Daschner, Stefan; Obermaier, Robert

Do We Use Relatively Bad (Algorithmic) Advice? The Effects of Performance Feedback and Advice Representation on Advice Usage

Journal of Behavioral Decision Making

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Daschner, Stefan; Obermaier, Robert (2024) : Do We Use Relatively Bad (Algorithmic) Advice? The Effects of Performance Feedback and Advice Representation on Advice Usage, Journal of Behavioral Decision Making, ISSN 1099-0771, Wiley Periodicals, Inc., Hoboken, NJ, Vol. 37, Iss. 5,
https://doi.org/10.1002/bdm.70001

Journal of Behavioral Decision Making

WILEY

RESEARCH ARTICLE OPEN ACCESS

Do We Use Relatively Bad (Algorithmic) Advice? The Effects of Performance Feedback and Advice Representation on Advice Usage

Stefan Daschner[1] | Robert Obermaier[1,2]

[1]Chair of Business Economics, Accounting and Control, University of Passau, Passau, Germany | [2]Corvinus Institute of Advanced Studies (CIAS), Corvinus University Budapest, Budapest, Hungary

Correspondence: Stefan Daschner (stefan.daschner@uni-passau.de)

Received: 25 July 2023 | **Revised:** 27 September 2024 | **Accepted:** 17 October 2024

Keywords: confidence intervals | forecast | perfect automation schema | performance feedback | relatively bad advice | source of advice

ABSTRACT

Algorithms are capable of advising human decision-makers in an increasing number of management accounting tasks such as business forecasts. Due to expected potential of these (intelligent) algorithms, there are growing research efforts to explore ways how to boost algorithmic advice usage in forecasting tasks. However, algorithmic advice can also be erroneous. Yet, the risk of using relatively bad advice is largely ignored in this research stream. Therefore, we conduct two online experiments to examine this risk of using relatively bad advice in a forecasting task. In Experiment 1, we examine the influence of performance feedback (revealing previous relative advice quality) and source of advice on advice usage in business forecasts. The results indicate that the provision of performance feedback increases subsequent advice usage but also the usage of subsequent relatively bad advice. In Experiment 2, we investigate whether advice representation, that is, displaying forecast intervals instead of a point estimate, helps to calibrate advice usage towards relative advice quality. The results suggest that advice representation might be a potential countermeasure to the usage of relatively bad advice. However, the effect of this antidote weakens when forecast intervals become less informative.

1 | Introduction

Artificial intelligence (AI), that is, algorithms imitating human intelligence, is increasingly considered as a fundamental pillar in the business landscape, fundamentally changing the way businesses operate and compete (Iansiti and Lakhani 2020). As accurate forecasts of future economic developments are considered as one of the most important business competencies (Hogarth and Makridakis 1981; Önkal, Gönül, and de Baets 2019), software providers come up with forecasting algorithms that can advise a human decision-maker to improve their initial forecasts (Chacon, Kausel, and Reyes 2022, Agrawal, Gans, and Goldfarb 2018; Dawes, Faust, and Meehl 1989; Meehl 1954; for an overview, see Grove et al. 2000; Kaufmann and Wittmann 2016).

At the same time, recent research (e.g., Önkal, Gönül, and de Baets 2019; Dietvorst, Simmons, and Massey 2018) increasingly focusses on how to boost trust towards algorithmic advice, an antecedent of advice usage (McKnight, Choudhury, and Kacmar 2002). Nevertheless, greater usage of advice is only beneficial if it outperforms the initial forecast of the human decision-maker. And therein lies the risk: Algorithmic advice can also be erroneous (e.g., Dietvorst, Simmons, and Massey 2015), resulting in a deterioration of forecast quality when using relatively bad advice.

This study is concerned that a one-sided view on the expected potentials arising from the adaption of algorithmic forecasting systems alone fails to recognize the risk that come along with it: the usage of relatively bad advice that may deteriorate forecast

quality within such a judge–advisor system (JAS). Merely boosting advice usage might increase the risk of using relatively bad advice without calibrating usage to relative advice quality.

Trust, and in turn advice usage, is subject to a dynamic process in repeated interactions (Hoff and Bashir 2015) and develops on the basis of expectation–disconfirmation comparisons (e.g., Oliver 1977). According to the expectation–disconfirmation theory (EDT), decision-makers adapt advice usage to the most recent information available if they are provided with performance feedback. We regard the provision of performance feedback as a pivotal influencing factor on advice usage within a JAS. Although there is plenty of empirical evidence that performance feedback on previous relatively good advice increases subsequent advice usage (e.g., Daschner and Obermaier 2022; Yu et al. 2019), we are among the first who critically investigate whether performance feedback also increases the risk of using subsequent relatively bad advice, that is, advice that decreases forecast quality.

In prior literature, the source of advice (algorithm vs. human) is another factor that seems to influence advice usage. However, empirical findings are mixed (for an overview, see Jussupow, Benbasat, and Heinzl 2020; Mahmud et al. 2022) and indicate the highly task-dependent influence of algorithmic advice on human advice usage behavior (Castelo, Bos, and Lehman 2019). Recently, there emerged two opposing phenomena: algorithm aversion, that is, human advice is used more than algorithmic advice (Dietvorst, Simmons, and Massey 2015; Onkal et al. 2009; Longoni, Bonezzi, and Morewedge 2019), and algorithm appreciation, that is, algorithmic advice is used more than human advice (Logg, Minson, and Moore 2019; You, Yang, and Li 2022; Gunaratne, Zalmanson, and Nov 2018).

Yet, this literature commonly neglects relative advice quality and is thus silent on the risk of using relatively bad advice. However, without considering relative advice quality, algorithm appreciation could indicate a relatively higher usage of bad algorithmic advice. First evidence comes from Dietvorst, Simmons, and Massey (2015) who document algorithm aversion in forecasts only when algorithms err. However, this is in contradiction with findings in highly automated environments (e.g., Mosier et al. 2001; Parasuraman and Riley 1997; Robinette et al. 2016; Robinette, Howard, and Wagner 2017; Salem et al. 2015) showing overtrust, that is, using even obviously erroneous algorithms.

This study investigates on the usage of bad algorithmic advice by building on the perfect automation schema (PAS) framework (Dzindolet et al. 2002; Prahl and Van Swol 2017). This framework is developed in the overtrust literature in highly automated environments and is now also applied in the forecast domain. It suggests relatively higher algorithmic advice usage unless explicit performance feedback reveals that they err, violating the decision-makers assumption of perfect algorithms. We shed light on an important research gap and investigate the influence of two factors, that is, performance feedback and source of advice, on (bad) advice usage.

In this study, we additionally investigate on how to mitigate the usage of relatively bad advice in forecasting tasks. Based on prior literature (Fernandes et al. 2018; Roulston et al. 2006; Joslyn and LeClerc 2012; Greis et al. 2016), we experimentally examine the influence of advice representation on advice usage while considering for relative advice quality. We thereby answer the following research questions: (1) How does performance feedback and the source of advice influence the usage of advice in repeated interactions? (2) How does advice representation influence advice usage?

We conduct two online experiments. In Experiment 1, we follow prior literature (You, Yang, and Li 2022) and manipulate performance feedback as a between-subjects factor where either continuous performance feedback after each forecast or no performance feedback at all is provided. Continuous performance feedback made sure that participants were aware that they have received a relatively good or bad advice before proceeding to the subsequent forecast. Advice is merely labeled as either coming from a human colleague of the same company or from an algorithm, but are both based on an autoregressive integrating moving average (ARIMA) forecasting model. Advice quality is dependent on the individual initial forecast and is thus a within-subjects factor. This is why we refer to relative advice quality. As advice in general outperformed participants in nine different forecast scenarios (Countries 1–9) based on a pretest with 145 students, we deliberately manipulated advice in Country 7 where relatively bad advice with a 50% error rate was provided.

In Experiment 2, we manipulate advice representation as between-subjects factor (point forecast vs. 70% confidence interval [CI] vs. 99% CI vs. best/worst case) and relative advice quality as a within-subjects factor. In the first forecasting task, relatively good advice is provided, which is followed by a relatively bad advice in the second forecasting task. Relative advice quality is based on a pretest, and participants in this experiment were not informed about relative advice quality at any time. In contrast to Experiment 1, we do not provide any performance feedback and focus exclusively on algorithmic advice.

We make four contributions: First, this study is among the first that examines on the usage of relatively bad advice in the field of forecasting. Instead of solely comparing advice usage among different sources of advice, we additionally consider relative advice quality. We therefore contribute to the algorithm appreciation literature by taking a critical perspective on this empirical phenomenon. Second, we find that performance feedback on previous relatively good advice increases the usage of the subsequent advice (compared to no performance feedback); however, this is also the case when the subsequent advice is of bad quality. As a design parameter of the human–algorithm interaction, provision of performance feedback should therefore be wisely considered before implementation. It might influence a critical judgment of advice quality of the human decision-maker at the time of decision. Third, we generally document support for the generalizability of the PAS (Prahl and Van Swol 2017), that is, a bias towards algorithmic advice in highly automated environments; however, this schema is attenuated for forecasting algorithms. Forth, providing informative forecast intervals can help to calibrate advice usage to relative advice quality and thus can mitigate the risk of using relatively bad advice. Our results are of interest for designers of forecasting systems regarding advice representation and for decision-makers interacting with algorithmic advisors.

2 | Trust and Relatively Bad Advice in Forecasting Tasks

2.1 | Trust Calibration Towards Relative Advice Quality

Trust is a major predictor of advice usage (McKnight, Choudhury, and Kacmar 2002; Pavlou and Gefen 2004) and should be calibrated to the quality of algorithmic advice (Lee and See 2004). In highly automated environments, trust calibration is facilitated by the possibility to cross-check the advice with available information at the time of decision. For example, in the aircraft cockpit, an automated decision advice requiring to increase altitude can be verified by comparing the altimeter of the aircraft with the altitude specifications according to the check lists. These "routine" tasks follow a deterministic logic, and the decision outcome can thus be generalized as being either right or wrong.

Trust calibration is impeded by the nature of forecasting tasks. Forecasts are prepared prior to "corporate action," which in turn influences the quality of the forecast. The true value of the forecasting task is realized only after a certain time lag. Cross-checks are therefore not possible at the time of the decision. Human decision-makers can thus not objectively verify whether they face a good or a bad forecasting advice but can only gain experience from previous performance feedbacks.

Moreover, it is not only a question of whether the advice will be right or wrong—as in "routine" tasks—but also to what degree it will be right. The JAS serves as a valuable tool for understanding how decision-makers engage with and incorporate advice into their final decisions and has gained recognition in the field of algorithm appreciation literature, with researchers often referring to it when examining advice usage (Himmelstein and Budescu 2022; Fildes and Petropoulos 2015; Logg, Minson, and Moore 2019). In a JAS, decision-makers are asked to make an initial forecast. Subsequently, they receive advice intended to help them refine and adjust their initial forecast. The evaluation of relative advice quality can only be made when the true value of the forecast is available. Furthermore, Daschner and Obermaier (2022) indicate that the decision-maker's own initial forecast is the benchmark to make this evaluation. Advice can thus either be good (better-than-own initial forecast) or bad (worse-than-own initial forecast), as evaluated *ex post* when the true value of the forecasting task materialized. Due to this time lag, there is uncertainty about the relative advice quality at the time of decision.

2.2 | Relatively Bad Advice

Errors are inevitable in uncertain decision domains such as forecasting (Dietvorst and Bharti 2020). As decision-makers are no crystal ball readers, they do not know the quality of advice at the time of decision, that is, *ex ante* of the materialization of the true value. However, there is little research on what constitutes an error in forecasts. We therefore follow Daschner and Obermaier (2022) who define bad advice as advice that deteriorates the initial forecast of an average decision-maker. This leads to a violation of performance expectations once explicit performance feedback is provided. Consequently, bad advice is not objectively bad but depends on the initial forecast quality of each decision-maker and the degree of advice usage. Yet, we know little how these relatively bad advices influence future advice usages.

In highly automated tasks, there is empirical evidence that decision-makers even use obviously bad advice (Salem et al. 2015; Robinette et al. 2016). This phenomenon is termed as overtrust, a heuristic replacement of vigilant information seeking and processing in favor of the usage of algorithmic advice (Mosier et al. 1996, 2001; Parasuraman and Manzey 2010). Situation unawareness and a loss of manual skills are long-term consequence when human decision-makers heuristically resort to algorithmic advice (Endsley and Kiris 1995; Billings 1991).

A critical evaluation of advice thus seems to be key in forecasting tasks. For example, there is the risk that decision-makers may heuristically accept an advice for the forecast of the demand of a product in the next period, even though this demand is extraordinarily high or low compared to historical data. There might be reasons for this; however, decision-makers should not adapt this advice without critical verification. Consequently, algorithmic advice can make decision-makers overly vulnerable if they provide relatively bad advice. In the long run, the usage of relatively bad advice could be an indication of overtrust.

3 | Experiment 1: Factors Influencing Advice Usage

3.1 | Theoretical Background and Hypotheses

3.1.1 | Performance Feedback

According to the EDT, trust develops on the basis of expectation–disconfirmation comparisons (Oliver 1977, 1980; Bhattacherjee and Premkumar 2004). Following the work of McKnight, Choudhury, and Kacmar (2002), we regard trust as an antecedent of advice usage. Consequently, an expectation confirmation maintains or even increases subsequent advice usage. This suggestion is empirically supported by the study of De Baets and Harvey (2020) where decision-makers follow inputs of good performing forecasting models more than poorly performing models. In contrast, a disconfirmation of expectations probably leads to decreases in subsequent advice usage (see Figure 1). Trust will then immediately be withdrawn as it gets transparent that it was misplaced (Hardin 1993; Lewicki, Tomlinson, and Gillespie 2006). This threshold between expectation confirmation and disconfirmation in a forecasting task seems to be dependent on the decision-maker's own initial forecasting quality (Daschner and Obermaier 2022). Unfortunately, a lack of availability of performance feedback is inherent in the forecasting domain and complicates the trust development process. Particularly in strategic business decisions, there is commonly no performance feedback available. In contrast, the accuracy of monthly costs, sales revenue, or profit forecasts is available. As these forecast errors can be distorted by unexpected or abnormal events, the provision of performance feedback can be a choice of design of the JAS. However, it should be noted that performance feedback can only be provided for previous advices given, making it notoriously unhelpful for guiding subsequent advice usages.

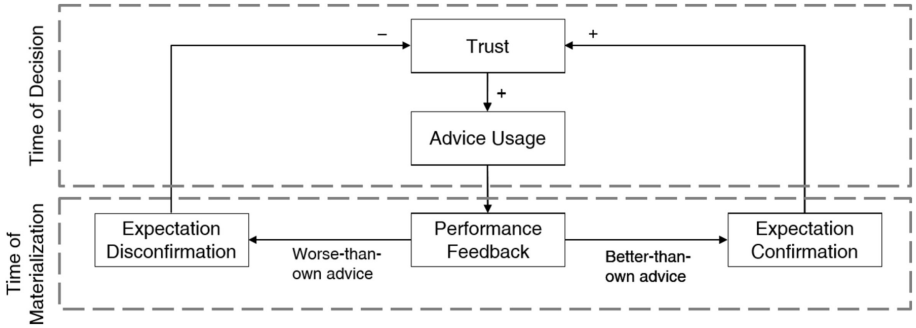

FIGURE 1 | Expectation–disconfirmation process enabled by performance feedback in repeated interactions with a forecasting advisor.

Rational decision-makers are assumed to adapt future advice usages according to previous performance information (Deutsch 1958; Kramer 2010). Performance feedback easily conveys this information (Harvey and Fischer 2005) and enables trust updates through expectation–disconfirmation comparisons (Barber 1983; Mayer, Davis, and Schoorman 1995; Rempel, Holmes, and Zanna 1985). By resolving uncertainty about prior forecast quality (of the advisor and the decision-maker's own performance), decision-makers can *ex post* verify whether the advice quality was better or worse than their own initial forecast. Again, this possibility depends highly on the forecasting setting and when the true value is observable, if at all.

Based on this framework, we hypothesize a moderating effect of performance feedback on the relationship between previous relative advice quality and subsequent advice usage in repeated interactions:

Hypothesis 1. *Performance feedback on relatively good advice increases subsequent advice usage more than without performance feedback.*

Hypothesis 2. *Performance feedback on relatively bad advice decreases subsequent advice usage more than without performance feedback.*

Due to the fact that performance feedback only reveals previous relative advice quality, it cannot provide guidance for the usage of subsequent advices. Any behavioral changes to this feedback might reflect a delayed reaction. However, if performance feedback on good prior advice increases trust and the usage of subsequent advice, it may also increase the risk of using a future relatively bad advice. This study is particularly concerned with this potential risk. We further hypothesize:

Hypothesis 3. *Revealing relatively good advice increases the subsequent usage of relatively bad advice more than without performance feedback.*

3.1.2 | The PAS

The PAS represents a cognitive structure that helps human decision-makers to organize and interpret information when interacting with an algorithmic decision supporting system (Dzindolet et al. 2002, 2003; Madhavan and Wiegmann 2007). As algorithms do never get tired, distracted, and stressed and do not have human needs, they are perceived as perfect and infallible. This perception of perfection is associated with high expectations towards relative advice quality. High expectations increase trust towards advisors. Consequently, the PAS seems to positively influence algorithmic advice usage. In contrast, humans are expected to fail as erring is perceived as a human and not as an algorithmic trait (Renier, Schmid Mast, and Bekbergenova 2021). The expectations towards human advice are thus relatively lower than towards algorithmic advice (Daschner and Obermaier 2022).

Dietvorst, Simmons, and Massey (2015) give indication that this higher expectations facet of the PAS is violated when seeing them err. Daschner and Obermaier (2022) refine the threshold when forecasting advice is erroneous. They find that subsequent advice usage decreases only if advice quality is worse than the decision-maker's own forecast quality. To violate this higher expectations facet, the provision of performance feedback is a prerequisite because this enables the expectation–disconfirmation process. We therefore hypothesize a moderating effect of performance feedback, previous relative advice quality and source of advice:

Hypothesis 4. *Algorithmic advice usage is higher than human advice usage if no relatively bad advice was previously revealed.*

Additionally, PAS suggests an all-or-none logic, that is, the algorithm is either perfect or flawed with errors that could recur in the future interaction. Therefore, we expect that relatively bad algorithmic advice leads to a relatively harsher decrease in the usage of subsequent advice when revealed by performance feedback (Prahl and Van Swol 2017) but can recover with relatively good advice again. We therefore hypothesize:

Hypothesis 5. *After performance feedback on relatively bad advice, there is a harsher decrease in advice usage for algorithmic than human advice.*

As the focus of this study is on the usage of relatively bad advice, we analogously hypothesize that algorithmic advice usage is higher than human advice usage in cases of relatively bad advice.

Hypothesis 6. *The usage of relatively bad algorithmic advice is higher than relatively bad human advice when no relatively bad advice was previously revealed.*

3.2 | Method

3.2.1 | Task and Procedure

We conducted a 2×2×9 mixed design experiment at a public university in Germany with 205 online participants (118 women; 84 men; 3 diverse; mean age = 22.73 years with an SD of 3.41) with students enrolled in the faculty of law (10.2%), economics (29.3%), philosophy (47.3%), computer science and mathematics (12.2%), and others (1.0%). Participants mainly used a laptop (84.9%) or a tablet (10.7%). Only 4.4% used their smartphones to run this experiment.

Performance feedback and source of advice are independent between-subjects factors, and relative advice quality is the within-subjects factor. All participants received the same sequence of countries and advice. The series was randomly generated to cover different seasonality and noise and to reduce carryover effects due to learning. After successful completion, participants were paid two euros fixed. To further motivate participants, they were informed that the best five participants, measured on the mean absolute percentage error (MAPE), get paid an extra bonus of 25 euros each. Informed written consent to take part in this experiment was obtained before the commencement by each participant (Data S1, Appendix S1 and Data S2).

Participants were asked to imagine themselves as managers of a fictive company that recently launched a new product in different regions. The task was to forecast the upcoming week's demand of a product based on a bar chart of the past 14 weeks in nine different countries (within-subjects factor). Each bar of the chart accumulated and disclosed the daily demands. No more extraneous information was disclosed to limit potential biases to a minimum. Participants got familiar with the forecast task in a training round including performance feedback on their absolute percentage error (APE). Afterwards, the source of advice (human colleague vs. algorithm) was introduced, and participants were informed about the availability of performance feedback (continuous performance feedback after each forecast vs. no performance feedback). If provided, performance feedback included the participant's own initial and final APE as well as the APE of the advice provided. We kept the performance feedback representation as simple as possible so that participants could easily process this information (You, Yang, and Li 2022). Figure 2 summarizes the experimental procedure. The experiment took on average 18 min and 41 s to complete with a minimum of 5 min and a maximum of approximately 70 min. All entries were required to continue in the experiment to avoid missing values.

3.2.2 | Relative Advice Quality and Relative Bad Advice

Advice in all nine countries was based on an ARIMA forecasting model and was based on the same historical data. To control for relative advice quality, we run a pretest based with 145 participants. These results revealed that the advice outperformed the average participant in eight out of nine countries (except in Country 2 with an APE of 14% of the advice compared to a MAPE of approximately 4% of the average participant). As the focus of this study is on the usage of relatively bad advice, we deliberately provide a manipulated advice in Country 7 with a percentage error of 50% to integrate strong negative consequences of advice usage. Experimental material is attached in Appendix S1.

We operationalize relative advice quality as a dichotomous variable with relatively good advice and relatively bad advice as factor expressions. In order to take account of a possible improvement in one's own initial forecast through an advice with a higher APE than the participant's initial forecast, but with a different sign (Soll and Larrick 2009), we consider an advice to be bad if the participant's final APE is higher than his/her initial APE. This is in line with Daschner and Obermaier (2022),

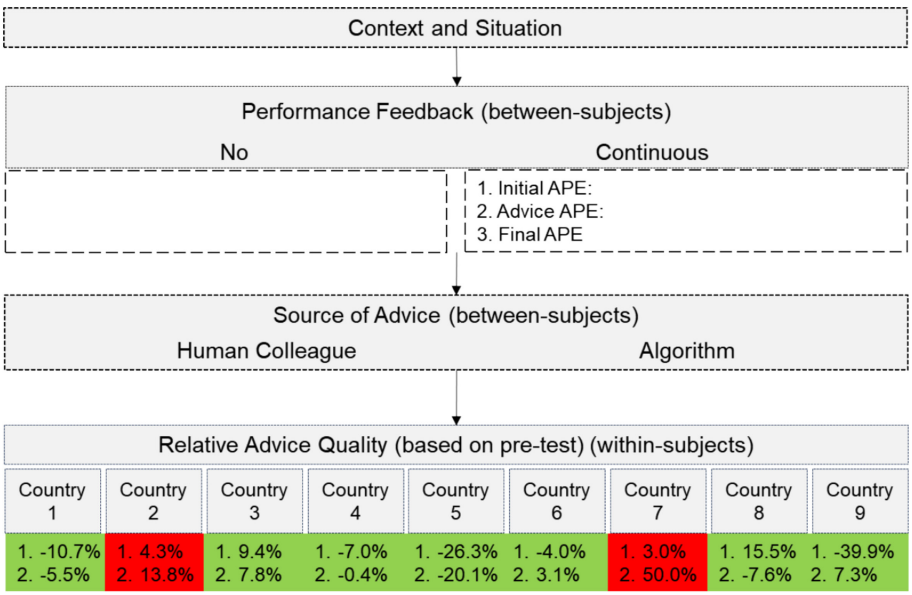

FIGURE 2 | Experimental procedure with performance feedback and source of advice as between-subjects and relative advice quality as within-subjects factor (1. = initial APE; 2. = advice APE). The majority of the pretest participants deteriorated their final APE by advice usage only in Country 2 and Country 7.

suggesting that the general perception of bad advice is subject to the participants' initial forecast quality.

3.2.3 | Dependent Variable

We follow prior literature and apply a weight on advice (WOA) variable that is commonly used in advice research (Harvey and Fischer 1997; Sniezek, Schrah, and Dalal 2004; Kaufmann et al. 2023). The change in the point forecast is set in relation to the difference between the advice and the initial point forecast. That is, a WOA of 100% indicates that the advice is completely adopted, a WOA of 0% indicates that the advice is ignored. WOA is computed via the following equation:

$$\text{WOA} = \frac{\text{final forecast} - \text{initial own forecast}}{\text{advice} - \text{initial own forecast}} \%$$

The initial own forecast is the participant's forecast before receiving advice, the final forecast the participant's adjusted forecast after receiving and weighting advice.

3.3 | Results

3.3.1 | Descriptive Statistics

3.3.1.1 | WOA Development. We follow prior literature (Logg, Minson, and Moore 2019) and winsorized any WOA values greater than 1 or less than 0. Figure 3 depicts the number of participants receiving relatively good or bad advice for the respective country, separated into treatments of source of advice and performance feedback. Also, it shows the mean WOAs in each country, separated into participants currently receiving relatively good (bad) advice. Descriptive statistics can be found in Appendix S2.

3.3.2 | Hypothesis Tests

If not otherwise stated, we test in the following analyses for the 5% (10%) significance level to determine a statistically (marginally) significant effect.

3.3.2.1 | The Influence of Performance Feedback (Hypotheses 1 and 2). Hypothesis 1 posits an interaction between performance feedback and previous relative advice quality, that is, performance feedback increases subsequent WOA relatively more than without performance feedback when previous relative advice quality was good. Vice versa, Hypothesis 2 posits a decrease in subsequent WOA after performance feedback compared to no performance feedback when previous relative advice quality was bad.

We conduct a linear mixed model analysis to examine the influence of performance feedback and previous relative advice quality on recent WOA. The fixed effects of the model included the main effects of performance feedback and previous relative advice quality, as well as the interaction term. Additionally, the effects of country, age, and gender were considered as fixed effects. The convergence criteria were set according to the Satterthwaite method. Model parameter estimation was performed using the maximum likelihood method. Random effects were accounted for by specifying a random intercept for each participant. Repeated measurements for the variable country were considered using participant as the grouping variable.

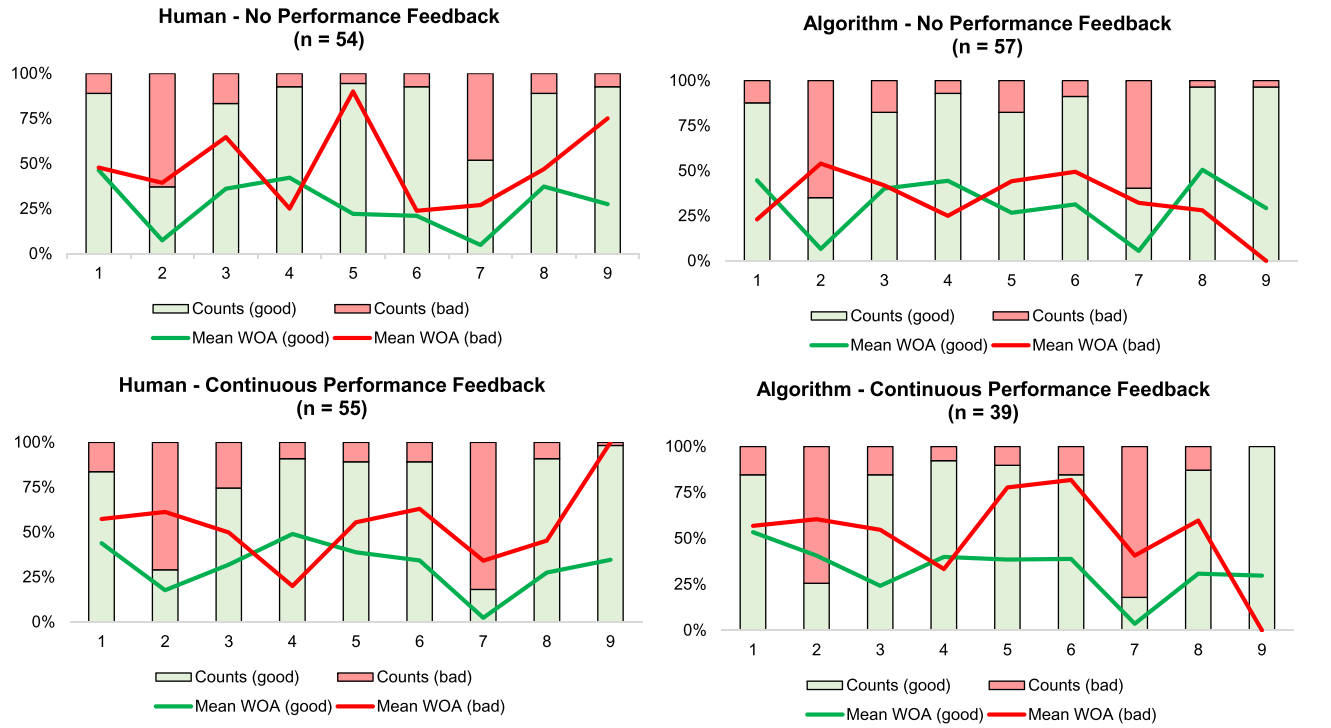

FIGURE 3 | Descriptive statistics on the share of participants receiving relatively good (bad) advice and their respective WOA in each country, separated by treatments of source of advice and performance feedback.

Results show a significant interaction between performance feedback and previous relative advice quality ($t = -4.318$, $b = -14.035$, $p < 0.001$), as well as significant main effects of performance feedback ($t = 3.942$, $b = 10.456$, $p < 0.001$) and previous relative advice quality ($t = 3.914$, $b = 8.737$, $p < 0.001$). Performance feedback positively moderates the relationship between previous relative advice quality and WOA, if relatively good advice was provided. Reversely, this moderation effect is negative if relatively bad advice was provided in the previous forecast (see Figure 4). We find support for Hypotheses 1 and 2.

3.3.2.2 | Does Performance Feedback Increase the Usage of Relatively Bad Advice? (Hypothesis 3). Hypothesis 3 states that participants use relatively bad advice more when prior performance feedback on relatively good advice was provided compared to no performance feedback on relatively good advice. We therefore select only cases where relatively bad advice was provided after a relatively good advice ($n = 311$; see Table 1).

Analogously to Hypothesis 1/Hypothesis 2, we conduct a linear mixed model with performance feedback as well as country, age, and gender as fixed factors. The effect of performance feedback is statistically significant ($t = 3.390$, $b = 12.568$, $p < 0.001$), indicating that revealing previous relatively good advice quality leads to a higher usage of subsequent relatively bad advice compared to no performance feedback. This result is in line with Hypothesis 1 and supports Hypothesis 3.

3.3.2.3 | Additional Analysis on the Number of Relatively Bad Advices Received. We further examine on the influence of the cumulated amount of relatively bad advices received previously on the effect of the interaction between performance feedback and previous relative advice quality. Again, we conduct a linear mixed model to analyze the effect of the interaction between performance feedback, previous relative advice quality, and number of relatively bad advices on WOA. There is a marginally significant three-way interaction ($t = 1.826$, $b = 3.901$, $p = 0.068$).

We therefore run this model for the two performance feedback treatments separately and excluded cases where more than four relatively bad advices were previously received due to a lack of observations. There is a significant interaction between previous relative advice quality and number of relatively bad advices received so far ($t = 3.041$, $b = 7.061$, $p = 0.002$), as well as a significant main effect of previous relative advice quality ($t = -3.566$, $b = -16.627$, $p < 0.001$) and a marginally significant main effect of number of relatively bad advices ($t = -1.856$, $b = -3.425$, $p = 0.064$). This result implies that the influence of performance feedback on relatively good previous advice quality on subsequent WOA diminishes with an increasing number

FIGURE 4 | Interaction between performance feedback and previous relative advice quality. Error bars indicate the 95% confidence interval.

TABLE 1 | Mean WOA, standard deviation (SD), and counts of participants (n) receiving relatively bad advice after a relatively good advice, separated into countries and performance feedback.

	Country	2	3	4	5	6	7	8	9
No performance feedback	Mean WOA	45, 7	60, 3	40, 0	51, 0	38, 5	29, 7	78, 1	40, 0
	SD	28, 5	21, 7	54, 8	36, 8	34, 7	26, 1	31, 0	54, 8
	n	62	2	5	12	8	57	2	5
Continuous performance feedback	Mean WOA	65, 7	66, 7	16, 7	64, 1	75, 4	36, 1	55, 9	100, 0
	SD	27, 9	38, 0	40, 8	33, 1	34, 2	26, 8	41, 4	n.a.
	n	56	5	6	9	11	66	4	1

of relatively bad advices received (see Figure 5a). As expected, no significant differences were identified in the no performance feedback treatment (previous relative advice quality with $t = 0.958$, $b = 3.448$, $p = 0.339$; number of relatively bad advices with $t = 0.311$, $b = 0.569$, $p = 0.756$; interaction with $t = 1.468$, $b = 3.297$, $p = 0.143$) (see Figure 5b).

3.3.2.4 | The Influence of Source of Advice (Hypotheses 4 and 5). Hypotheses 4 and 5 posit a moderating effect of source of advice on the interaction effect between performance feedback and previous relative advice quality. We therefore conduct a linear mixed model with source of advice, performance feedback, and previous relative advice quality as fixed factors, as well as the three-way interaction term. Additionally, the effects of country, age, and gender were considered as fixed effects. The remaining model parameters were the same as above.

The results show a significant interaction between source of advice, performance feedback, and previous relative advice quality ($t = -4.696$, $b = -9.770$, $p < 0.001$), as well as significant main effects of performance feedback ($t = 4.107$, $b = 10.791$, $p < 0.001$), previous relative advice quality ($t = 4.056$, $b = 8.627$, $p < 0.001$), and a marginally significant main effect of source of advice ($t = 1.725$, $b = 4.346$, $p = 0.086$). Therefore, we run the model separately for no performance feedback and continuous performance feedback. We find no significant difference between human and algorithmic advisors if no performance feedback is provided ($t = 1.158$, $b = 4.285$, $p = 0.249$). Neither previous relative advice quality ($t = 0.485$, $b = 3.323$, $p = 0.628$) nor the interaction with source of advice ($t = 0.732$, $b = 3.176$, $p = 0.464$) is significant. If continuous performance feedback is provided, we neither find a significant main effect of source of advice (t = 0.989, b = 3.672, $p = 0.325$), nor of previous relative advice quality (t = 0.895, b = 6.523, $p = 0.371$), however, there is a marginally significant interaction between source of advice and previous relative advice quality (t = −1.667, b = −8.182, $p = 0.096$). When previous advice was relatively good, algorithmic advice usage is higher than human advice usage. Vice versa, algorithmic advice usage is lower than human advice usage if previous advice was relatively bad (see Figure 6). In sum, the results provide very weak to no support to Hypotheses 4 and 5.

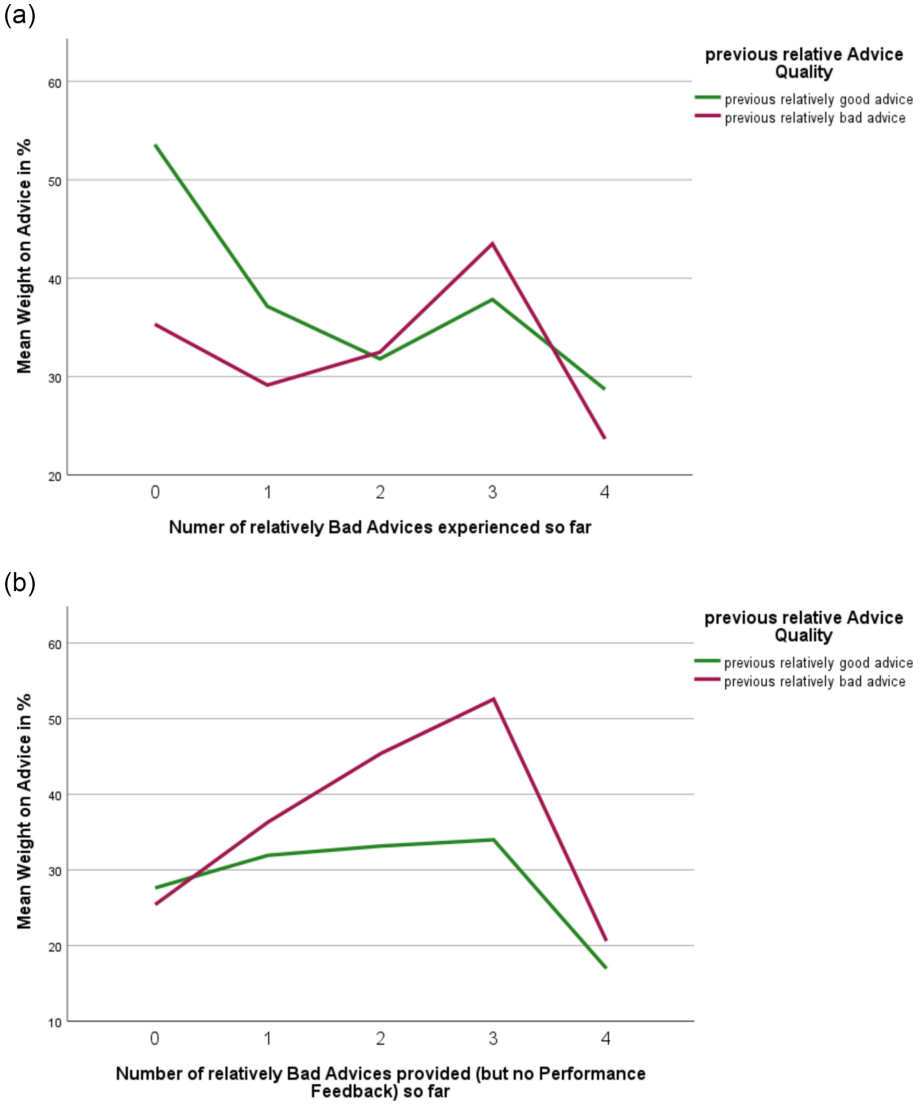

FIGURE 5 | (a) Interaction between number of relatively bad advices and previous relative advice quality, if performance feedback is provided. (b) No interaction between number of relatively bad advices and previous relative advice quality, if no performance feedback is provided.

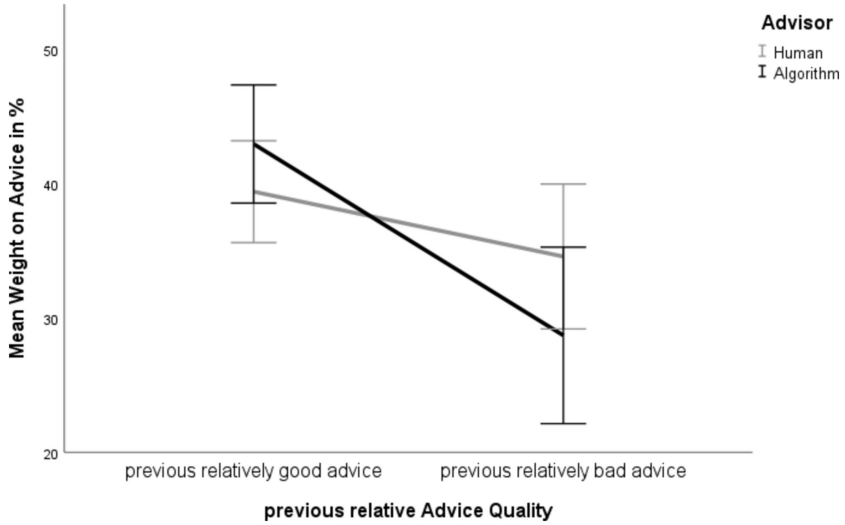

FIGURE 6 | Interaction between source of advice and previous relative advice quality, if performance feedback is provided. Error bars indicate the 95% confidence interval.

TABLE 2 | Mean WOA, standard deviation (SD), and counts of participants receiving relatively bad advice after a relatively good advice, separated into countries, source of advice, and performance feedback.

		Country	2	3	4	5	6	7	8	9
Human	No performance feedback	Mean WOA	38, 9	75, 7	33, 3	89, 9	20, 3	25, 7	100, 0	66, 7
		SD	26, 4	n.a.	57, 7	15, 5	27, 2	19, 5	n.a.	57, 7
		n	28	1	3	3	3	25	1	3
	Continuous performance feedback	Mean WOA	68, 0	75, 5	0, 0	53, 3	67, 8	31, 8	32, 9	100, 0
		SD	23, 8	37, 6	0, 0	40, 0	44, 3	22, 6	46, 3	n.a.
		n	31	4	3	5	5	37	2	1
Algorithm	No performance feedback	Mean WOA	51, 3	45, 0	50, 0	38, 0	49, 5	32, 9	56, 1	0, 0
		SD	29, 3	n.a.	70, 7	32, 3	36, 6	30, 2	n.a.	0, 0
		n	34	1	2	9	5	32	1	2
	Continuous Performance Feedback	Mean WOA	62, 8	31, 6	33, 3	77, 7	81, 8	41, 6	78, 9	n.a.
		SD	32, 7	n.a.	57, 7	18, 7	25, 7	30, 8	29, 9	n.a.
		n	25	1	3	4	6	29	2	0

3.3.2.5 | Does Algorithmic Advice Increase the Usage of Bad Advice? (Hypothesis 6). Hypothesis 6 states that participants use relatively bad advice more when it comes from an algorithm and no relatively bad advice was revealed previously. Analogously to Hypothesis 3, we select only cases where relatively bad advice was provided after a relatively good advice ($n = 311$, 155 cases in human advisor treatment, 156 cases in algorithmic advisor treatment) (see Table 2).

Again, we conduct a linear mixed model analysis to examine the influence of source of advice on the weight on relatively bad advice. The fixed effects of the model included the main effect of source of advice as well as the effects of country, age, and gender. The remaining model parameters were the same as above.

The effect of source of advice, however, is not statistically significant ($t = 1.470$, $b = 5.600$, $p = 0.143$), indicating that relatively bad algorithmic advice is tendentially, but not significantly, used more than relatively bad human advice. This result does not support Hypothesis 6.

3.4 | Discussion

3.4.1 | The Influence of Performance Feedback on Advice Usage

We provide empirical evidence that the provision of performance feedback influences subsequent advice usages in repeated interactions within a JAS. That is, performance feedback on relatively good advice increases subsequent advice usages and decreases advice usage after a previous relatively bad advice. This finding supports the suggestions derived from EDT (Oliver 1977, 1980). There is also some indication that the positive effect of revealing

relatively good advice quality diminishes with an increasing number of relatively bad advices. This might indicate learning effects that might result of an increasing skepticism towards the advisor. To further analyze on potential learning effects, we call for longitudinal studies that investigate the interaction within a JAS.

Nevertheless, an adaption of advice usage to performance feedback is a delayed reaction to a decision made in the past and might be notoriously unhelpful in forecasting tasks. A positive influence of performance feedback on subsequent advice usages is useful if the advisor outperforms human decision-makers. However, decision-makers might get increasingly vulnerable in situations where the advice is relatively bad or even flawed. Then, the positive influence of performance feedback results in a higher usage of relatively bad advice compared to decision-makers who did not receive any performance feedback that in turn tendentially decreases their final forecast quality. In the end, providing performance feedback represents a trade-off between increasing relatively good advice usage and decreasing performance due to the usage of relatively bad advice and developers of forecasting systems need to consider this undesirable side effect. In addition, the performance feedback provided in this laboratory setting was kept very simple and included only initial and final APE of the participant as well as the APE of the advisor's forecast. Nevertheless, in a more realistic setting, performance feedback may not include all of these components. In combination with a larger time lag between making the forecast and getting performance feedback, the hindsight bias (Hawkins and Hastie 1990) might become a relevant issue that could distort the findings of this laboratory experiment. This study is among the first that points to this drawback of performance feedback that might be an indication for the potential risk of overtrust, that is, to an erosion of a critical verification of advice quality, due to performance feedback. Further research on this is urgently needed.

The effect of performance feedback is particularly problematic in the phase of implementation of a new forecasting system. In this early stage, it is important to verify the quality of advice and to override it if deemed necessary. Thus, designers of forecasting systems should take caution whether and how to provide performance feedback. Merely providing performance feedback does not help to calibrate advice usage to relative advice quality but helps to increase advice usage instead or results in delayed but unhelpful reactions to relatively bad advice in the past. Providing a summary performance feedback could represent one way to mitigate these unappreciated delayed reactions.

3.4.2 | The Influence of Source of Advice on Advice Usage

The results indicate tendencies of algorithm appreciation and show parallels to the study of Prahl and Van Swol (2017) who do not find a significant difference between algorithmic and human advisors when implementing relatively bad advice. Even though the higher performance expectations facet of the PAS is only weakly supported, the finding indicates a bias in favor of algorithmic advisors that is in line with the findings in the algorithm appreciation literature (Logg, Minson, and Moore 2019). Analogously, the all-or-none facet of the PAS seems to be very weak for forecasting algorithms. Again, we do observe a weakly harsher decrease of algorithmic advice usage after revealing relatively bad advice (compared to human advice usage).

In addition, our results are in contradiction with the findings of Dietvorst, Simmons, and Massey (2015) who find algorithm aversion when seeing algorithms err. One main explanation for these diverging results might come from the definition of an error. As errors are inevitable in forecasting tasks (Dietvorst and Bharti 2020), but yet not further specified, we follow the definition of Daschner and Obermaier (2022). We thus define erroneous advice, that is, bad advice, as an advice that increases the decision-maker's initial forecasting error when using. This operationalization seems to be different from the study of Dietvorst, Simmons, and Massey (2015) and might influence the findings. Another explanation seems to be the different type of algorithmic advisors. Dietvorst, Simmons, and Massey (2015) applied a performative algorithm, that is, algorithms that directly make decisions without human intervention. In contrast, we apply an advisory algorithm, that is, algorithms that provide advice to human decision-makers (see Jussupow, Benbasat, and Heinzl 2020 for this type differentiation). It seems that decision-makers have higher sensitivity to errors of performative algorithms.

Furthermore, our results deviate from the findings of the algorithm aversion literature (e.g., Önkal et al. 2009; Longoni, Bonezzi, and Morewedge 2019). Research already documented that the usage of algorithmic advice is task-dependent (Castelo, Bos, and Lehman 2019). Our findings in business forecasts might thus not be generalizable to other forecasting domains such as medical forecasts (Longoni, Bonezzi, and Morewedge 2019). Also, the framing of the advisor might impact the usage. For example, Önkal et al. (2009) framed the human advisor as a financial expert, whereas we framed it as a colleague from the same company. However, this framing was deliberately chosen because the value of high expertise in forecasting is limited (Armstrong 1980) and can distort advice usage due to responsibility shifting strategies.

The PAS is a finding in the trust in automation research that predominantly applies deterministic algorithms, such as a simple signal detection software (Madhavan and Wiegmann 2007; Dzindolet et al. 2003; Yu et al. 2019). It suggests an all-or-none logic, that is, the algorithm is either perfect or flawed with errors that could recur in the future interaction. This logic of the PAS initially seemed to hold for forecasting algorithms (Prahl and Van Swol 2017). However, yet there is first evidence that this facet is not generalizable (see also Daschner and Obermaier 2022). This study gives evidence that this schema is present but weakened. The probabilistic nature and the ability of forecasting algorithms to learn from mistakes might erode the all-or-none logic and could increase the acceptance of algorithmic errors, just like for human errors (Berger et al. 2021). This perception of self-learning algorithms might be further boosted by the increasing use of machine learning algorithms in private and professional life (Jordan and Mitchell 2015). As our experiment is

conducted with university students, new course offerings such as machine learning regression models and data analytics could have enhanced the understanding of the functionality of forecasting algorithms. Yet, more research is required to investigate whether this finding is generalizable to professional decision-makers. Particularly, we call for further research on the effect of timing and frequency of relatively bad advice and the duration of interaction on advice usage. The timing of the relatively bad advice might have a central influence on subsequent advice usages. Future research could apply advisors that are in practical application. Also, it could examine differences between individual characteristics such as gender, age, or cultural background.

To conclude, this study highlights the risk of using relatively bad advice when implementing algorithmic advisors in forecasting. Experiment 1 shows that performance feedback and algorithmic source of advice increase advice usage but do not necessarily calibrate it to relative advice quality. Yet, literature on overtrust in forecasting tasks is scarce. Confronted with the trend towards boosting algorithmic advice usage, we point to its importance in this task domain.

4 | Experiment 2: Representation of Uncertainty Information

Experiment 1 demonstrated that decision-makers use relatively bad advice, particularly when performance feedback on relatively good advice has been provided previously. We now investigate how to improve the calibration of advice usage towards relative advice quality in a forecasting task.

4.1 | Theoretical Background and Hypotheses

There is inherent uncertainty in forecasting tasks and consequently in forecasting advice as well (Winkler 2015; Zhou et al. 2017). Therefore, representing inherently uncertain advice with uncertainty information seems to be more congruent than a point forecast. According to this congruence principle (Du et al. 2011), decision-makers prefer an advice if its representation matches with the nature of the underlying uncertainty about the future. Uncertainty in forecasts is often represented as CI in prior research (Greis et al. 2016; Zhou et al. 2017) or as standard deviations of the mean (e.g., six sigma). Nevertheless, prior research shows that decision-makers often misinterpret CI (see Padilla, Kay, and Hullman 2020 for an overview). A main reason is that decision-makers have general difficulties interpreting probabilities and struggle in correct understanding (Belia et al. 2005; Hoekstra et al. 2014).

Forecast intervals provide additional decision-relevant information (Leffrang and Müller 2021) and might help to evaluate advice quality *ex ante* at the time of decision. Prior studies indicate that the representation of a forecast interval can increase advice usage (Fernandes et al. 2018; Roulston et al. 2006; Joslyn and LeClerc 2012; Greis et al. 2016). Yet, literature is silent on whether it also increases the usage of relatively bad advice in forecasting tasks. Additionally, not every representation of uncertainty information is useful for the decision-maker. Less informative intervals might not increase advice usage (Yaniv and Foster 1995). The informativeness of a forecast interval thereby depends on its confidence level or width (narrow vs. wide). Advice represented with narrow forecast intervals is perceived as competent and in turn might be used more than a point forecast (Du et al. 2011).

There is evidence in highly automated decision tasks that the representation of uncertainty information helps to improve final decision accuracy (Dzindolet et al. 2003; Lee and See 2004; Wang, Jamieson, and Hollands 2009; Mercado et al. 2016). This suggests that relatively bad advice might be used less when representing advice with narrow forecast intervals compared to a point forecast. Reversely, relatively good advice might be used more when representing advice with narrow forecast intervals compared to a point forecast. Consequently, it seems to be one promising way to calibrate advice usage to relative advice quality. We follow this research and investigate whether this finding is generalizable to forecasting tasks and follow Daschner and Obermaier (2022) to explicitly consider relative advice quality. We therefore hypothesize:

Hypothesis 7a. *When advice is relatively good, representing algorithmic advice as a narrow forecast interval results in a higher advice usage than a point forecast.*

Hypothesis 7b. *When advice is relatively bad, representing algorithmic advice as a narrow forecast interval results in a lower advice usage than a point forecast.*

Forecast intervals can also be wide. If uncertainty of a forecast advice is perceived as high, that is, the interval of the forecast is wide, advice usage even decreases compared to a point forecast (Du et al. 2011). Then, the informativeness of the forecast decreases. Less informative forecasts are, in turn, less appreciated by decision-makers (Yaniv and Foster 1995; Yaniv 1997; Goodwin, Gönül, and Önkal 2013). We hypothesize:

Hypothesis 8. *Representing algorithmic advice as a wide forecast interval results in lower advice usage compared to a point forecast, irrespective of relative advice quality.*

4.2 | Method

4.2.1 | Task and Procedure

We conducted a 4×2 mixed design experiment with 221 online participants (116 women; 99 men; 6 diverse; mean age = 27.81 with an SD of 7.91 years) recruited via the platform Prolific with uncertainty representation of advice (point forecast vs. 70% CI vs. 99% CI vs. best/worst case) as between-subjects factor and relative advice quality (good vs. bad) as within-subjects factor. Each participant was paid a predetermined amount of £1.50 after completing the survey. All participants were obliged to provide written informed consent prior to the experiment. They were allowed to withdraw at any time and without giving a reason. The average duration of the experiment was 11 min and 10 s. Figure 7 summarizes the experimental procedure (Data S2).

The task was similar to the task in Experiment 1 except three important changes. First, the historical data were represented

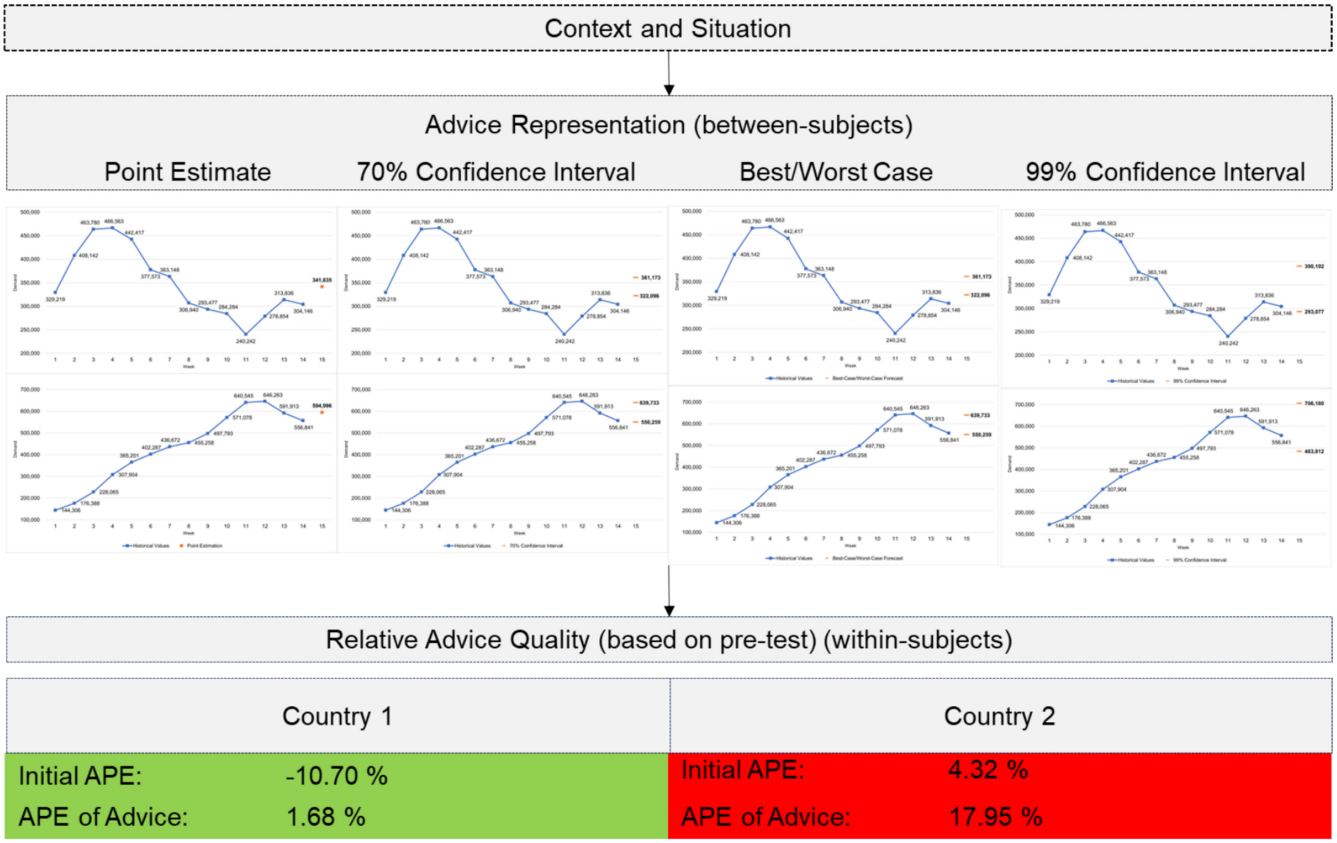

FIGURE 7 | Experimental procedure with advice representation as between-subjects and relative advice quality as within-subjects factor. The majority of the pretest participants deteriorated their final APE by advice usage only in Country 2.

as a line graph because it facilitates the identification of a trend (Washburne 1927) and might be a common visualization format in forecasting tasks. Second, participants only had to make two subsequent forecasts instead of nine (within-subjects factor). The Countries 1 and 2 of Experiment 1 were applied here again. Third, the advice in Countries 1 and 2 is based on the Excel forecasting model called "Forecast Sheet" as Excel might be a commonly used software of decision-makers. Its forecast is very close to the forecast based on the ARIMA model in Experiment 1 and was applied to reduce the percentage error of the advice in Country 1 and to increase it in Country 2. Comparisons of relative advice quality with pretest results mentioned in Experiment 1 indicate relatively good advice in Country 1 and relatively bad advice in Country 2.

4.2.2 | Independent and Dependent Variables

Advice representation is an independent variable. To generate a significant difference in the width of the CI, we choose a 70% confidence level for the narrow, informative forecast interval (approximately one standard deviation of the mean) and a 99% confidence level for a wide, relatively less informative CI (approximately three standard deviations of the mean).

We follow prior literature (Goodwin, Gönül, and Önkal 2013) and label the upper bound of the 70% CI as best case and the lower bound of the 70% CI as worst case. This reframing of the same information should be jargon-free (Goodwin, Gönül, and Önkal 2013). We therefore also examine an additional representation format of uncertainty information that does not disclose the quantitative information included in a CI. In sum, the representation of uncertainty information is manipulated as between-subjects factor (point forecast vs. 70% CI vs. 99% CI vs. best/worst case).

As a second independent variable, relative advice quality is manipulated as within-subjects factor (relatively good advice, then relatively bad advice). The initial deviation of the participants' forecast from the actual value of the product demand in Week 15 in the pretest was 10.70%. In comparison, the algorithmic advice provided by the Excel forecasting model had a percentage deviation of 1.68%. Thus, the advice in Country 1 is described as relatively good advice. In the pretest for Country 2, the deviation of the algorithmic advice from the actual product demand value in Week 15 is higher (17.95%) than for the average participant (4.32%). Therefore, advice in Country 2 is referred to as relatively bad advice (see Appendix S3). No performance feedback is provided to avoid any additional distortions.

WOA as the dependent variable remained the same as in Experiment 1. As WOA requires a point forecast, we use the middle of the interval forecasts, assuming symmetry of the advised forecast intervals. Thus, the value of advice is identical for all advice representation treatments. Nevertheless, the provision of forecast intervals complicates the calculation of WOA. As there are a considerable number of WOA values that exceed 1 or are beneath 0, we therefore deviate from prior literature and performed a 90% winsorization, that is, 90% of the data remained unchanged, and observations greater than the 95th percentile are set to the value

at the 95th percentile, and all observations less than the 5th percentile are set to the value at the 5th percentile. In sum, we winsorized 23 and 22 WOA values for Countries 1 and 2, respectively.

4.3 | Results

4.3.1 | Manipulation Check on Relative Advice Quality

In order to ensure that the relative advice quality is in line with the pretest, two separate one-sample Wilcoxon signed rank tests are performed. Accordingly, it can be tested whether the deviation of the algorithmic advice is smaller (larger) than the average deviation of the participants' initial forecast from the actual product demand value in Week 15 in Country 1 (Country 2). In Country 1, the algorithmic deviation (1.68%) is smaller than the participants' initial deviation (Mdn 11.05%, SD ± 6.04%) ($p < 0.001$), that is, the algorithm outperformed the participants in this forecasting task. In Country 2, the participants' forecast (Mdn 7.05%, SD ± 7.31%) was better than the algorithmic advice (17.95%) ($p < 0.001$). The manipulation was successful.

4.3.2 | Descriptive Statistics

Table 3 shows the mean WOAs in each country, separated into treatments of advice representation.

4.3.3 | Hypothesis Tests

We run a two-way mixed ANOVA to examine the influence of advice representation and relative advice quality (good vs. bad) on WOA. There is a statistically significant interaction between advice representation and relative advice quality, $F_{3,217} = 3.676$, $p = 0.013$, partial $\eta^2 = 0.048$. Pairwise comparisons in Country 1 (relatively good advice) show a mean difference in WOA of 3.52% between a point forecast and a 70% CI ($p = 1.000$, 95% CI [−15.15%; 22.19%]) and a mean difference of −4.98% between a point forecast and a best/worst case scenario ($p = 1.000$, 95% CI [−25.33%; 15.36%]). Both differences are non-significant. Thus, we do not find support for Hypothesis 7a.

Analogously, pairwise comparisons in Country 2 (relatively bad advice) show a non-significant mean difference in WOA of 21.05% between a point forecast and a 70% CI ($p = 0.158$, 95% CI [−4.02%; 46.11%]), indicating tendentially lower WOAs in the latter advice representation group. Interestingly, there is a significant mean difference in WOA of 29.50% between a point forecast and a best/worst case scenario ($p = 0.027$, 95% CI [2.19%; 56.82%]) (see Figure 8). This result provides weak support for Hypothesis 7b.

We further examine on the difference between point forecast and 99% CI. The results of the two-way mixed ANOVA indicate significantly lower WOAs in Country 1 (mean difference of 38.50%, $p < 0.001$, 95% CI [19.16%; 57.84%]) and in Country 2 (mean difference of 63.65%, $p < 0.001$, 95% CI [37.69%; 89.62%]) (see Figure 8). The results support Hypothesis 8.

In a supplementary analysis, in the point forecast group, WOA is not statistically significantly different between relatively good and bad advice ($p = 0.758$). In contrast, in the 70% CI group, WOA in Country 2 (relatively bad advice) is statistically significantly reduced (mean = −15.8%, $p = 0.009$). Analogously, in the best/worst case scenario group, we also find a significant decrease in WOA (mean = −32.7%, $p < 0.001$). This finding further supports the suggestion that providing uncertainty information helps to calibrate advice usage to relative advice quality.

We further run a two-way mixed ANOVA with the same independent variables on final APE. Table 4 depicts the respective descriptive statistics. There is a statistically significant interaction between advice representation and relative advice quality, $F_{3,217} = 17.172$, $p < 0.001$, partial $\eta^2 = 0.192$. Pairwise comparisons in Country 1 (relatively good advice) show a significantly higher final APE in the 99% CI group compared to a point forecast (mean difference = 6.09%, $p < 0.001$), the 70% CI (mean difference = 4.99%, $p < 0.001$), and the best/worst case scenario (mean difference = 5.04%, $p < 0.001$). Vice versa, the final APE in the 99% CI group is lower in Country 2 compared to a point forecast (mean difference = −4.65%, $p = 0.002$), however, not significantly lower compared to the 70% CI group (mean difference = −2.78%, $p = 0.140$) and the best/worst case scenario group (mean difference = −2.53%, $p = 0.349$).

TABLE 3 | Mean WOA for the respective uncertainty representations of advice. Error bars indicate the 95% confidence interval.

Advice representation	*N*	WOA 1 (Good advice)	WOA 2 (Bad advice)	Delta
Point forecast	54	55.34% (34.41%)	57.08% (35.00%)	1.74% (0.59%)
70% confidence interval	65	51.83% (38.58%)	36.04% (46.39%)	−15.79% (7.81%)
99% confidence interval	56	16.85% (39.52%)	−6.57%(68.53%)	−23.41% (29.01%)
Best/worst case	46	60.33% (39.65%)	27.58% (48.59%)	−32.75% (8.94%)
Total	221	45.59% (41.48%)	28.62% (55.76%)	−16.97% (14.28%)

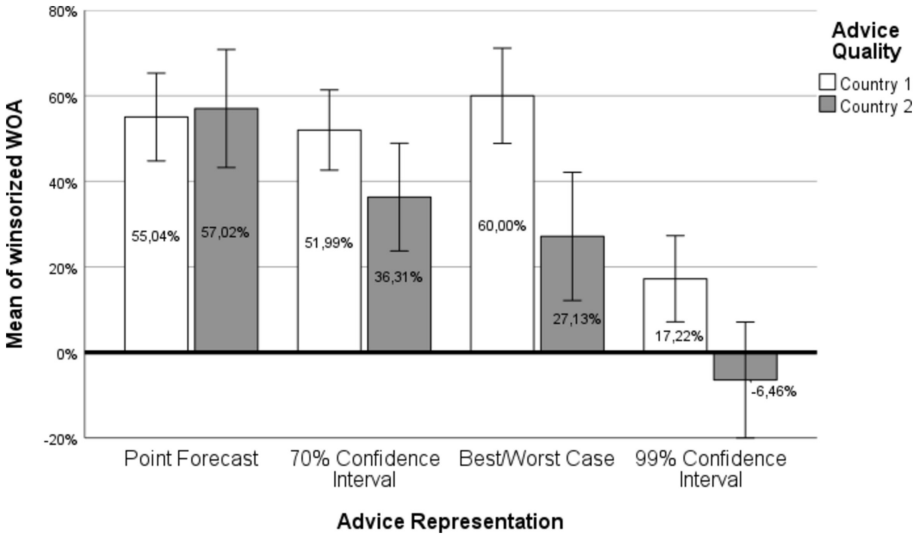

FIGURE 8 | Mean WOAs separated into groups of advice representation and relative advice quality. Error bars indicate the 95% confidence interval.

TABLE 4 | Descriptive statistics on final APE, separated into advice representations and relative advice quality.

Final APE		Country 1(relatively good advice)	Country 2(relatively bad advice)
Point forecast	Mean	4.61%	13.49%
	Median	4.47%	13.99%
	SD	3.74%	4.82%
	n	54	54
70% confidence interval	Mean	5.70%	11.61%
	Median	4.17%	9.09%
	SD	4.55%	7.09%
	n	65	65
99% confidence interval	Mean	10.69%	8.83%
	Median	11.16%	5.07%
	SD	6.72%	8.51%
	n	56	56
Best/worst case	Mean	5.65%	11.36%
	Median	4.17%	10.17%
	SD	4.47%	5.23%
	n	46	46

In general, pairwise comparisons between Country 1 and Country 2 show that the final APE of a decision-maker increases the most in the case of a point forecast advice (mean difference = 8.88%, $p < 0.001$), followed by the narrow forecast interval treatments (70% CI with a mean difference of 5.91%, $p < 0.001$; best/worst case scenario with a mean difference of 5.72%, $p < 0.001$). Only the 99% CI (wide forecast interval) even reduces the final APE marginally with a mean difference of −1.86%, $p = 0.093$.

4.4 | Discussion

The results of this experiment demonstrate that the representation of uncertainty information affects advice usage in forecasting tasks. This effect depends on the informativeness of uncertainty information. Even though a 70% CI and a 99% CI represent in general the same advice, the representation of a (visually) wider forecast interval decreases advice usage. It might be perceived as too uninformative, even if the confidence level is

then accordingly higher (Uggirala et al. 2004; Yaniv 1997). This 99% CI mitigates the usage of relatively bad advice, but simultaneously, it reduces the usage of relatively good advice as well. Consequently, a calibration is not achieved by too uninformative forecast intervals. Nevertheless, the calculation of WOA requires this centered value and reaches its limits when analyzing forecast intervals. The chosen approach to middle the forecast intervals results in an equal baseline for all treatments; however, it does not account for the participant's understanding of range estimates. We follow prior literature (Goodwin, Gönül, and Önkal 2013) and did not represent the centered advice to avoid any anchoring effects. However, little is known about the influence of the anchoring effect in advice taking (e.g., Schultze, Mojzisch, and Schulz-Hardt 2017). Accordingly, it is difficult to rule out that the respective lower and upper bounds of uncertain range advice were taken as advice. We further calculated WOA based on the lower and upper bounds to validate our insights. The analyses in general support the potential of forecast intervals to calibrate usage to relative advice quality and can be found in Appendix S4.

The results of Experiment 2 provide some indications that advices with a narrow forecast interval, particularly with a best/worst case framing, help the decision-maker to reduce the usage of relatively bad advice. The difference between the usage of relatively good and bad advice is greater than when a point forecast advice is given, but the final APE also tends to be lower when relatively bad advice is given. In the case of wide forecast intervals, the usage of advice is generally lower compared to the other advice representations and seems to be independent of the relative advice quality. This explains why there is even a decrease in the final APE between relatively good and bad advice. Too wide forecast intervals could be interpreted as a sign of incompetence. The findings are consistent with previous literature on the representation of uncertainty information, which found that transparent communication of system uncertainty at least partially leads to calibration of trust (Mercado et al. 2016; Dzindolet et al. 2003; Lee and See 2004; Wang, Jamieson, and Hollands 2009). These results seem to be generalizable to forecasting tasks. The provision of forecast intervals seems to cause the participants to be more careful about advice usage, interestingly, only in case of relatively bad advice (see also Mercado et al. 2016). This finding does not imply that decision-makers reduce advice usage in the case of relatively good advice.

In addition, we find no statistically significant difference between labeling a forecast interval as either a 70% CI or as best/worst case. This finding contradicts the suggestion of Goodwin, Gönül, and Önkal (2013) that this reframing of the same uncertainty information may be perceived as more competent. In both cases, these narrow, informative forecast intervals help to reduce the usage of relatively bad advice by calibrating advice usage to relative advice quality.

5 | General Discussion and Conclusion

In Experiment 1 of this study, we address the influence of performance feedback and source of advice on advice usage in repeated interactions while considering relative advice quality. On average, advice usage rates are between 40% and 50%. This reflects the wish of decision-makers to improve their initial judgment (Harvey and Fischer 1997). Due to advice discounting or overconfidence of decision-makers, we might have observed advice usage rates below 50% (Harvey and Fischer 1997; Yaniv and Kleinberger 2000; Wiczorek and Meyer 2019). We also find empirical evidence that the provision of performance feedback on good advice increases subsequent advice usage, whereas performance feedback on relatively bad advice reduces it. In addition, it seems that the positive effect of performance feedback after relatively good advice decreases with an increasing number of revealed relatively bad advices. Thus, we address the call for further research to better understand reactions to relatively bad advice in forecasting (Prahl and Van Swol 2017).

Furthermore, this study is particularly concerned with the risk of using relatively bad advice in forecasting tasks. Our findings indicate higher usage of relatively bad advice when performance feedback is provided (compared to no performance feedback). The design and application of forecasting systems therefore require the consideration of both the upside potentials and the downside risk of providing performance feedback. This study therefore highlights the drawback of recent research efforts on how to increase advice usage (Önkal, Gönül, and de Baets 2019). In addition to ethnic decision problems (e.g., Krügel, Ostermaier, and Uhl 2022), our results now provide evidence in a business context that algorithmic advisors might be overtrusted rather than distrusted. Interacting with a highly reliable algorithm increases the risk of overtrust, which is commonly associated with negative impacts on decision quality (Rieger and Manzey 2022a, 2022b; Bahner 2008; Mosier et al. 1996, 1997, 2001). Overtrust could jeopardize the purpose of the human-in-the-loop, which is to override advice where necessary. Instead, the human could degenerate to a "zombie" (Krügel, Ostermaier, and Uhl 2022). Therefore, the maintenance of a critical human mind within the human–algorithm dyad is essential to avoid excessive vulnerability when advisor systems perform poorly. To the best of our knowledge, this study is among the first that investigates the risk of using relatively bad advice as a potential antecedent of overtrust in the field of forecasting.

Experiment 2 explores whether the representation of uncertainty information (advice representation) helps to calibrate advice usage to relative advice quality in order to reduce the usage of relatively bad advice. Our findings suggest that an informative forecast interval can mitigate the usage of relatively bad advice in forecasting tasks. However, they also point to the importance of an appropriate representation of uncertainty information, as wide forecast intervals seem to be little informative and result in a general decrease in advice usage. In general, our findings lend support to the congruence principle (Du et al. 2011). Moreover, the provision of forecast intervals can contribute to fading the all-or-none notion decision-makers have of algorithms applied in automation. This could mitigate the effect of the algorithmic origin of advice on the usage of relatively bad advice as examined in Experiment 1.

This study does not come without some limitations. First, the order of countries was not randomized so that we cannot rule out any order effects. Second, the frequency of relatively bad advice provided in Experiment 1 might confound the effect of relative advice quality. As we follow prior literature investigating the threshold between good and bad advice (Daschner and Obermaier 2022), we cannot fully control the frequency of relatively bad advice. Third, we operationalized the width of forecast

interval by choosing confidence levels at which a difference in the width is easy to perceive by the participant. However, the threshold at which advice becomes less informative is a matter for future research. Fourth, it should also be noted that the results of this study are based on a short-term interaction. However, we find indications that the cumulated experience with the advisor influences subsequent advice usage. It is questionable whether the observed effects also hold in a longitudinal experiment. For example, this study provides indications that in the beginning of an interaction, performance feedback increases advice usage in subsequent advices, which, however, diminishes with an increasing number of relatively bad advices. Longitudinal studies are required to examine whether the provision of performance feedback is really a problem. Finally, we applied a business forecast scenario where no severe consequences of using relatively bad advice were implemented in the experiment. This lack of consequences might influence the behavior of the participants and might not be representative for real-word forecasting problems. Therefore, future research could address whether the findings of this study are generalizable to other forecasting domains. For example, a relatively bad advice in the medical domain might have much more severe consequences. Reactions to relatively bad advice might deviate from reactions in the business context. Also, algorithm aversion is particularly documented in these highly consequential task domains (e.g., Longoni, Bonezzi, and Morewedge 2019) and questions the generalizability of the algorithm appreciation tendency observed in this study.

As practical implications, the representation of uncertainty information is indeed a potential countermeasure to reduce the usage of relatively bad advice in forecasting tasks. Forecast intervals are one option to represent inherent uncertainty; however, they must be informative to decision-makers. Lower CI convey advice in narrow intervals and thus help to calibrate advice usage to relative advice quality. In practice, developers of forecasting systems should provide informative uncertainty information to decision-makers and consider an appropriate width of a CI for the algorithmic advice. The conveyance of uncertainty information has also implications for the provision of performance feedback, widening the range of performance feedback options. For example, the decision system can inform decision-makers about the hit rate, that is, the share of forecast intervals that has included the true value. The appropriate provision of algorithmic and human performance feedback is an issue that requires further exploration.

Acknowledgments

Open Access funding enabled and organized by Projekt DEAL.

Data Availability Statement

The data that support the findings of this study are openly available in Figshare at https://figshare.com/articles/journal_contribution/Do_we_use_Bad_Algorithmic_Advice_/25765023.

References

Agrawal, A., J. Gans, and A. Goldfarb. 2018. *Prediction Machines: The Simple Economics of Artificial Intelligence*. Boston, MA: Harvard Business Review Press.

Armstrong, J. S. 1980. "The Seer-Sucker Theory: The Value of Experts in Forecasting." *Technology Review* 82, no. 7: 16–24.

Bahner, J. E. 2008. "Übersteigertes Vertrauen in automation: Der Einfluss von Fehlererfahrungen auf complacency und automation bias [Overtrust in Automation: The Influence of Error Experiences on Complacency and Automation Bias]." Doctoral diss., Technische Universität Berlin.

Barber, B. 1983. *The Logic and Limits of Trust*. New Brunswick, NJ: New Rutgers University Press.

Belia, S., F. Fidler, J. Williams, and G. Cumming. 2005. "Researchers Misunderstand Confidence Intervals and Standard Error Bars." *Psychological Methods* 10, no. 4: 389–396.

Berger, B., M. Adam, A. Rühr, and A. Benlian. 2021. "Watch Me Improve—Algorithm Aversion and Demonstrating the Ability to Learn." *Business & Information Systems Engineering* 63, no. 1: 55–68.

Bhattacherjee, A., and G. Premkumar. 2004. "Understanding Changes in Belief and Attitude Toward Information Technology Usage: A Theoretical Model and Longitudinal Test." *MIS Quarterly* 28, no. 2: 229–254.

Billings, C. E. 1991. "Human-centered Aircraft Automation: A Concept and Guidelines." National Aeronautics and Space Administration, Ames Research Center.

Castelo, N., M. W. Bos, and D. R. Lehman. 2019. "Task-Dependent Algorithm Aversion." *Journal of Marketing Research* 56, no. 5: 809–825. https://doi.org/10.1177/0022243719851788.

Chacon, A., E. E. Kausel, and T. Reyes. 2022. "A Longitudinal Approach for Understanding Algorithm Use." *Journal of Behavioral Decision Making* 35, no. 4: e2275.

Daschner, S., and R. Obermaier. 2022. "Algorithm Aversion? On the Influence of Advice Accuracy on Trust in Algorithmic Advice." *Journal of Decision Systems* 31: 1–21.

Dawes, R. M., D. Faust, and P. E. Meehl. 1989. "Clinical Versus Actuarial Judgment." *Science* 243, no. 4899: 1668–1674.

De Baets, S., and N. Harvey. 2020. "Using Judgment to Select and Adjust Forecasts From Statistical Models." *European Journal of Operational Research* 284, no. 3: 882–895.

Deutsch, M. 1958. "Trust and Suspicion." *Journal of Conflict Resolution* 2, no. 4: 265–279.

Dietvorst, B. J., and S. Bharti. 2020. "People Reject Algorithms in Uncertain Decision Domains Because They Have Diminishing Sensitivity to Forecasting Error." *Psychological Science* 31, no. 10: 1302–1314.

Dietvorst, B. J., J. P. Simmons, and C. Massey. 2015. "Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err." *Journal of Experimental Psychology: General* 144, no. 1: 114–126.

Dietvorst, B. J., J. P. Simmons, and C. Massey. 2018. "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms if They Can (Even Slightly) Modify Them." *Management Science* 64, no. 3: 1155–1170. https://doi.org/10.1287/mnsc.2016.2643.

Du, N., D. V. Budescu, M. K. Shelly, and T. C. Omer. 2011. "The Appeal of Vague Financial Forecasts." *Organizational Behavior and Human Decision Processes* 114, no. 2: 179–189.

Dzindolet, M. T., S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck. 2003. "The Role of Trust in Automation Reliance." *International Journal of Human-Computer Studies* 58, no. 6: 697–718.

Dzindolet, M. T., L. G. Pierce, H. P. Beck, and L. A. Dawe. 2002. "The Perceived Utility of Human and Automated Aids in a Visual Detection Task." *Human Factors* 44, no. 1: 79–94.

Endsley, M. R., and E. O. Kiris. 1995. "The Out-of-the-Loop Performance Problem and Level of Control in Automation." *Human Factors* 37, no. 2: 381–394.

Fernandes, M., L. Walls, S. Munson, J. Hullman, and M. Kay. 2018. "Uncertainty Displays Using Quantile Dotplots or cdfs Improve Transit Decision-Making." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12.

Fildes, R., and F. Petropoulos. 2015. "Improving Forecast Quality in Practice." *Foresight: The International Journal of Applied Forecasting* 36: 5–12.

Goodwin, P., M. S. Gönül, and D. Önkal. 2013. "Antecedents and Effects of Trust in Forecasting Advice." *International Journal of Forecasting* 29, no. 2: 354–366.

Greis, M., P. E. Agroudy, H. Schuff, T. Machulla, and A. Schmidt. 2016. "Decision-Making Under Uncertainty: How the Amount of Presented Uncertainty Influences User Behavior." In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, 1–4.

Grove, W. M., D. H. Zald, B. S. Lebow, B. E. Snitz, and C. Nelson. 2000. "Clinical Versus Mechanical Prediction: A Meta-Analysis." *Psychological Assessment* 12, no. 1: 19–30.

Gunaratne, J., L. Zalmanson, and O. Nov. 2018. "The Persuasive Power of Algorithmic and Crowdsourced Advice." *Journal of Management Information Systems* 35, no. 4: 1092–1120.

Hardin, R. 1993. "The Street-Level Epistemology of Trust." *Politics and Society* 21, no. 4: 505–529.

Harvey, N., and I. Fischer. 1997. "Taking Advice: Accepting Help, Improving Judgment, and Sharing Responsibility." *Organizational Behavior and Human Decision Processes* 70, no. 2: 117–133.

Harvey, N., and I. Fischer. 2005. "Development of Experience-Based Judgment and Decision Making: The Role of Outcome Feedback." In *The Routines of Decision Making*, edited by T. Betsch and S. Haberstroh, 119–137. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Hawkins, S. A., and R. Hastie. 1990. "Hindsight: Biased Judgments of Past Events After the Outcomes Are Known." *Psychological Bulletin* 107, no. 3: 311–327. https://doi.org/10.1037/0033-2909.107.3.311.

Himmelstein, M., and D. V. Budescu. 2022. "Preference for Human or Algorithmic Forecasting Advice Does Not Predict if and How It Is Used." *Journal of Behavioral Decision Making* 36, no. 1: e2285.

Hoekstra, R., R. D. Morey, J. N. Rouder, and E. J. Wagenmakers. 2014. "Robust Misinterpretation of Confidence Intervals." *Psychonomic Bulletin & Review* 21: 1157–1164.

Hoff, K. A., and M. Bashir. 2015. "Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust." *Human Factors* 57, no. 3: 407–434.

Hogarth, R. M., and S. Makridakis. 1981. "Forecasting and Planning: An Evaluation." *Management Science* 27, no. 2: 115–138.

Iansiti, M., and K. R. Lakhani. 2020. *Competing in the Age of AI: Strategy and Leadership When Algorithms and Networks Run the World*. MA: Harvard Business Review Press.

Jordan, M. I., and T. M. Mitchell. 2015. "Machine Learning: Trends, Perspectives, and Prospects." *Science* 349, no. 6245: 255–260. https://doi.org/10.1126/science.aaa8415.

Joslyn, S. L., and J. E. LeClerc. 2012. "Uncertainty Forecasts Improve Weather-Related Decisions and Attenuate the Effects of Forecast Error." *Journal of Experimental Psychology: Applied* 18, no. 1: 126–140.

Jussupow, E., I. Benbasat, and A. Heinzl. 2020. "Why Are We Averse Towards algorithms? A Comprehensive Literature Review on Algorithm Aversion."

Kaufmann, E., A. Chacon, E. E. Kausel, N. Herrera, and T. Reyes. 2023. "Task-Specific Algorithm Advice Acceptance: A Review and Directions for Future Research." *Data and Information Management* 7: 100040.

Kaufmann, E., and W. W. Wittmann. 2016. "The Success of Linear Bootstrapping Models: Decision Domain-, Expertise-, and Criterion-Specific Meta-Analysis." *PLoS ONE* 11, no. 6: e0157914.

Kramer, R. M. 2010. "Trust Barriers in Cross-Cultural Negotiations: Psychological Analysis." In *Organizational Trust: A Cultural Perspective*, edited by M. N. K. Saunders, D. Skinner, G. Dietz, N. Gillespie, and R. J. Lewicki, 182–204. UK: Cambridge University Press.

Krügel, S., A. Ostermaier, and M. Uhl. 2022. "Zombies in the Loop? Humans Trust Untrustworthy AI-Advisors for Ethical Decisions." *Philosophy and Technology* 35, no. 1: 17. https://doi.org/10.1007/s13347-022-00511-9.

Lee, J. D., and K. A. See. 2004. "Trust in Automation: Designing for Appropriate Reliance." *Human Factors* 46, no. 1: 50–80.

Leffrang, D., and O. Müller. 2021. "Should I Follow This Model? The Effect of Uncertainty Visualization on the Acceptance of Time Series Forecasts." In *2021 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)* (IEEE, 20–26).

Lewicki, R. J., E. C. Tomlinson, and N. Gillespie. 2006. "Models of Interpersonal Trust Development: Theoretical Approaches, Empirical Evidence, and Future Directions." *Journal of Management* 32, no. 6: 991–1022.

Logg, J. M., J. A. Minson, and D. A. Moore. 2019. "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment." *Organizational Behavior and Human Decision Processes* 151: 90–103.

Longoni, C., A. Bonezzi, and C. K. Morewedge. 2019. "Resistance to Medical Artificial Intelligence." *Journal of Consumer Research* 46, no. 4: 629–650.

Madhavan, P., and D. A. Wiegmann. 2007. "Effects of Information Source, Pedigree, and Reliability on Operator Interaction With Decision Support Systems." *Human Factors* 49, no. 5: 773–785.

Mahmud, H., A. N. Islam, S. I. Ahmed, and K. Smolander. 2022. "What Influences Algorithmic Decision-Making? A Systematic Literature Review on Algorithm Aversion." *Technological Forecasting and Social Change* 175: 121390.

Mayer, R. C., J. H. Davis, and F. D. Schoorman. 1995. "An Integrative Model of Organizational Trust." *Academy of Management Review* 20, no. 3: 709–734.

McKnight, D. H., V. Choudhury, and C. Kacmar. 2002. "Developing and Validating Trust Measures for e-Commerce: An Integrative Typology." *Information Systems Research* 13, no. 3: 334–359.

Meehl, P. E. 1954. "Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence."

Mercado, J. E., M. A. Rupp, J. Y. Chen, M. J. Barnes, D. Barber, and K. Procci. 2016. "Intelligent Agent Transparency in Human–Agent Teaming for Multi-UxV Management." *Human Factors* 58, no. 3: 401–415.

Mosier, K. L., L. J. Skitka, M. D. Burdick, and S. T. Heers. 1996. "Automation Bias, Accountability, and Verification Behaviors." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 40, no. 4: 204–208.

Mosier, K. L., L. J. Skitka, S. Heers, and M. Burdick. 1997. "Automation Bias: Decision Making and Performance in High-Tech Cockpits." *International Journal of Aviation Psychology* 8, no. 1: 47–63. https://doi.org/10.1207/s15327108ijap0801_3.

Mosier, K. L., L. J. Skitka, M. Dunbar, and L. McDonnell. 2001. "Aircrews and Automation Bias: The Advantages of Teamwork?." *International Journal of Aviation Psychology* 11, no. 1: 1–14.

Oliver, R. L. 1977. "Effect of Expectation and Disconfirmation on Postexposure Product Evaluations: An Alternative Interpretation." *Journal of Applied Psychology* 62, no. 4: 480–486.

Oliver, R. L. 1980. "A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions." *Journal of Marketing Research* 17, no. 4: 460–469.

Önkal, D., M. S. Gönül, and S. de Baets. 2019. "Trusting Forecasts." *Futures & Foresight Science* 1, no. 3–4: e19.

Önkal, D., P. Goodwin, M. Thomson, S. Gönül, and A. Pollock. 2009. "The Relative Influence of Advice From Human Experts and Statistical Methods on Forecast Adjustments." *Journal of Behavioral Decision Making* 22, no. 4: 390–409.

Padilla, L., M. Kay, and J. Hullman. 2020. "Uncertainty Visualization." In *Handbook of Computational Statistics and Data Science*. Hoboken, NJ: Wiley.

Parasuraman, R., and D. H. Manzey. 2010. "Complacency and Bias in Human use of Automation: An Attentional Integration." *Human Factors* 52, no. 3: 381–410.

Parasuraman, R., and V. Riley. 1997. "Humans and Automation: Use, Misuse, Disuse, Abuse." *Human Factors* 39, no. 2: 230–253.

Pavlou, P. A., and D. Gefen. 2004. "Building Effective Online Marketplaces With Institution-Based Trust." *Information Systems Research* 15, no. 1: 37–59.

Prahl, A., and L. M. Van Swol. 2017. "Understanding Algorithm Aversion: When Is Advice From Automation Discounted?." *Journal of Forecasting* 36, no. 6: 691–702.

Rempel, J. K., J. G. Holmes, and M. P. Zanna. 1985. "Trust in Close Relationships." *Journal of Personality and Social Psychology* 49, no. 1: 95–112.

Renier, L. A., M. Schmid Mast, and A. Bekbergenova. 2021. "To Err Is Human, Not Algorithmic – Robust Reactions to Erring Algorithms." *Computers in Human Behavior* 124: 106879.

Rieger, T., and D. Manzey. 2022a. "Human Performance Consequences of Automated Decision Aids: The Impact of Time Pressure." *Human Factors* 64, no. 4: 617–634.

Rieger, T., and D. Manzey. 2022b. "Understanding the Impact of Time Pressure and Automation Support in a Visual Search Task." *Human Factors* 66: 770–786.

Robinette, P., A. M. Howard, and A. R. Wagner. 2017. "Effect of Robot Performance on Human–Robot Trust in Time-Critical Situations." *IEEE Transactions on Human-Machine Systems* 47, no. 4: 425–436.

Robinette, P., W. Li, R. Allen, A. M. Howard, and A. R. Wagner. 2016. "Overtrust of Robots in Emergency Evacuation Scenarios." In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (IEEE, 101–108).

Roulston, M. S., G. E. Bolton, A. N. Kleit, and A. L. Sears-Collins. 2006. "A Laboratory Study of the Benefits of Including Uncertainty Information in Weather Forecasts." *Weather and Forecasting* 21, no. 1: 116–122.

Salem, M., G. Lakatos, F. Amirabdollahian, and K. Dautenhahn. 2015. "Would You Trust a (Faulty) Robot?." In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, edited by J. A. Adams, W. Smart, B. Mutlu, and L. Takayama, 141–148. New York, NY: ACM.

Schultze, T., A. Mojzisch, and S. Schulz-Hardt. 2017. "On the Inability to Ignore Useless Advice: A Case for Anchoring in the Judge-Advisor-System." *Experimental Psychology* 64, no. 3: 170–183. https://doi.org/10.1027/1618-3169/a000361.

Sniezek, J. A., G. E. Schrah, and R. S. Dalal. 2004. "Improving Judgement With Prepaid Expert Advice." *Journal of Behavioral Decision Making* 17, no. 3: 173–190.

Soll, J. B., and R. P. Larrick. 2009. "Strategies for Revising Judgment: How (and How Well) People Use Others' Opinions." *Journal of Experimental Psychology. Learning, Memory, and Cognition* 35, no. 3: 780–805. https://doi.org/10.1037/a0015145.

Uggirala, A., A. K. Gramopadhye, B. J. Melloy, and J. E. Toler. 2004. "Measurement of Trust in Complex and Dynamic Systems Using a Quantitative Approach." *International Journal of Industrial Ergonomics* 34, no. 3: 175–186.

Wang, L., G. A. Jamieson, and J. G. Hollands. 2009. "Trust and Reliance on an Automated Combat Identification System." *Human Factors* 51, no. 3: 281–291.

Washburne, J. N. 1927. "An Experimental Study of Various Graphic, Tabular, and Textual Methods of Presenting Quantitative Material." *Journal of Educational Psychology* 18, no. 6: 361–376.

Wiczorek, R., and J. Meyer. 2019. "Effects of Trust, Self-Confidence, and Feedback on the Use of Decision Automation." *Frontiers in Psychology* 10: 519.

Winkler, R. L. 2015. "The Importance of Communicating Uncertainties in Forecasts: Overestimating the Risks From Winter Storm Juno." *Risk Analysis* 35, no. 3: 349–353.

Yaniv, I. 1997. "Weighting and Trimming: Heuristics for Aggregating Judgments Under Uncertainty." *Organizational Behavior and Human Decision Processes* 69, no. 3: 237–249.

Yaniv, I., and D. P. Foster. 1995. "Graininess of Judgment Under Uncertainty: An Accuracy-Informativeness Trade-Off." *Journal of Experimental Psychology: General* 124, no. 4: 424–432.

Yaniv, I., and E. Kleinberger. 2000. "Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation." *Organizational Behavior and Human Decision Processes* 83, no. 2: 260–281.

You, S., C. L. Yang, and X. Li. 2022. "Algorithmic Versus Human Advice: Does Presenting Prediction Performance Matter for Algorithm Appreciation?." *Journal of Management Information Systems* 39, no. 2: 336–365.

Yu, K., S. Berkovsky, R. Taib, J. Zhou, and F. Chen. 2019. "Do I Trust My Machine Teammate? An investigation From Perception to Decision." In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 460–468.

Zhou, J., S. Z. Arshad, S. Luo, and F. Chen. 2017. "Effects of Uncertainty and Cognitive Load on User Trust in Predictive Decision Making." In *HumanComputer Interaction–INTERACT 2017: 16th IFIP TC 13 International Conference, Mumbai, India, September 25–29, 2017, Proceedings, Part IV 16*, 23–39. Cham, Switzerland: Springer International Publishing.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.