2021) argue that averaging between the initial forecast and advice is a good strategy when there is no strong reason to favor one or the other. In the transparent-simple treatment, the weight on advice was 37% which is significantly lower than that in the nontransparent treatment ($p < 0.001$). In the transparent-complex treatment, the weight on advice was 56% which is significantly higher than that in the nontransparent treatment ($p = 0.019$).

As a robustness check, we conduct an ordinary least squares (OLS) regression (Table 2) with the two transparent treatments as independent variables and the weight on advice as the dependent variable (Model 1). The results indicate that the observed effects remain significant when controlling for the age, gender, and level of education of the participants (Model 2; detailed effects of controls are reported in the Supporting Information).

Because the use of algorithmic advice should improve the final forecast, we analyze the accuracy of the final forecast. As a measure, we use the mean absolute error (MAE), that is, the absolute deviation of the forecast from the mean demand, averaged over all products. We find that participants in the transparent-simple treatment performed worse than those in the other two treatments (Figure 4b). Compared to the nontransparent treatment, the MAE of the final forecast was 21% higher in the transparent-simple treatment ($p < 0.001$). The MAE of the participants in the transparent-complex

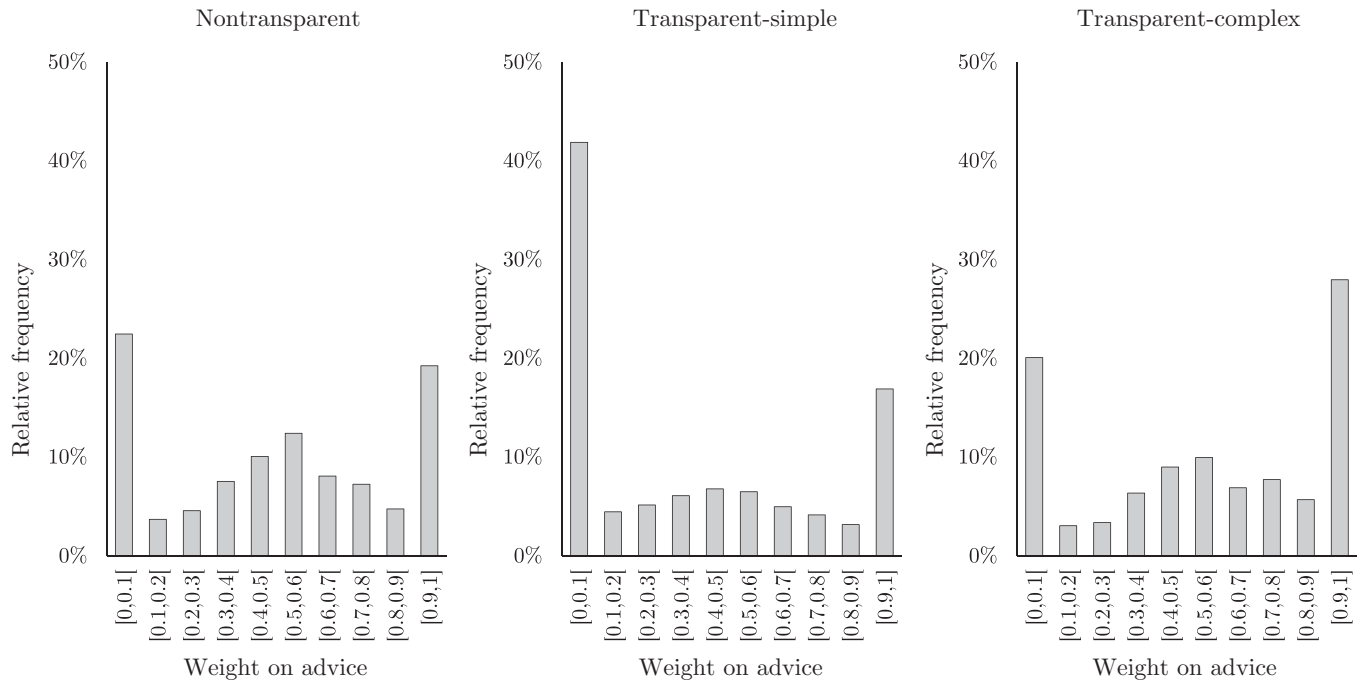


FIGURE 5 Relative frequency of the weight on advice among the three treatments

TABLE 2 Effect of algorithm transparency in the two treatments on weight on advice

	Dependent variable	
	Weight on advice	
	(1)	(2)
Transparent-simple	-0.128*** (0.028)	-0.138*** (0.027)
Transparent-complex	0.063* (0.028)	0.055* (0.027)
Constant	0.493*** (0.020)	0.490*** (0.063)
Observations	509	509
Controls	N	Y
R ²	0.087	0.183

Note: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

treatment was slightly lower than that of the participants in the nontransparent treatment, but—other than for the weight on advice—the difference was not significant ($p = 0.173$). In line with the weight on advice, the participants did not exploit the full potential of the algorithmic forecast, with an MAE of 66.2 units.

Figure 5 displays the relative frequency of the weight on advice among the three different treatments. We can identify the typical trimodal pattern in which participants either fully ignore the advice, average the initial forecast and advice, or fully accept advice and ignore their initial forecast (Ache et al., 2020; Soll & Larrick, 2009). In the nontransparent treatment, for 23% of all decisions, participants ignored the

advice, for 12% of all decisions they averaged and for 19% of the decisions they fully adopted the advice. This picture flips in the transparent-complex treatment, where for only 20% of all decisions, participants ignored the advice, for 11% of the decisions they averaged and for 28% of the decisions they fully adopted the advice. In the transparent-simple treatment, however, for 42% of the decisions, the advice was ignored—the highest share of neglected advice among all treatments—for 7% of the decisions, the advice was averaged, and for 17% of the decisions, the advice was fully adopted. We can conclude that the differences between the mean weight on advice of the treatments were mainly caused by different shares of participants that either fully accepted or neglected the algorithmic advice.

In summary, providing transparency reduced the weight on advice for the simple algorithm and increased the weight on advice for the complex algorithm. In our experiment, in which both advice-giving algorithms were optimal and provided identical forecasts, increasing the transparency reduced the forecast accuracy for the simple algorithm without significantly affecting the forecast accuracy for the complex algorithm.

4.3.3 | Perceived appropriateness of complexity as a moderator of the effects of algorithm transparency on the weight on advice

The previous analyses focused on the effect of algorithm transparency on the mean weight on advice for two specific algorithms with different complexity levels. We next analyze the extent to which the effect of algorithm

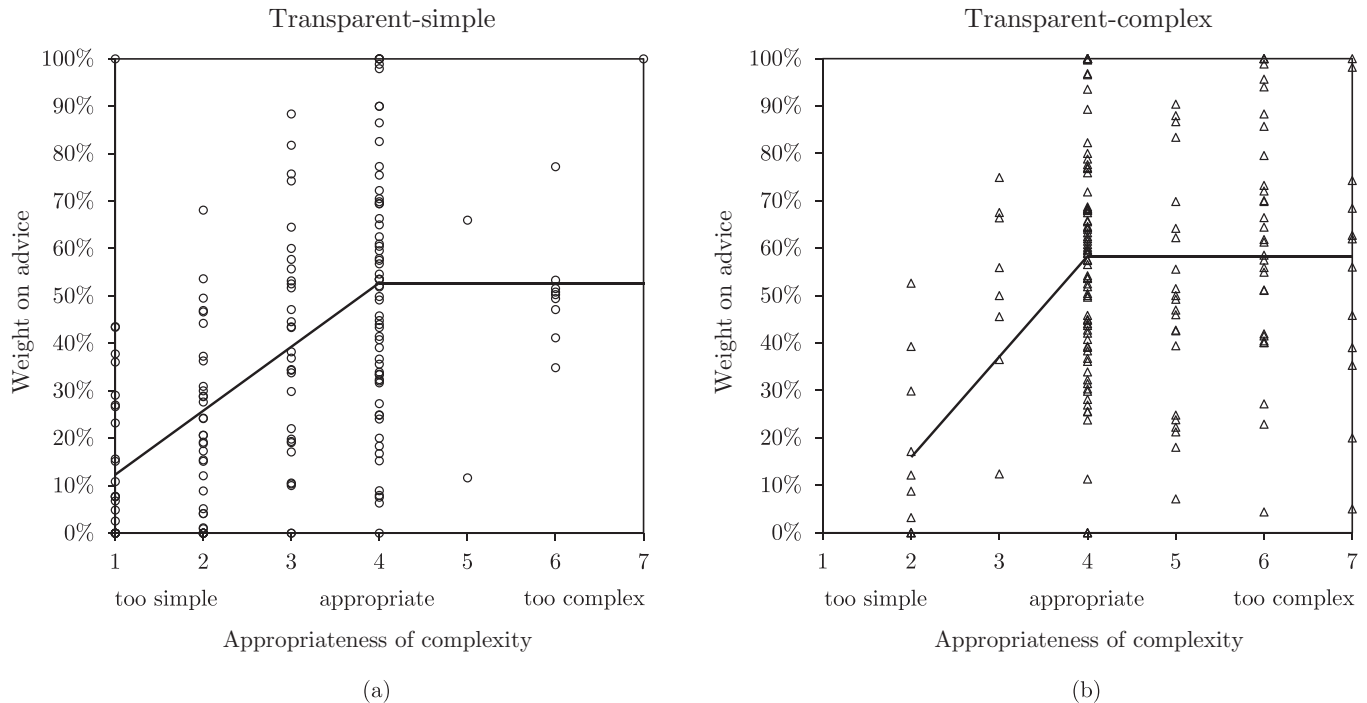


FIGURE 6 Piecewise linear regression for the effect of the perceived appropriateness of algorithmic complexity on weight on advice for (a) the transparent-simple treatment and (b) the transparent-complex treatment, indicating a plateau curve

transparency on weight on advice is moderated by the perceived appropriateness of algorithm complexity.

Figure 6 shows how individuals perceived the appropriateness of algorithm complexity in relation to their weight on advice in the two transparent treatments. To analyze whether a deviation from the appropriate level of complexity in both directions has an influence on the weight on advice, we run piecewise linear regressions with unknown breakpoints (Muggeo, 2003) (Figure 6). The model estimates a plateau curve with a breakpoint at the value of 4, the *appropriate* level of complexity for both treatments (Davies' test: transparent-simple treatment, $p = 0.06$; transparent-complex treatment, $p < 0.001$; Davies, 1987). For values smaller than the breakpoint (*too simple*), we observe a significantly positive slope of the regression model (transparent-simple treatment: $\beta_1 = 0.1276$, $p < 0.001$; transparent-complex treatment: $\beta_1 = 0.3486$, $p = 0.004$). For levels greater than the breakpoint (*too complex*), the slope of the regression model is not significant (transparent-simple treatment: $\beta_2 = 0.0178$, $p = 0.898$; transparent-complex treatment: $\beta_2 = 0.0199$, $p = 0.997$). In other words, the weight on advice significantly decreases and the more the participants perceive the transparent algorithm to be simpler than appropriate. Perceiving the algorithm as more complex than appropriate has no significant effect on the weight on advice, independent of the applied algorithm.

To further explore the relationship between the perceived appropriateness of complexity and the weight on advice, we run an OLS regression with the nontransparent treatment as constant. In addition to the two transparent treatments as control variables, we introduce three independent variables that capture the degree to which the algorithm was perceived as

appropriate (assigning a value of 1 if the reported value for the appropriateness question was 4 and 0 otherwise), *too simple* (translating a value of 1, 2, or 3 for the appropriateness question to 3, 2, or 1), or *too complex* (translating a value of 5, 6, or 7 for the appropriateness question to 1, 2, or 3). We computed the interaction between transparency and the perception of the appropriateness of complexity and present the results in Table 3. The differences in the weight on advice are no longer explained by the different algorithms but by the individuals' perception of the appropriateness of algorithm complexity (Model 1). We find that the appropriateness of complexity is a moderator for the effect of transparency on the weight on advice when the algorithm is perceived as too simple (significant negative effect). In contrast, perceiving an algorithm as too complex or appropriate does not have a significant effect on the weight on advice. The results remain robust when we control for the gender, age, and level of education of the participants (Model 2) and for models containing only one of the two transparent treatments (Models 3 and 4 for the transparent-simple treatment; Models 5 and 6 for the transparent-complex treatment).

In summary, we find indications for a plateau curve: Participants who perceive a transparent algorithm as too simple place a lower weight on advice on the recommendation of the algorithm compared to participants who perceive the algorithm as being appropriate or too complex. Perceiving an algorithm as too complex has no significant effect on the weight on advice. This suggests that not the objective complexity of an algorithm itself (Research Question 1), but the subjective perception of the appropriateness of complexity (Research Question 2) drives the results. We

TABLE 3 Effect of transparency, appropriateness of algorithm complexity, and treatments on the weight on advice with the nontransparent treatment as the constant

	<i>Dependent variable</i>					
	<i>Weight on advice</i>					
	(1)	(2)	(3)	(4)	(5)	(6)
Transparent × Perceived as too simple	−0.115*** (0.025)	−0.100*** (0.025)	−0.110*** (0.029)	−0.104*** (0.029)	−0.166** (0.056)	−0.117* (0.056)
Transparent × Perceived as too complex	0.033 (0.027)	0.041 (0.026)	0.048 (0.045)	0.061 (0.045)	0.018 (0.039)	0.036 (0.038)
Transparent × Perceived as appropriate	0.064 (0.052)	0.065 (0.051)	0.052 (0.069)	0.040 (0.068)	0.047 (0.080)	0.075 (0.081)
Transparent-simple	−0.027 (0.054)	−0.054 (0.053)	−0.031 (0.064)	−0.044 (0.064)		
Transparent-complex	0.027 (0.053)	0.013 (0.052)			0.055 (0.078)	0.015 (0.078)
Constant	0.493*** (0.018)	0.484*** (0.059)	0.493*** (0.018)	0.445*** (0.069)	0.493*** (0.018)	0.543*** (0.076)
Observations	509	509	342	342	338	338
Controls	N	Y	N	Y	N	Y
R ²	0.233	0.298	0.228	0.294	0.092	0.186

Note: Models (1) and (2)/(3) and (4)/(5) and (6) contain the responses of participants in all/the nontransparent and transparent-simple / the nontransparent and transparent-complex treatments, respectively.

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

disentangle both explanations with a second experiment in the following section.

5 | EXPERIMENT 2

We conducted a second experiment to further explore the observation that the perception of the appropriateness of algorithmic complexity rather than the algorithm itself drives the effect of transparency on the weight on advice. Furthermore, we show that this effect remains present in a setting where decision-makers are aware of the underlying demand pattern.

5.1 | Experimental design

The second experiment was identical to the first one, with two exceptions. First, in the beginning, we informed the participants that “[t]he products are known to follow a stationary normally distributed demand” and that “[t]he demand of each product randomly fluctuates around a constant level.” Thus, we created an artificial setting in which all demand characteristics of the products are known to the forecaster beforehand. Second, we used a linear regression algorithm in the transparent-complex treatment. The algorithm computes a recommendation via the linear regression with the time as the independent variable and the demand as the dependent variable. It then evaluates the p -value of the estimated trend coefficient. If $p < 0.05$, the trend is ignored, and the

algorithm runs another linear regression without a trend to estimate the demand in period 11. With the trend, the algorithm uses the initial linear regression to forecast the demand in period 11. Since our products exhibited a stationary normally distributed demand, the algorithm never observed any trend and always computed a linear regression without a trend to forecast the demand in period 11. This results in advice that is identical to the advice of the simple algorithm that uses the arithmetic mean demand as forecast.

In this setting, there is little reason to believe that the simple algorithm is simpler than appropriate and the linear regression we use as a complex algorithm does not add value to the task at hand. Thus, we expect a smaller, if any, treatment effect of the weight on advice compared to Experiment 1. On the individual level, we expect the plateau curve to be the general relationship between the weight on advice and perceived appropriateness of complexity to hold, as in Experiment 1.

5.2 | Experimental protocol

The experiment was programmed in oTree (Chen et al., 2016) and conducted on Amazon’s Mechanical Turk (MTurk) on September 13 and September 14, 2021. We collected data from at least 500 participants to detect a medium-sized effect ($f = 0.15$) at a power of 0.85 and $\alpha = 0.05$ (ANOVA with three groups). A total of 1112 MTurk started our experiment, of which 65 did not proceed to the comprehension questions.

TABLE 4 Result overview: Means of the relevant measures in the different treatments (standard error)

Measure	Nontransparent	Transparent-simple	Transparent-complex
Understanding	5.05 (0.15)	6.22 (0.06)	5.75 (0.10)
Appropriateness of complexity	-	4.98 (0.11)	5.25 (0.09)
Weight on advice (%)	54.32 (2.06)	56.57 (2.07)	60.34 (1.97)
Performance (MAE)	118.06 (4.93)	121.07 (4.84)	122.13 (5.28)

Abbreviation: MAE, mean absolute error.

A total of 442 participants failed to answer the comprehension questions correctly, and 70 participants did not complete the entire experiment. This left us with 534 participants. The participants were randomly assigned to one of the three treatments. A total of 163 participants were assigned to the nontransparent treatment, 192 participants were assigned to the transparent-simple treatment, and 179 participants were assigned to the transparent-complex treatment. On average, the participants took 10.5 min to complete the experiment and earned \$1.35.

5.3 | Results

We present an overview of the main results of the second experiment in Table 4. In the following, we check whether our treatment manipulation was successful. Then, we analyze the effects of transparency on the weight on advice for the simple and the complex algorithm. Finally, we assess the effect of the individually perceived appropriateness of complexity on the weight on advice. Unless stated otherwise, we conduct two-sided Mann–Whitney U tests with the Holm–Bonferroni correction (Holm, 1979).

5.3.1 | Validation of treatment manipulation

The level of understanding was lowest in the nontransparent treatment where participants indicated it on average as 5.05 on a 7-point Likert-type scale (Table 4). The level of understanding was 6.22 in the transparent-simple treatment, which is significantly higher than that in the nontransparent treatment ($p < 0.001$). In the transparent-complex treatment, the level of understanding was reported to be 5.75 on average, which is significantly higher than in the nontransparent treatment ($p = 0.002$) and significantly lower than in the transparent-simple treatment ($p < 0.001$). Participants in the transparent-simple treatment and the transparent-complex treatment did not report significantly different appropriateness of complexity. Participants in the transparent-simple treatment rated the appropriateness of complexity at 4.98 on average, whereas participants in the transparent-complex treatment rated it at 5.25. The ratings are not different from each other ($p = 0.17$). The lack of difference in the perceived level of complexity between the two treatments can be explained by the salience of the demand pattern: If the demand is known to follow a stationary normal distri-

TABLE 5 Effect of algorithm transparency in the two treatments on weight on advice

	Dependent variable	
	Weight on advice	
	(1)	(2)
Transparent-simple	0.023 (0.029)	0.021 (0.029)
Transparent-complex	0.060* (0.029)	0.061* (0.030)
Constant	0.543*** (0.021)	0.625*** (0.125)
Observations	534	534
Controls	N	Y
R^2	0.008	0.030

Note: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

bution, any algorithm might be perceived to be rather over complex.

The results suggest that our treatment manipulation was successful. The manipulations of transparency and complexity led to significant differences between the self-reported levels of understanding.

5.3.2 | Effects of transparency of simple and complex algorithms on the weight on advice

Considering our main insights from the first experiment and the fact that the perceived appropriateness of complexity is similar in the two transparent treatments, we do not expect to see large treatment differences in the weight on advice or the MAE. Indeed, Figure 7a reveals that the weight on advice among the three treatments is similar. The average weight on advice in the nontransparent treatment does not significantly differ from the weight on advice in the nontransparent treatment ($p = 0.439$). The weight on advice in the transparent-complex treatment significantly differs from the weight on advice in the nontransparent treatment ($p = 0.013$) but does not significantly differ from the weight on advice in the transparent-simple treatment ($p = 0.130$). We conducted an additional OLS regression that supports these initial results (Table 5).

We do not observe significant differences between the MAE of the three treatments (Figure 7b). In the

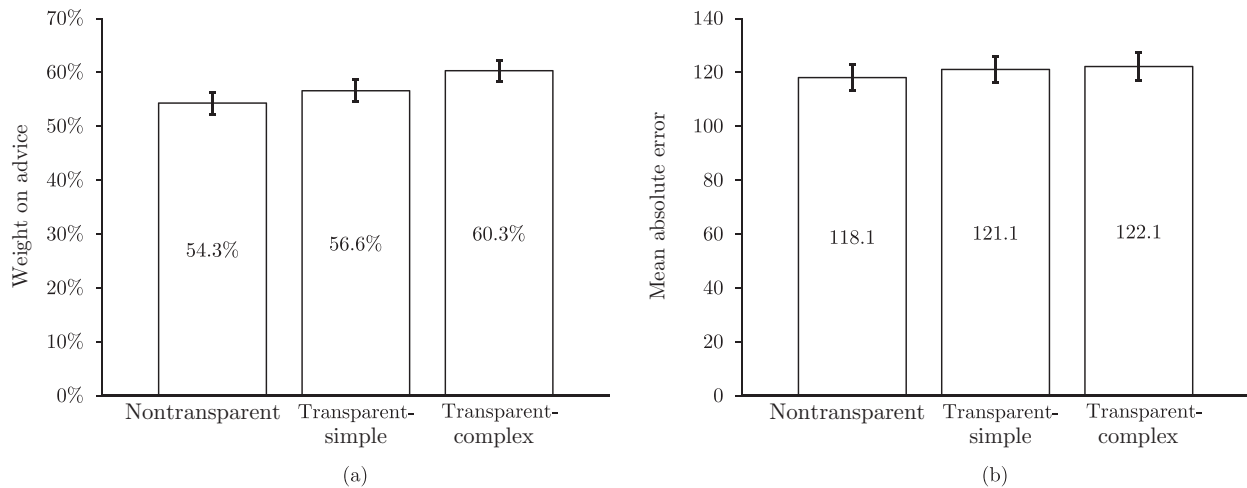


FIGURE 7 Comparison of average (a) weight on advice and (b) MAE for the three treatments. Error bars indicate standard errors

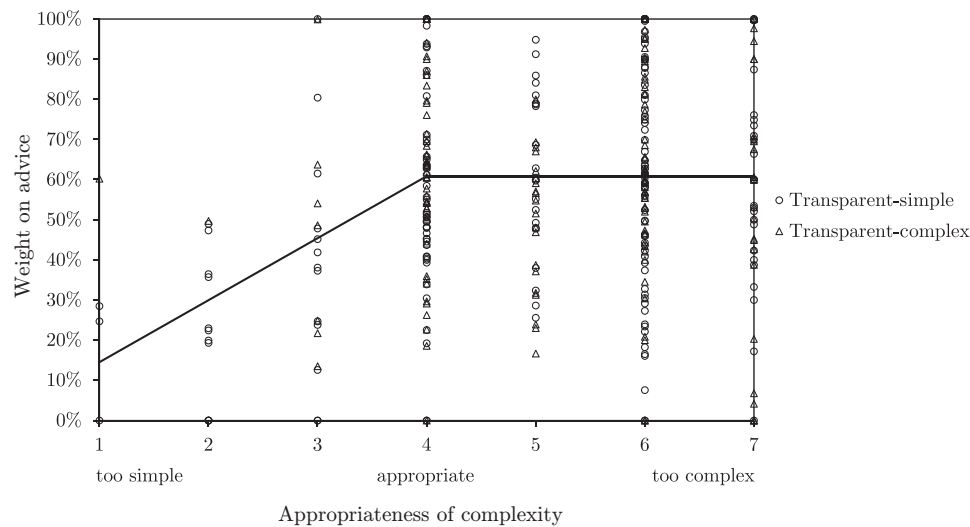


FIGURE 8 Piecewise linear regression for the effect of the level of perceived appropriateness of algorithmic complexity on the weight on advice, indicating a plateau curve

nontransparent treatment the MAE is 118.1, and in the transparent-simple treatment it is 121.1. These values are not significantly different from each other ($p = 0.737$). In the transparent-complex treatment, the MAE is 122.1, which is neither significantly different from the nontransparent treatment ($p = 0.764$) nor from the transparent-simple treatment ($p = 0.9903$).

5.3.3 | Perceived appropriateness of complexity as a moderator of the effects of algorithm transparency on the weight on advice

To analyze the individually perceived appropriateness of complexity and its effect on the weight on advice, we run a piecewise linear regression model with unknown break-

points (Muggeo, 2003) (Figure 8). The model estimates a plateau curve with a breakpoint at the value of 4, which resembles an *appropriate* level of complexity (Davies' test $p < 0.001$; Davies, 1987). On the left-hand side of the breakpoint, the model estimates a significant positive slope ($\beta_1 = 0.1533$, $p < 0.001$). On the right-hand side of the breakpoint, the model finds no significant slope ($\beta_2 = -0.0027$, $p = 0.922$). These results support our findings in the first experiment. The more participants perceived the algorithm to be simpler than appropriate, the lower the weight on advice. Perceiving an algorithm as more complex than appropriate has no effect on the weight on advice.

We validate the plateau curve with an OLS regression (Table 6). The appropriateness of complexity negatively affects the weight on advice, if the algorithm is perceived

TABLE 6 Effect of transparency, appropriateness of algorithm complexity, and treatments on the weight on advice with the nontransparent treatment as the constant

	<i>Dependent variable</i>					
	<i>Weight on advice</i>					
	(1)	(2)	(3)	(4)	(5)	(6)
Transparent × Perceived as too simple	−0.152*** (0.038)	−0.157*** (0.038)	−0.185*** (0.046)	−0.203*** (0.047)	−0.025 (0.074)	−0.0005 (0.075)
Transparent × Perceived as too complex	0.005 (0.024)	0.003 (0.024)	0.005 (0.033)	0.002 (0.034)	0.015 (0.035)	0.009 (0.035)
Transparent × Perceived as appropriate	0.002 (0.057)	0.004 (0.059)	−0.024 (0.079)	−0.033 (0.081)	0.048 (0.083)	0.037 (0.086)
Transparent-simple	0.052 (0.057)	0.052 (0.057)	0.067 (0.073)	0.071 (0.074)		
Transparent-complex	0.064 (0.057)	0.069 (0.058)			0.028 (0.078)	0.043 (0.079)
Constant	0.543*** (0.021)	0.716*** (0.123)	0.543*** (0.021)	0.746*** (0.123)	0.543*** (0.021)	0.777** (0.266)
Observations	534	534	355	355	342	342
Controls	N	Y	N	Y	N	Y
R ²	0.069	0.091	0.102	0.147	0.016	0.053

Note: Models (1) and (2)/(3) and (4)/(5) and (6) contain the responses of participants in all/the nontransparent and transparent-simple/the nontransparent and transparent-complex treatments, respectively.

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

to be too simple. We do not see this effect for models containing only the nontransparent and the transparent-complex treatment (Models 5 and 6). This might be explained because only ten out of 179 participants in the transparent-complex treatment perceived the algorithm to be simpler than appropriate (i.e., Likert score lower than 4). Perceiving the algorithm as more complex than appropriate or as appropriate still has no significant effect on the weight on advice.

In summary, the results support our conclusions from the first experiment. In both experiments, the treatment effect on the weight on advice vanished once we controlled for the appropriateness of complexity (Tables 3 and 6). Thus, with regard to Research Question 1, we can conclude that when making an algorithm transparent, there is no indication of an effect of the objective complexity on the use of advice. In both experiments, there is a strong indication that differences in the weight on advice are driven by the individually perceived appropriateness of algorithmic complexity. Thus, with regard to Research Question 2, we can summarize that the effect of the perceived appropriateness of complexity on the weight on advice can be visualized by a plateau curve. Perceiving an algorithm to be simpler than appropriate has a significant negative effect on the weight on advice. In contrast, perceiving an algorithm to be more complex than appropriate has no significant effect on the weight on advice.

6 | DISCUSSION

The concept that users should understand the algorithms they are working with is prevalent in many literature streams. For example, evolving research on explainable artificial intelligence focuses on the question of how to make complex machine learning algorithms appear simple and interpretable (Preece, 2018). Other studies, such as Ustun and Rudin (2019) and Jung et al. (2020), develop simple and easy-to-understand heuristics that provide decision support. Our findings suggest that these efforts could turn out to be a pitfall. Although an explanation improved the understanding of the algorithms in our prestudy and experiments, this did not translate into an increased use of algorithmic advice. We find indications that the use of advice of a transparent algorithm is moderated by the perceived appropriateness of its complexity, following a plateau curve. In cases where the transparent algorithm was seen to be too simple to solve a problem, the use of advice decreased in comparison to the nontransparent algorithm. Therefore, in settings where decision-makers may perceive an easy-to-understand algorithm as too simple, there is the risk that the use of advice is low.

We expected that the use of advice on a transparent algorithm depends on whether the algorithm's complexity goes beyond or falls short of the decision-maker's expectations. The results of both experiments confirm that making an algorithm transparent that is perceived as simpler than appropriate

leads to a lower use of advice compared with a nontransparent algorithm. This decrease in the use of advice might be caused by not meeting expectations that humans have when thinking about an algorithm. People expect some kind of complexity behind an algorithm (Logg et al., 2019), regardless of the problem they have to solve. Making an algorithm transparent that falls short of these expectations has detrimental effects on decision-makers' attitude toward the algorithm: they might feel disappointed, underwhelmed, and fear that the algorithm misses important procedural steps to give meaningful advice. This leads to lower use of advice compared with a nontransparent algorithm. Both experiments also show that perceiving an algorithm as more complex than appropriate has no significant effect on the weight on advice. Current research on the effect of model interpretability on trust in algorithms is consistent with our finding: For example, Ahn et al. (2021) show that explaining the results of an algorithm does not consistently result in greater or lower trust in the algorithm.

Our research is a first step in understanding how algorithms should be designed for successful human-machine collaboration. While we carefully designed the algorithms and experiments, we are aware of limitations that leave interesting areas for future research. For example, to isolate the effect of transparency and complexity on the use of advice from the potential effect of advice quality on the use of advice, we focused on algorithms that provided valuable advice. Future research could analyze settings in which this is not the case. Moreover, the decision-maker and the algorithm may have asymmetric or complementary information. Such situations pose additional challenges to the design of proper algorithms since mere compliance with an algorithm does not necessarily lead to successful human-machine collaboration (e.g. Fügener et al., 2021). How can such collaboration be designed so that decision-makers use algorithmic advice and adapt it only when necessary and appropriate? How does transparency affect such cooperation?

It could also be worthwhile to analyze how repeated interactions with feedback affect the use of advice when algorithms are made transparent. Seminal studies, such as Dietvorst et al. (2015) and Castelo et al. (2019), have shown that humans show an aversion toward the use of algorithms after seeing them err. This may also influence the effect of transparency on the use of advice. More research also needs to be conducted on the way that the lengthiness and complexity of explanations affect understanding.

Last, we see potential in analyzing how other dimensions, such as ethical standards and fairness, influence the use of algorithmic advice if the underlying principles of the algorithms are revealed. The analysis of the effect of partial transparency on the weight on advice could also be insightful.

7 | CONCLUSION

In this paper, we analyzed how algorithm complexity influences the effects of algorithm transparency on the use of advice in managerial decision tasks. We conducted two

laboratory experiments in which decision-makers received advice from algorithms of which we manipulated transparency and complexity.

In the first experiment, providing transparency on a simple algorithm reduces the use of advice, while providing transparency on a complex algorithm increases the use of advice. The treatment effect is mitigated when controlling for individual perceptions of the algorithmic complexity. We find that the effect of transparency on the weight on advice does not depend on the specific algorithm used but on the individually perceived appropriateness of complexity. Our second experiment validates this finding.

All our experiments lead to a central observation that is summarized by the plateau curve. The use of advice is affected by the perceived complexity of the algorithm. While perceiving an algorithm as too simple severely harms the use of its advice, the perception of an algorithm as being too complex has no significant effect on its use.

Our research has important implications for the interaction between humans and algorithms. Our results suggest that managers might not have to be concerned about revealing the principles of complex algorithms to decision-makers, even if the decision-makers do not fully comprehend them. However, revealing the underlying principles of an algorithm that might be perceived as being too simple can do more harm than good. Therefore, practitioners should carefully analyze and ponder the potential effects of providing transparency on advice-giving algorithms. In particular, if the algorithms apply methods that might be perceived as too simple, efforts to increase algorithm transparency can backfire and harm the use of advice and the performance of the final decision.

ACKNOWLEDGMENTS


Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2126/1 - 390838866.

ORCID

Cedric A. Lehmann  <https://orcid.org/0000-0003-1680-8250>

Christiane B. Haubitz  <https://orcid.org/0000-0002-7126-0606>

Andreas Fügener  <https://orcid.org/0000-0002-4580-7444>

Ulrich W. Thonemann  <https://orcid.org/0000-0002-3507-9498>

ENDNOTE

¹Note that while writing this paper, we slightly adapted the formulation of our research question after preregistration. Nevertheless, all preregistered hypotheses hold, and we summarize them in the [Supporting Information](#) of this paper.

REFERENCES

- Ache, F., Rader, C., & Hütter, M. (2020). Advisors want their advice to be used - but not too much: An interpersonal perspective on advice taking. *Journal of Experimental Social Psychology, 89*, 103979.

- Ahn, D., Almaatouq, A., Gulabani, M., & Hosanagar, K. (2021). Will we trust what we don't understand? Impact of model interpretability and outcome feedback on trust in AI. *SSRN Electronic Journal*.
- Arvan, M., Fahminia, B., Reisi, M., & Siemsen, E. (2019). Integrating human judgement into quantitative forecasting methods: A review. *Omega*, 86, 237–252.
- Bertino, E., Merrill, S., Nesen, A., & Utz, C. (2019). Redefining data transparency: A multidimensional approach. *Computer*, 52(1), 16–26.
- Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, 101(2), 127–151.
- Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.
- Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision-making. *Journal of Behavioral Decision Making*, 33(2), 220–239.
- Cai, C. J., Jongejan, J., & Holbrook, J. (2019). The effects of example-based explanations in a machine learning interface. In *International Conference on Intelligent User Interfaces, Proceedings IUI* (pp. 258–262). Association for Computing Machinery.
- Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5), 809–825.
- Chen, D. L., Schonger, M., & Wickens, C. (2016). otree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9, 88–97.
- Cramer, H., Evers, V., Ramlal, S., Van Someren, M., Rutledge, L., Stash, N., Aroyo, L., & Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, 18(5), 455–496.
- Dantzig, G. B. (1957). Discrete-variable extremum problems. *Operations Research*, 5(2), 266–277.
- Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 74(1), 33–43.
- De Baets, S., & Harvey, N. (2020). Using judgment to select and adjust forecasts from statistical models. *European Journal of Operational Research*, 284(3), 882–895.
- 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.
- Diubin, G., & Korbut, A. (2008). Average behavior of greedy algorithms for the minimization knapsack problem: General coefficient distributions. *Computational Mathematics and Mathematical Physics*, 48(9), 1521–1535.
- Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6), 697–718.
- Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *Management Information Systems Quarterly*, 45(3), 1527–1556.
- Fildes, R. (2006). The forecasting journals and their contribution to forecasting research: Citation analysis and expert opinion. *International Journal of Forecasting*, 22(3), 415–432.
- Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces*, 37(6), 570–576.
- Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, 25(1), 3–23.
- Fildes, R., Goodwin, P., & Önkald, D. (2019). Use and misuse of information in supply chain forecasting of promotion effects. *International Journal of Forecasting*, 35(1), 144–156.
- Fildes, R., & Petropoulos, F. (2015). Improving forecast quality in practice. *Foresight: The International Journal of Applied Forecasting*, 36, 5–12.
- Fischer, I., & Harvey, N. (1999). Combining forecasts: What information do judges need to outperform the simple average? *International Journal of Forecasting*, 15(3), 227–246.
- Franses, P. H., & Legerstee, R. (2011). Experts' adjustment to model-based SKU-level forecasts: Does the forecast horizon matter? *Journal of the Operational Research Society*, 62(3), 537–543.
- Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, 62(1), 451–482.
- Gino, F. (2008). Do we listen to advice just because we paid for it? The impact of advice cost on its use. *Organizational Behavior and Human Decision Processes*, 107(2), 234–245.
- Gino, F., & Moore, D. A. (2007). Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making*, 20(1), 21–35.
- Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 627–660.
- Goodwin, P., & Fildes, R. (1999). Judgmental forecasts of time series affected by special events: Does providing a statistical forecast improve accuracy? *Journal of Behavioral Decision Making*, 12(1), 37–53.
- Goodwin, P., Gönül, M. S., & Önkald, D. (2013). Antecedents and effects of trust in forecasting advice. *International Journal of Forecasting*, 29(2), 354–366.
- Harvey, N., & Bolger, F. (1996). Graphs versus tables: Effects of data presentation format on judgemental forecasting. *International Journal of Forecasting*, 12(1), 119–137.
- Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes*, 70(2), 117–133.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.
- Jung, J., Concannon, C., Shroff, R., Goel, S., & Goldstein, D. G. (2020). Simple rules to guide expert classifications. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(3), 771–800.
- Khosrowabadi, N., Hoberg, K., & Imdahl, C. (2022). Evaluating human behaviour in response to ai recommendations for judgemental forecasting. *European Journal of Operational Research*. Advance online publication. <https://doi.org/10.1016/j.ejor.2022.03.017>
- Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 2390–2395). Association for Computing Machinery.
- Kremer, M., Moritz, B., & Siemsen, E. (2011). Demand forecasting behavior: System neglect and change detection. *Management Science*, 57(10), 1827–1843.
- Lawrence, M., Goodwin, P., O'Connor, M., & Önkald, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, 22(3), 493–518.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444.
- Lehmann, C. A., Haubitz, C. B., Fügener, A., & Thonemann, U. W. (2020). Keep it mystic? - The effects of algorithm transparency on the use of advice. In *Proceedings of the International Conference on Information Systems, ICIS '20* (pp. 1–9). Association for Information Systems.
- Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, 90–103.
- Lu, J., Lee, D., Kim, T. W., & Danks, D. (2020). Good explanation for algorithmic transparency. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, (p. 93) Association for Computing Machinery.
- Muggeo, V. M. R. (2003). Estimating regression models with unknown breakpoints. *Statistics in Medicine*, 22(19), 3055–3071.
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 381–410.

- Perera, H. N., Hurley, J., Fahimnia, B., & Reisi, M. (2019). The human factor in supply chain forecasting: A systematic review. *European Journal of Operational Research*, 274(2), 574–600.
- Prahl, A., & Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, 36(6), 691–702.
- Preece, A. (2018). Asking ‘why’ in AI: Explainability of intelligent systems - perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management*, 25(2), 63–72.
- Seymour, W. (2018). Detecting bias: Does an algorithm have to be transparent in order to be fair? In J. Bates, P. D. Clough, R. Jäschke, & J. Otterbacher (Eds.), *Proceedings of the International Workshop on Bias in Information, Algorithms, and Systems, 2018*, 25, 7. CEUR-WS.
- Sinha, R., & Swearingen, K. (2002). The role of transparency in recommender systems. In *CHI'02 Extended Abstracts on Human Factors in Computing Systems* (pp. 830–831). Association for Computing Machinery.
- Sniezek, J. A., Schrah, G. E., & Dalal, R. S. (2004). Improving judgement with prepaid expert advice. *Journal of Behavioral Decision Making*, 17(3), 173–190.
- Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others’ opinions. *Journal of Experimental Psychology: Learning, Memory and, Cognition*, 35(3), 780–805.
- Soll, J. B., Palley, A. B., & Rader, C. A. (2021). The bad thing about good advice: Understanding when and how advice exacerbates overconfidence. *Management Science*, 68(4), 2949–2969.
- Springer, A., & Whittaker, S. (2018). What are you hiding? algorithmic transparency and user perceptions. In *The 2018 AAAI Spring Symposium Series* (pp. 455–459). MIT Press.
- Springer, A., & Whittaker, S. (2019). Making transparency clear. In *Joint Proceedings of the ACM IUI 2019 Workshops* (pp. 1–5). CEUR-WS.
- Ustun, B., & Rudin, C. (2019). Learning optimized risk scores. *Journal of Machine Learning Research*, 20, 1–75.
- Wang, W., & Benbasat, I. (2007). Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems*, 23(4), 217–246.
- Webby, R., & O’Connor, M. (1996). Judgemental and statistical time series forecasting: A review of the literature. *International Journal of Forecasting*, 12(1), 91–118.
- Yaniv, I. (2004a). The benefit of additional opinions. *Current Directions in Psychological Science*, 13(2), 75–78.
- Yaniv, I. (2004b). Receiving other people’s advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, 93(1), 1–13.
- Yu, K., Berkovsky, S., Taib, R., Zhou, J., & Chen, F. (2019). Do I trust my machine teammate? an investigation from perception to decision. In *IUI'19: Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 460–468). Association for Computing Machinery.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Lehmann, C. A., Haubitz, C. B., Fügener, A., & Thonemann, U. W. (2022). The risk of algorithm transparency: How algorithm complexity drives the effects on the use of advice. *Production and Operations Management*, 31, 3419–3434. <https://doi.org/10.1111/poms.13770>