Watch Me Improve—Algorithm Aversion and Demonstrating the Ability to Learn

Benedikt Berger · Martin Adam · Alexander Rühr · Alexander Benlian

Abstract Owing to advancements in artificial intelligence (AI) and specifically in machine learning, information technology (IT) systems can support humans in an increasing number of tasks. Yet, previous research indicates that people often prefer human support to support by an IT system, even if the latter provides superior performance – a phenomenon called algorithm aversion. A possible cause of algorithm aversion put forward in literature is that users lose trust in IT systems they become familiar with and perceive to err, for example, making forecasts that turn out to deviate from the actual value. Therefore, this paper evaluates the effectiveness of demonstrating an AI-based system's ability to learn as a potential countermeasure against algorithm aversion in an incentive-compatible online experiment. The experiment reveals how the nature of an erring advisor (i.e., human vs. algorithmic), its familiarity to the user (i.e., unfamiliar vs. familiar), and its ability to learn (i.e., non-learning vs. learning) influence a decision maker's reliance on the advisor's judgement for an objective and non-personal decision task. The results reveal no difference in the reliance on unfamiliar human and algorithmic advisors, but differences in the reliance on familiar human and algorithmic advisors that err. Demonstrating an advisor's ability to learn, however, offsets the effect of familiarity. Therefore, this study contributes to an enhanced understanding of algorithm aversion and is one of the first to examine how users perceive whether an IT system is able to learn. The findings provide theoretical and practical implications for the employment and design of AI-based systems.

Keywords Algorithm aversion · Artificial intelligence · Machine learning · Decision support · Advice taking

B. Berger (✉) · A. Rühr
Institute for Information Systems and New Media, Ludwig-Maximilians-Universität München, Ludwigstr. 28, 80539 Munich, Germany
e-mail: benedikt.berger@bwl.lmu.de

A. Rühr
e-mail: ruehr@bwl.lmu.de

M. Adam · A. Benlian
Institute of Information Systems and E-Services, Technical University of Darmstadt, Hochschulstr. 1, 64289 Darmstadt, Germany
e-mail: adam@ise.tu-darmstadt.de

A. Benlian
e-mail: benlian@ise.tu-darmstadt.de

1 Introduction

Artificial intelligence (AI) research has extended the capabilities of information technology (IT) systems to support or automate tasks, such as medical diagnosis, credit card fraud detection, and advertising budget allocation (Anthes 2017). Accordingly, the deployment of AI-based systems, i.e. IT systems employing capabilities developed in AI research, is supposed to change substantially how businesses operate and people work (vom Brocke et al. 2018; Ransbotham et al. 2017). AI researchers employ various approaches to realize new capabilities, yet many promising achievements are based on machine learning (Jordan and Mitchell 2015). Market research companies predict the market for IT systems employing machine learning to grow with double-digit rates over the upcoming

years (Columbus 2020). Regardless of the specific problem domain, machine learning allows equipping IT systems with the ability to learn, i.e. the capability to improve in performance over time (Faraj et al. 2018). Such AI-based systems can assist users in various situations concerning their business and private life (Maedche et al. 2019).

A particularly important machine learning use case is the support of decisions. Decision support systems (DSS) have evolved in several waves in the past and the application of machine learning promises another leap forward (Watson 2017). Yet, existent decision-making and information systems (IS) research suggests that people can be reluctant to accept support from or delegate decisions to DSS – a phenomenon called algorithm aversion (Dietvorst et al. 2015; Castelo et al. 2019).[1] This phenomenon constitutes a serious issue for businesses employing DSS: Since even simple algorithms can outperform humans in many decision tasks (Kuncel et al. 2013; Elkins et al. 2013), rejecting the advice of DSS often leads to inferior decisions. Furthermore, potential gains of combining human and algorithmic insights (Dellermann et al. 2019) cannot be realized if decision makers are unwilling to take algorithmic advice into account.

Previous studies, however, reached conflicting conclusions regarding the conditions under which algorithm aversion emerges (Logg et al. 2019). Specifically, there is a debate about whether people generally prefer human advice to algorithmic advice (Castelo et al. 2019) or whether people's reliance on an algorithm only decreases after becoming familiar with the algorithm, which means to observe its performance (Dietvorst et al. 2015). Existing research has suggested various causes for algorithm aversion (Burton et al. 2020). A prominent idea among those studies is that people are less forgiving toward erring algorithms than toward erring humans (Dietvorst et al. 2015) because people disregard the possibility that algorithms can overcome their shortcomings and grow from them. In the context of decision support, erring means to provide advice that in the end turns out to be not fully accurate as is common for decisions under uncertainty. Accordingly, people tend to rely less on an algorithmic advisor than on a human advisor after becoming familiar with the advisor and observing the advisor to err, even if the erring algorithmic advisor objectively outperforms the erring human advisor. In this study, we juxtapose the two

understandings of algorithm aversion in the context of decision support to answer the following research question:

RQ1 *Do people exhibit a general algorithm aversion or do they prefer human to algorithmic decision support only after observing that the decision support errs?*

If people indeed shun erring algorithmic support because they disregard the possibility that algorithms can improve, demonstrating the opposite (i.e., an algorithm's ability to learn) should alleviate algorithm aversion. However, existing research has not examined whether there are differences in algorithm aversion to DSS with and without the ability to learn. Therefore, we specifically investigate whether demonstrating an algorithm's ability to learn can contribute to overcoming algorithm aversion. We focus on the ability to learn for two reasons: Demonstrating an algorithm's ongoing improvement in performance to users is theoretically intriguing because this design feature may counter users' algorithm aversion and consequently increase their willingness to rely on particular AI-based systems. Moreover, the increasing application of machine learning in practice is especially relevant for tasks that algorithms can support, such as classification or forecasting (Jordan and Mitchell 2015). Therefore, we pose a second research question:

RQ2 *Does demonstrating an algorithm's ability to learn alleviate algorithm aversion?*

To answer our research questions, we conducted an incentive-compatible online experiment with 452 subjects. Within this experiment, participants had to solve a forecasting task while deciding to what degree they would rely on an erring advisor to increase their odds of receiving a bonus. We manipulated the advisor to examine how its nature (i.e., human vs. algorithmic), its familiarity to the participants (i.e., unfamiliar vs. familiar), and its ability to learn (i.e., non-learning vs. learning) affect the participants' reliance on the advice. Our results do not indicate a general aversion to algorithmic advice but a negative effect of familiarity on the participants' willingness to accept algorithmic advice. However, if the algorithm is able to learn, the negative effect of familiarity disappears.

Our study makes a major, threefold contribution to research on algorithm aversion and the interaction with AI-based systems: First, we shed light on the algorithm aversion phenomenon by substantiating that becoming familiar with an erring algorithm is an important boundary condition for this phenomenon. Second, we demonstrate that the experience during the familiarization with an algorithm plays a key role in relying on an algorithm's advice. Third, we provide first insights on an AI-based system's ability to learn as an increasingly important but hitherto underexplored design characteristic, which may counter algorithm

---

[1] We acknowledge that an algorithm is a processing logic to solve a task and that it is, thus, only a part of an IT system. Since previous research has labeled the phenomenon addressed in this study as algorithm aversion although users rather interact with IT systems than with algorithms, we use the terms algorithm and IT system interchangeably in this study. The same applies to the terms algorithmic advisors and DSS.

aversion. Thereby, we answer the call for research on individuals' interaction with AI-based systems (Buxmann et al. 2019). Our findings also hold important implications for the design and employment of continually learning systems. Specifically, developers may seek to emphasize these systems' ability to learn in order to enhance users' tolerance for erring advice and, thus, reliance on support from AI-based systems.

2 Theoretical Foundations

2.1 Algorithm Aversion

The literature on algorithm aversion is rooted in the controversy over the merits of clinical (i.e., based on deliberate human thought) and actuarial (i.e., based on statistic models) judgement in different domains, such as medical diagnosis and treatment (Meehl 1954; Dawes 1979; Dawes et al. 1989; Grove et al. 2000). Overall, this research concludes that actuarial data interpretation is superior to clinical analysis but that humans nevertheless show a tendency to resist purely actuarial judgement. This resistance extends to the use of algorithmic decision support when compared to human advice (Promberger and Baron 2006; Alvarado-Valencia and Barrero 2014). Evidence from IS research supports these findings: For instance, Lim and O'Connor (1996) demonstrate that people underutilize information from DSS when making decisions. Elkins et al. (2013) find that expert system users feel threatened by system recommendations contradicting their expertise and thus tend to ignore these recommendations. Furthermore, the results by Leyer and Schneider (2019) indicate that managers are less likely to delegate strategic decisions to an AI-based DSS than to another human. While most empirical evidence supports the existence of algorithm aversion, other studies observed an appreciation of algorithmic advice (Dijkstra et al. 1998; Logg et al. 2019) or even an exaggerated reliance on AI-based systems (Dijkstra 1999; Wagner et al. 2018). Similarly, Gunaratne et al. (2018) reveal that humans tend to follow algorithmic financial advice more closely than identical crowdsourced advice. Therefore, our study seeks to contribute toward untangling these contradicting findings. Table A1 in the digital online appendix (available online via http://link.springer.com) provides an overview of recent studies on algorithm aversion.

When comparing studies on algorithm aversion, it is important to note that two differing understandings of the term algorithm aversion exist (Dietvorst et al. 2015; Logg et al. 2019). Dietvorst et al. (2015) coined the term for the choice of inferior human over superior algorithmic judgement. However, their study specifically shows that people

shun algorithmic decision making after having interacted and thus becoming familiar with the particular system. The commonly proposed reason for this behavior is that users devalue algorithmic advice after observing the algorithm to err, which means following the algorithmic advice still holds the risk of making suboptimal decisions (Prahl and Van Swol 2017; Dietvorst et al. 2015; Dzindolet et al. 2002). In contrast, other studies require participants to decide about their reliance on algorithmic advice before becoming familiar with the algorithm's performance (Castelo et al. 2019; Longoni et al. 2019; Logg et al. 2019). These differences result in two varying understandings of what algorithm aversion is: unwillingness to rely on an algorithm that a user has experienced to err versus general resistance to algorithmic judgement. Our study aims at improving our understanding of algorithm aversion by investigating both understandings of this phenomenon in one common setting.

Previous research has suggested manifold predictors of algorithm aversion, such as the perceived subjectivity and uniqueness of tasks (Castelo et al. 2019; Longoni et al. 2019), the decision maker's expertise (Whitecotton 1996) as well as the algorithm's understandability (Yeomans et al. 2019). Burton et al. (2020) assorted possible causes of algorithm aversion into five categories: decision makers' false expectations regarding the algorithms' capabilities and performance, lack of control residing with the decision maker, incentive structures discriminating against the use of algorithmic decision support, incompatibility of intuitive human decision making and algorithmic calculations, and conflicting concepts of rationality between humans and algorithms. This study addresses the first of these categories: It specifically deals with the reasoning that people are less lenient toward algorithms than toward other humans because people expect algorithms to be perfect and do not believe that algorithms can overcome their errors (Dietvorst et al. 2015; Dawes 1979), whereas humans gain experience over time (Highhouse 2008). If this reasoning were true, then people should exhibit lower aversion toward an erring algorithm demonstrating the ability to learn than toward an erring algorithm that does not demonstrate the ability to learn. Existent studies have suggested several measures to enhance the use of DSS: allowing for minor adjustments of the algorithm by the decision maker (Dietvorst et al. 2018); improving the system design (Fildes et al. 2006; Benbasat and Taylor 1978); and training decision makers in the DSS use (Green and Hughes 1986; Mackay and Elam 1992). However, despite the increasing application of machine learning, we do not yet know how decision makers react to advice by AI-based systems that demonstrate the ability to learn and, thus, perceivably improve over time. We address this research gap in this study.

2.2 Ability to Learn

AI researchers employ various approaches to realize computational capabilities (Russell and Norvig 2010). The approach that enabled most of the recent breakthroughs in AI research is machine learning (Jordan and Mitchell 2015). Mitchell (1997, p. 2) defines machine learning as follows: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." Specifically, machine learning allows equipping systems with functionalities via data-based training instead of manual coding. We refer to such systems as AI-based systems because machine learning is part of the AI domain. Owing to algorithmic improvements, the increasing availability of training data, and decreasing costs of computation, machine learning has spurred substantial progress in the realization of several computational capabilities, such as computer vision, speech recognition, natural language processing, and decision support within AI-based systems (Jordan and Mitchell 2015).

When incorporating machine learning in IT systems, we can distinguish between training prior to system deployment (until the system meets specific performance thresholds) and ongoing (i.e., continual) learning after system deployment (Parisi et al. 2019). The latter is necessary if the available data is insufficient to train the system up to a desired level or if the system must be able to adapt to varying environmental conditions or user characteristics. For instance, DSS that depend on their users' personal information suffer from a cold-start problem at the beginning of their use (Liebman et al. 2019). Among continually learning systems, we can further differentiate between those that explicitly involve users in the learning process (i.e., interactive or cooperative learning) and those that implicitly improve over time (Amershi et al. 2014; Saunders et al. 2016). In case of explicit learning, the user is part of the training loop and can exert influence on the process. Examples of explicit learning applications are data labeling (Wang and Hua 2011) and video game design (Seidel et al. 2018). Implicit learning systems improve over time without depending on explicit user feedback by either relying on other data sources or observing users' behavior. Search engines, for instance, optimize the ranking of their search results by drawing upon clickstream data (Joachims and Radlinski 2007).

Whereas previous research has investigated the human role in interactive learning settings (Amershi et al. 2014), little is known about users' reactions toward implicit learning systems. Zhang et al. (2011) show that retailer learning conceptualized as the quality of personalized product recommendations on an e-commerce website reduces customers' product screening and evaluation costs while enhancing decision-making quality. Besides this study, we are not aware of research that has investigated whether humans perceive the ability to learn of AI-based systems and, if so, which consequences these perceptions have. Given the increasing use of machine learning and early calls for research on this matter (Liang 1987), our study seeks to provide first evidence on AI-based system users' perceptions of the ability to learn.

3 Hypotheses Development

The common result of early research on the appreciation of algorithmic judgement is that people generally prefer human judgement despite its oftentimes inferior quality (Dawes et al. 1989). However, recent research findings on the effects of advisor nature (i.e., human or algorithmic) challenge this conclusion: On the one hand, underutilization of algorithmic advice may at least partially reflect the decision maker's overconfidence and egocentric advice discounting (Logg et al. 2019; Soll and Mannes 2011). Therefore, it is important to compare the reliance on algorithmic judgement not against the decision maker's own judgement but against the reliance on another human's judgement. On the other hand, several studies on algorithm aversion have employed tasks, such as recommending jokes (Yeomans et al. 2019) or medical treatment (Promberger and Baron 2006). The quality of decisions in these settings is often subjective or depends on the decision maker's personal characteristics (Longoni et al. 2019; Castelo et al. 2019). However, Castelo et al. (2019) show that algorithm aversion reduces if people perceive tasks to be more objective. Logg et al. (2019) and Gunaratne et al. (2018) even gather evidence for algorithm appreciation in several numeric forecasting tasks. The control group findings by Dietvorst et al. (2015) and the results by Prahl and Van Swol (2017) substantiate the idea of algorithm appreciation unless people have become familiar with an erring algorithm's performance. Following this recent evidence, we suggest that in a decision task with an objectively measurable outcome that is independent of the decision maker's personal characteristics, there is no reason to generally devalue advice from an algorithm the decision maker is not familiar with. Instead, human decision makers may even favor algorithmic advice because an algorithmic advisor's abilities are complementary to their own while those of a human advisor are not (Prahl and Van Swol 2017; Dellermann et al. 2019; Dawes 1979). Accordingly, we hypothesize:

H1 *For an objective and non-personal decision task, human decision makers exhibit algorithm appreciation if they are unfamiliar with the advisor's performance.*

While the literature generally offers mixed results regarding preferences for advisor nature, there is clear evidence of experience with an erring algorithmic advisor having a negative effect on the reliance on this advisor (Dietvorst et al. 2015; Prahl and Van Swol 2017). An important precondition for this effect is that the experience with the algorithmic advisor allows decision makers to determine that the advisor errs. Otherwise, people tend to continually rely on incorrect advice (Dijkstra 1999). A common explanation for this phenomenon is that people expect an algorithm's advice to be perfect (Dzindolet et al. 2002; Highhouse 2008). However, in decisions under uncertainty neither humans nor algorithms can provide perfect advice. A disconfirmation of this expectation then leads to lower reliance on the algorithm compared to a similarly performant or even inferior human (Dietvorst et al. 2015). This reasoning is also in line with IS research on continued system use (Bhattacherjee and Lin 2015). Following the call by Castelo et al. (2019) for more research on how experience with algorithms influences their use, we thus propose:

H2a *For an objective and non-personal decision task, familiarity with an advisor's performance moderates the effect of advisor nature on a human decision maker's reliance on the advice if the advisor errs.*

H2b *For an objective and non-personal decision task, human decision makers rely more on the advice of an unfamiliar algorithm than on the advice of a familiar algorithm if the advisor errs.*

H2c *For an objective and non-personal decision task, human decision makers exhibit algorithm aversion if they are familiar with the advisor's performance and the advisor errs.*

If experiencing an algorithm to err causes a deterioration of reliance on this algorithm's advice owing to unmet performance expectations (Dzindolet et al. 2002; Prahl and Van Swol 2017), a positive experience of an algorithm's performance may conversely encourage a decision maker to rely on the algorithm (Alvarado-Valencia and Barrero 2014). In their study of algorithm aversion, Dietvorst et al. (2015) measured a set of beliefs about differences between human and algorithmic forecasts from the participants' perspective. While the participants thought that algorithms outperformed humans in avoiding obvious mistakes and weighing information consistently, they strongly believed that humans were much better than algorithms at learning from mistakes and improving with practice. However, in

light of the recent technological advances in AI and the increasing use of machine learning (Jordan and Mitchell 2015), these beliefs are not necessarily accurate, especially in the domain of objective and non-personal decision tasks. Likewise, we suggest that an algorithm's ability to learn (i.e., to improve over time) can reduce the detrimental effect that familiarity with an erring algorithm has on the decision maker's reliance on the algorithm's advice. Naturally, this is only possible if users can recognize the algorithm's ability to learn, which means the algorithm must demonstrate this ability. Furthermore, we expect this effect to hold only for algorithmic advisors because human advisors are expected to be able to learn. Therefore, our last hypotheses are:

H3a *For an objective and non-personal decision task, demonstrating an advisor's ability to learn moderates the effect of advisor nature on the reliance on a familiar and erring advisor.*

H3b *For an objective and non-personal decision task, human decision makers rely more on the advice of a familiar and erring algorithm with the ability to learn than on the advice of a familiar and erring algorithm without this ability.*

H3c *For an objective and non-personal decision task, human decision makers do not exhibit algorithm aversion if they are familiar with the advisor's performance and the advisor is erring but has the ability to learn.*

Overall, we suggest that the nature of an advisor in an objective and non-personal decision task has an effect on the decision maker's reliance on the advice in favor of an algorithmic advisor (H1). However, becoming familiar with the advisor's performance before deciding whether to rely on its advice moderates this effect if the advisor errs (H2a). As a result, the reliance on a familiar and erring algorithmic advisor is lower than the reliance on an unfamiliar algorithmic advisor (H2b) and lower than the reliance on a familiar and erring human advisor (H2c). Lastly, an algorithm's ability to learn moderates the effect of advisor nature on the reliance on a familiar advisor (H3a) such that the ability to learn increases the reliance on a familiar and erring algorithmic advisor (H3b) and resolves algorithm aversion (H3c).

4 Method

4.1 Experimental Design and Procedure

To test the hypotheses, we conducted an incentive-compatible online experiment in accordance with most research on algorithm aversion (Burton et al. 2020). An online

experiment fitted the purpose of our study because it allowed us to measure the potential effects precisely and with high internal validity. Our experiment had a between-subject design with manipulations of advisor nature (human vs. algorithmic), familiarity (non-familiar vs. familiar), and ability to learn (non-learning vs. learning). Since the ability to learn can affect decision makers' behavior only if they are familiar with the advisor, we could not employ a traditional full factorial design. Instead, we subdivided the experimental groups becoming familiar with the advisor into those experiencing a non-learning and those experiencing a learning advisor. Table 1 depicts our experimental design.

In our online experiment, we asked the participants to make a forecast within a business setting. The experimental procedure comprised six steps (see Fig. 1): In the first step, we welcomed all the participants and instructed them to answer all questions thoroughly. Furthermore, we informed them about the presence of attention checks, the monetary incentivization, and the payment modalities. The second step encompassed the introduction of our experimental setting. We asked the participants to imagine working as a call center manager and being responsible for the staffing. The call center had just recently acquired a new client. The number of incoming calls for this client's hotline operations now had to be estimated on a regular basis to make appropriate staffing decisions. The participants' task was to estimate the number of incoming calls for a specific day and the accuracy of their estimation partly determined their remuneration for taking part in the study. Accordingly, the participants had an incentive to put their best effort in the estimations. We chose this task because it is a common forecasting problem in business, of objective and non-personal nature, and suitable for machine-learning-based IT support (Ebadi Jalal et al. 2016; Fukunaga et al. 2002). The participants received several aids, which enabled them to make sophisticated estimations. First, we told the participants that the number of calls on an average day would be 5000. Second, the participants received information about six variables influencing the number of calls on a specific day:

- The quarter of the year (ranging from Q1 to Q4);
- The day of the month (ranging from 1 to 31);
- The day of the week (ranging from Monday to Friday);
- The running of a promotion campaign (either yes or no);
- The recent sales (in percent below or above average); and
- The recent website traffic (in percent below or above average).

For all of these variables, we provided the participants with a short explanation about their effects on the number of incoming calls. Third, the participants received an advisor's estimation based on the six variables' specific values on the date for which the participants had to make their forecast. Lastly, we told the participants that they had to make eight training estimations before their final and incentivized estimation.

After receiving this information, the participants had to answer several comprehension questions before proceeding to the third step of the experimental procedure (see Table A7 in the digital appendix). This third step comprised the eight training estimations and was inspired by Dietvorst et al.'s (2015) experimental setting. This training phase was necessary to ensure that the participants could become familiar with the advisor. The participants in the conditions of not becoming familiar with the advisor had a training phase without the advisor to prevent confounds that could potentially distort the results. After completing the training estimations, we once more informed the participants that their ninth and final estimation (i.e., serious phase) would determine the variable share of their payment. The final forecast constituted the fourth step of the experiment and included the measurement of the dependent variable. This step was followed by a post-experimental questionnaire containing control and demographic variables (step 5). In the sixth and last step, we informed the participants about the accuracy of their final estimation and provided them with payment details.

4.2 Experimental Treatments

We administered our experimental treatments in the second and third step of the experimental procedure. In H-U and A-U (i.e., *the unfamiliar advisor conditions*), the participants read in the scenario explanation that they would familiarize themselves with the task during the eight training estimations. However, for the ninth estimation, which determined the participants' bonus payments, they would receive an advice from an advisor. Depending on the experimental condition, we introduced the advisor either as an "Industry Expert" (H-U) with long-standing experience in the field or as a "Prediction Software" (A-U) with a

Table 1 Experimental design

		Human	Algorithmic
Unfamiliar		H-U	A-U
Familiar	Non-learning	H-F-N	A-F-N
	Learning	H-F-L	A-F-L

Fig. 1 Experimental procedure
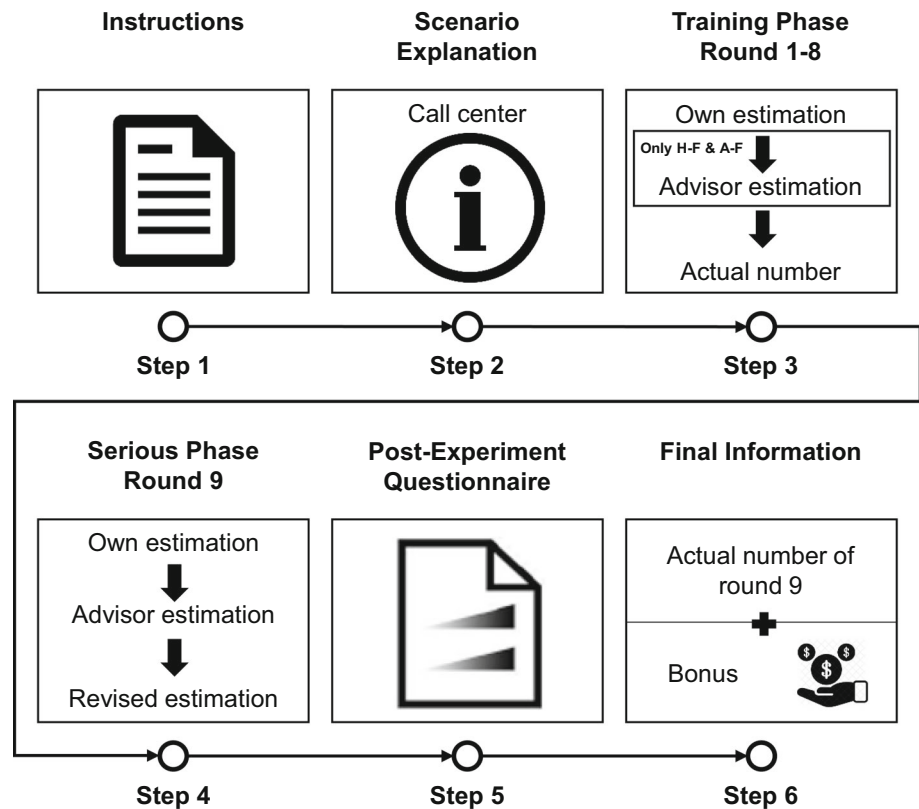
long-standing product history in the field. As such, the participants knew from the beginning that they would encounter an advisor in the last stage of the scenario. In contrast to H-U and A-U, we explained the remaining groups (i.e., *the familiar advisor conditions*) that they would also receive advice throughout the training estimations in the scenario. The advisor introductions for H-F-N and H-F-L were the same as for H-U (i.e., *the human advisor conditions*) and the advisor introductions for A-F-N and A-F-L were the same as for A-U (i.e., *the algorithmic advisor conditions*).

The eight training estimations in the third step of the experiment proceeded as follows: For each round, we showed the participants a specific date along with the levels of the six variables listed in the last section. We selected the eight dates randomly and presented them to the participants in chronological order. While the levels of the first three variables (quarter of the year, day of the month, and day of the week) depended on the chosen date, we generated the levels for the remaining variables (promotion campaign, recent sales, and website traffic) randomly for each of the dates. All the participants saw the same dates and the same variable levels in the same order. Based on the six variable levels, we calculated a true number of calls for each date, which the participants did not know but had to estimate. The digital appendix contains an explanation of the exact calculations (Tables A2–A6). At the end of

each training round, we revealed the true value to the participants such that they could evaluate the accuracy of their estimation. In H-U and A-U, this happened immediately after the participants had submitted their estimations because no advisor was involved in these conditions. In the other groups, the participants received the advisor estimation after submitting their estimation but before receiving the true value. The participants in these conditions could, thus, not only evaluate their own but also the advisor's estimation performance, i.e. becoming familiar with the advisor. The advisor estimations did not differ between human and algorithmic advisors but between non-learning (H-F-N and A-F-N) and learning advisors (H-F-L and A-F-L). While advisors of either nature erred, the learning advisor continually improved in performance, whereas the non-learning advisor did not (see Fig. 2). The average change in prediction errors from round to round (i.e., the error fluctuation) was the same for both types of advisors (12.3%). However, the non-learning advisor had a lower average prediction error (5.5%) than the learning advisor (6.3%). Furthermore, both types of advisors had the same accuracy in the eighth round (4.5%). Accordingly, a favorable perception of the learning advisor is attributable neither to overall performance advantages during the training nor to a lower prediction error in the eighth training round, which might have caused unintended recency effects on the following incentivized estimation.

The convex shape of the learning curve with decreasing performance gains based on additional data corresponds to the learning pattern of machine learning algorithms when applied in new contexts as represented by the new call center client in our scenario (NVIDIA Corporation 2020).

To ascertain that the advisor's estimations were not too far from or too close to the participants' estimations and thus created unintended confound, we conducted a pretest with 248 participants from Amazon Mechanical Turk. The participants had to provide estimations in a scenario similar to our final experiment. The average prediction error of the participants' estimations was 5.5%. Therefore, we designed the advisors' estimations in our actual experiment to have a similar prediction error on average.

Lastly, we conducted a second pretest with 267 participants from Amazon Mechanical Turk experiencing one of the six treatments to examine whether our experimental treatments would work as intended. We used manipulation checks for perceived learning (e.g., "*The Prediction Software gained a good understanding of how to properly estimate the number of calls.*") by Alavi et al. (2002), anthropomorphism (e.g., "*The source of advice is …*" "*machinelike … humanlike*") by Bartneck et al. (2009) and Benlian et al. (2020), and familiarity with the advisor (e.g., "*Overall, I am familiar with the Industry Expert*") by Gefen (2000) and Kim et al. (2009). Table A11 in the digital appendix contains the manipulation checks. The results of the second pretest indicated that all treatments would work as intended: First, H-F-L and A-F-L exhibited a significantly higher level of perceived learning than H-F-N and A-F-N (F = 5.53, $p < 0.05$). Second, H-F-N, H-F-L, A-F-N, and A-F-L exhibited a significantly higher level of familiarity than H-U and A-U (F = 3.00, $p < 0.1$). Lastly, H-U, H-F-N, and H-F-L exhibited a significantly higher level of anthropomorphism than A-U, A-F-N, and A-F-L (F = 4.46, $p < 0.05$).

4.3 Measures

The measurement of our dependent variable was part of the incentive-compatible estimation (step 4) in our experiment. Previous studies on algorithm aversion made use of different instruments to measure a decision maker's reliance on advice. Whereas a number of studies required their participants to fully rely on either the advice or their own judgement (Dietvorst et al. 2015), we chose to use a more fine-grained measure. Specifically, we followed Logg et al. (2019) in employing the judge-advisor paradigm (Sniezek and Buckley 1995) to measure the advisor's influence on the decision maker. In the context of our experiment, this framework requires the decision maker to provide an initial estimation before receiving the advisor's estimation (like during the training phase in step 3) and an adjusted estimation after receiving the advice, which was not the case in the training estimations. The decision maker's initial estimation, adjusted estimation, and the advisor's estimation then allow calculating the weight of advice (WOA):

$$WOA = \frac{adjusted\ estimation - initial\ estimation}{advisor's\ estimation - initial\ estimation}$$

A WOA of 0 means that decision makers do not adjust but remain with their initial estimation and thus ignore the advice. In contrast, a WOA of 1 represents a full adoption of the advisor's estimation. Any values in between reflect the degree to which decision makers take their initial estimation and the advisor's estimation into account for their adjusted estimation. Values below 0 or above 1 may also occur if decision makers believe that the true value lies outside the interval of their initial and the advisor's estimation. Whereas several studies decide to winsorize such values (Logg et al. 2019), we retained these values as they were. Departing from the advisor's estimation (WOA < 0) or overweighting the advisor's estimation (WOA > 1) may be due to the participants' deliberate choices depending on their experience with their own and the advisor's

Fig. 2 Prediction errors of advisors for each training round
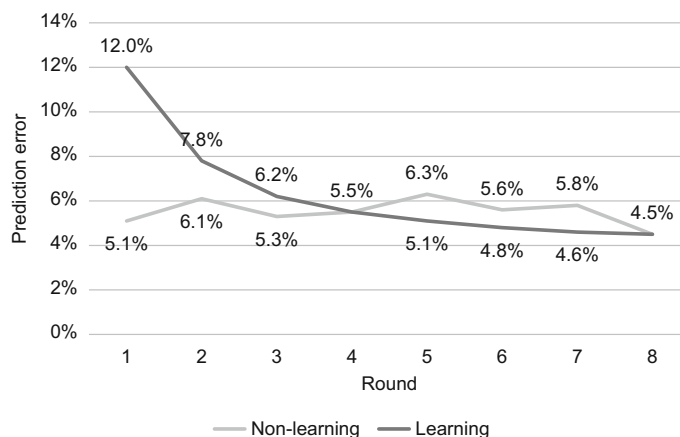
performance in the training phase (Prahl and Van Swol 2017). Based on the WOA values within the different experimental groups, we intended to apply bootstrapped moderation analyses to test H2a as well as H3a and ANOVAs with planned contrasts to test the remaining hypotheses.

Besides our dependent variable, we measured several control and demographic variables in the post-experimental questionnaire (see Tables A8 and A9 in the digital appendix). Among those were the participants' trusting disposition (Gefen and Straub 2004), personal innovativeness (Agarwal and Prasad 1998), experience in working for call centers as well as calling hotlines, and knowledge about call centers (based on Flynn and Goldsmith (1999)). Furthermore, we asked the participants for their age, gender, and education. Between the control and demographic variables, we placed an attention check (see Table A10 in the digital appendix). Lastly, we measured the participant's perceived realism of the scenario.

4.4 Data Collection

To collect sample data, we recruited participants from Amazon Mechanical Turk, a viable and reliable crowd-sourcing platform for behavioral research and experiments (Karahanna et al. 2018; Behrend et al. 2011; Steelman et al. 2014). Using Amazon Mechanical Turk is a suitable sampling strategy for our research, as it enables us to reach users who are internet savvy but not expert forecasters. Since experienced professionals have been shown to rely less on algorithmic advice than lay people, our sample is thus more conservative (Logg et al. 2019). We restricted participation to users who are situated in the U.S. and who exhibited a high approval rating (i.e., at least 95%) to ascertain high data quality (Goodman and Paolacci 2017). Moreover, we incentivized the attentive participation by mentioning that participants could receive up to twice the base payment as a bonus, depending on the accuracy of their final estimation.

From 636 participants completing the questionnaire, we removed those who failed the attention check or inserted values for the incentivized ninth estimation below 100. We further removed participants who exhibited outlier characteristics in the ninth estimation in the form of exceptionally fast (i.e., less than 7 s) or slow (i.e., more than 99 s) estimation times in any of the estimations. The final sample comprised 452 participants. Table 2 provides descriptive information of the analyzed data set.

To confirm the participants' random assignment to the different experimental conditions based on our control and demographic variables, we conducted Fisher's exact tests for the categorical variables and a MANOVA for the metric variables. There are no significant differences in

Table 2 Descriptive sample information

Groups	N	Mean		
		WOA (%)	Age	Gender (male) (%)
H-U	68	58.0	40.6	45.6
A-U	83	74.6	41.5	54.2
H-F-N	72	64.2	39.3	44.4
A-F-N	70	39.9	40.0	54.3
H-F-L	70	54.9	39.1	35.7
A-F-L	89	63.6	38.1	52.8

trusting disposition, personal innovativeness, experience in working for call centers as well as calling hotlines, and knowledge about call centers between the six experimental groups (all $p > 0.1$). We also did not find differences regarding demographics in terms of gender, age, or education (all $p > 0.1$). Lastly, the participants across all groups indicated that they perceived the experiment as realistic (mean = 5.6; std. dev. = 1.1).

5 Results

We tested our hypotheses by conducting a series of analyses in IBM SPSS Statistics 25.

To test H1 – the effect of advisor nature on WOA if the advisors are unfamiliar – we conducted an ANOVA comparing H-U with A-U. The test revealed no significant main effect between the two groups (F = 2.14, $p > 0.1$). As such, *H1 is not supported* in that the participants do not significantly rely more on the unfamiliar algorithmic advisor than on the unfamiliar human advisor.

For H2a, we conducted a bootstrap moderation analysis with 10,000 samples and a 95% confidence interval (CI) with data from H-U, A-U, H-F-N, and A-F-N to test whether familiarity moderates the effect of advisor nature (Hayes 2017, PROCESS model 1). The results of our moderation analysis (see Fig. 3) show that familiarity
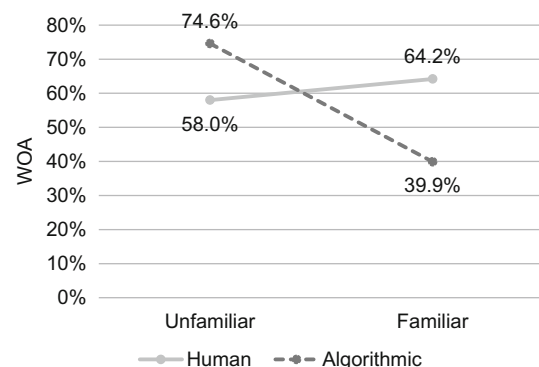

Fig. 3 Interaction plot for H-U, A-U, H-F-N, and A-F-N

moderates the effect of advisor nature on WOA (interaction effect = − 0.41, standard error = 0.18, $p < 0.05$). Specifically, the effect of advisor nature reverses when the advisor is familiar (effect = − 0.24, standard error = 0.13) compared to when the advisor is unfamiliar (effect = 0.17, standard error = 0.13), *supporting H2a*. To test H2b and H2c, we conducted a two-way independent ANOVA with planned contrasts among the same groups. The interaction effect between advisor nature and familiarity is significant ($F = 5.20$, $p < 0.05$), thus confirming the results of our moderation analysis. The pairwise comparison between A-U and A-F-N ($p < 0.01$) is significant and that between H-F-N and A-F-N ($p < 0.1$) is marginally significant. These results provide *support for H2b* and *weak support for H2c*.

For H3a, we conducted a bootstrap moderation analysis with 10,000 samples and a 95% confidence interval with data from groups H-F-N, A-F-N, H-F-L, and A-F-L to test whether demonstrating the ability to learn moderates the effect of advisor nature if the advisor is familiar (Hayes 2017, PROCESS model 1). The results of our moderation analysis (see Fig. 4) show that demonstrating the ability to learn moderates the effect of advisor nature on WOA (interaction effect = 0.33, standard error = 0.16, $p < 0.05$). Specifically, the negative effect of interacting with an algorithmic (vs. human) familiar advisor reverses when the familiar advisor demonstrates the ability to learn (effect = 0.09, standard error = 0.11) compared to when the familiar advisor does not learn (effect = − 0.24, standard error = 0.12), *supporting H3a*. We again conducted a two-way independent ANOVA with planned contrasts among the same groups to test H3b and H3c. The interaction effect between advisor nature and ability to learn is also significant ($F = 4.24$, $p < 0.05$), thus confirming the results of our moderation analysis. Similarly, the pairwise comparison between A-F-N and A-F-L is significant ($p < 0.05$), while the pairwise comparison between H-F-L and A-F-L is not ($p > 0.1$). These results *support both H3b and H3c*.

6 Discussion

Algorithm aversion has spurred controversial discussions in previous research, which resulted in differing understandings of this phenomenon. In this study, we set out to contribute toward clarifying what algorithm aversion is and under which conditions algorithm aversion occurs. Previous studies have produced conflicting findings about whether people are generally averse to algorithmic judgement or avoid algorithms only if they perceive these algorithms to err. Furthermore, we sought to investigate whether demonstrating an algorithm's ability to learn may serve as an effective countermeasure against algorithm
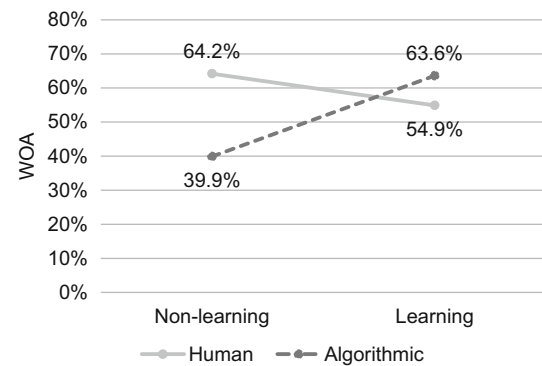

Fig. 4 Interaction plot for H-F-N, A-F-N, H-F-L, and A-F-L

aversion, given that this ability becomes increasingly prevalent in AI-based systems. We studied these questions by simulating a forecasting task within a business setting. The accuracy of both the decision makers' and the simulated advisors' estimations were objectively measurable and did not depend on the decision makers' personal characteristics. These important boundary conditions are true for many business decisions but should be considered when comparing our results with those of earlier studies (Castelo et al. 2019).

According to our results, humans do not generally (i.e., without being familiar with the advisor) prefer human to algorithmic advice. While we hypothesized the opposite (i.e., algorithm appreciation) to be true in the context of our study, our findings do not support this claim. Instead, the participants in our experiment relied to a similar degree on advice coming from an unfamiliar human and an unfamiliar algorithmic advisor. The role of familiarity, however, distinguishes the two understandings of algorithm aversion. Following the reasoning of Dietvorst et al. (2015), people shun algorithmic but not human advice after becoming familiar with the advisor. In other words, familiarity with the advisor interacts with nature of the advisor. The results of our experiment strongly support this claim and, thus, the understanding of algorithm aversion put forward by Dietvorst et al. (2015). Becoming familiar with the advisor reduced the reliance on the algorithmic but not on the human advisor despite their performance (i.e., the accuracy of their estimations in the training period) being identical.

What are possible reasons for this interaction? We adopted a line of argument from prior research, which suggests that erring weighs more severely for algorithmic than for human advisors because humans may overcome their weaknesses while algorithms may not (Highhouse 2008; Dzindolet et al. 2002; Dietvorst et al. 2015). If this were true, demonstrating an algorithm's ability to learn should reduce algorithm aversion. Therefore, we manipulated the advisor's performance during the training

estimations, which allowed the participants to become familiar with the advisor. Demonstrating the ability to learn requires the advisor to improve during the training estimations. This, in turn, means that the learning advisors initially must have a higher prediction error than the non-learning advisors to prevent strong performance differences between them. Accordingly, demonstrating the ability to learn means to elicit even higher algorithm aversion initially and compensating this disadvantage through improvement over time. Indeed, our results show that users honor an algorithm's ability to learn. The participants in our experiment relied more on the learning than on the non-learning algorithmic advisor in their incentivized estimation. Moreover, we could not find differences between reliance on the learning algorithmic advisor and reliance on the learning human advisor. These findings strongly support the idea that demonstrating an algorithm's ability to learn is a promising countermeasure against algorithm aversion. Furthermore, our results indicate that people's beliefs about differences between humans' and algorithms' abilities to learn contribute to algorithm aversion. However, these beliefs are not necessarily accurate, given that machine learning enables IT systems to overcome errors by considering subsequent feedback on their performance. Demonstrating an AI-based system's ability to learn may, thus, update these beliefs and prevent costly behavioral biases in decision making.

7 Implications

Our findings hold important implications for understanding decision makers' reliance on AI-based systems under uncertainty, thereby answering the call for research on individuals' reaction to and collaboration with AI-based systems (Buxmann et al. 2019).

We contribute to previous literature on algorithm aversion by comparing the reliance on an unfamiliar and a familiar algorithmic advisor to the reliance on an unfamiliar and a familiar human advisor of identical performance. Algorithm aversion was evident only if decision makers were familiar with the advisor. Therefore, we recommend using the term algorithm aversion only for the negative effect that familiarization with an algorithmic advisor has on reliance on this advisor, as was initially suggested by Dietvorst et al. (2015). Our results, furthermore, suggest that a general aversion to algorithmic judgement does not exist in objective and non-personal decision contexts. Different findings in early and recent studies on this topic may partly stem from the growing diffusion of algorithms in people's everyday life and a corresponding accustomation to algorithms.

Additionally, we show that the experience during familiarization determines the effect that becoming familiar with an algorithmic advisor has on relying on this advisor. We argue that if the experience does not meet decision makers' expectations of the advisor, the decision makers' reliance decreases. This is a reasonable reaction. However, if decision makers' expectations of an algorithmic advisor are overly high (i.e., decision makers expect an algorithmic advisor to provide perfect advice under uncertainty), this reaction may lead to an irrational discounting of algorithmic advice. Our results indicate this effect by contrasting the familiarization with an erring algorithmic advisor with the familiarization with an identical human advisor. Yet, improving the experience during the familiarization is an effective countermeasure against this effect. We find that demonstrating the ability to learn (i.e., improving over time) is such a countermeasure. Specifically, the continual improvement of the learning algorithmic advisor in our experiment outweighed its initial performance deficits in comparison to the non-learning advisor. To the best of our knowledge, our study is the first to show that users can recognize an algorithm's ability to learn and that this perception has positive effects on the users' behavior toward the algorithm.

Practitioners may also gain useful insights from our study. Companies that provide or seek to employ DSS in contexts under uncertainty should consider possible negative effects of users becoming familiar with IS. To counter these effects, companies employing DSS in such contexts should manage their employees' expectations of what IT systems can and cannot accomplish. Regarding the current debate on the effects of AI-based systems as black boxes (Maedche et al. 2019), our findings suggest that IS developers should invest in demonstrating and communicating the abilities of their IT systems to users. In case of AI-based systems with the ability to learn, this includes transparently demonstrating the system's performance improvements over time. Potential measures to emphasize these improvements are displaying performance comparisons to previous advices and periodical reports on the performance development of IT systems.

8 Limitations and Suggestions for Further Research

Like any research, our study has a few limitations, which also provide leads for further research. First, we designed our study as a simulated online experiment. Even though the experiment was incentive-compatible and of high internal validity, the results do not represent reliance on an algorithmic advisor with a highly consequential decision. Similarly, the participants in our experiment were crowdworkers acquired on Amazon Mechanical Turk. These

participants are likely to be more tech-savvy than the average population and may thus be less likely to exhibit algorithm aversion. As such, it would be interesting to test our hypotheses in a longitudinal field study with a real-world algorithmic advisor to strengthen external validity. Second, our study only constitutes an initial investigation into how demonstrating the ability to learn affects relying on algorithmic advice. Our experimental treatment confronted the participants with an algorithmic advisor exhibiting a stylized learning curve to allow for an unbiased comparison to a non-learning advisor. However, we know little about how bad the initial performance may be without losing decision makers' confidence in the advisor or how decision makers' marginal value of an additional advisor improvement develops over time. Furthermore, we modeled the ability to learn as a continual improvement from estimation to estimation but actual machine learning may also entail temporal performance losses. Therefore, future research should examine the role of boundary conditions in the effect of demonstrating the ability to learn, such as the impact of estimation volatility and mistakes during learning. Moreover, future research can explore possible mediators explaining the effect (e.g., trusting beliefs and expectation-confirmation) of the ability to learn on user behavior. Constructs potentially influencing the aforementioned effect as further moderators include user (e.g., personality types or culture), system (e.g., usefulness or transparency), and task (e.g., complexity or consequentialness) characteristics. Third, our study used a task of objective nature to study the relationship between algorithm aversion and demonstrating an algorithm's ability to learn. Since previous research has shown algorithm aversion to be more severe for tasks of a subjective and personal nature than for tasks of an objective and non-personal nature (Castelo et al. 2019; Longoni et al. 2019), future research may investigate whether demonstrating an algorithm's ability to learn can also alleviate algorithm aversion for subjective and personal tasks. Fourth, our results support the reasoning that people exhibit algorithm aversion for objective and non-personal tasks only if they experience the algorithm to err. However, algorithms, like humans, cannot provide perfect recommendations for decisions under uncertainty. Future research may therefore inspect people's expectations of algorithms and the conditions under which these expectations are disconfirmed.

9 Conclusion

Overall, our study is an initial step toward better understanding how users perceive the abilities of AI-based systems. Specifically, we shed light on how familiarity and demonstrating the ability to learn affect users' reliance on algorithmic decision support. Our findings not only show that familiarity with an erring algorithmic advisor reduces decision makers' reliance on this advisor but also that demonstrating an algorithm's ability to learn over time can offset this effect. We hope that our study provides an impetus for future research on collaboration with and acceptance of AI-based systems as well as actionable recommendations for designing and unblackboxing such systems.

References

Agarwal R, Prasad J (1998) A conceptual and operational definition of personal innovativeness in the domain of information technology. Inf Syst Res 9(2):204–215. https://doi.org/10.1287/isre.9.2.204

Alavi M, Marakas GM, Yoo Y (2002) A comparative study of distributed learning environments on learning outcomes. Inf Syst Res 13(4):404–415. https://doi.org/10.1287/isre.13.4.404.72

Alvarado-Valencia JA, Barrero LH (2014) Reliance, trust and heuristics in judgmental forecasting. Comput Hum Behav 36:102–113. https://doi.org/10.1016/j.chb.2014.03.047

Amershi S, Cakmak M, Knox WB, Kulesza T (2014) Power to the people: the role of humans in interactive machine learning. AI Mag 35(4):105–120. https://doi.org/10.1609/aimag.v35i4.2513

Anthes G (2017) Artificial intelligence poised to ride a new wave. Commun ACM 60(7):19–21. https://doi.org/10.1145/3088342

Bartneck C, Kulić D, Croft E, Zoghbi S (2009) Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int J Soc Robot 1(1):71–81. https://doi.org/10.1007/s12369-008-0001-3

Behrend TS, Sharek DJ, Meade AW, Wiebe EN (2011) The viability of crowdsourcing for survey research. Behav Res Methods 43(3):800–813. https://doi.org/10.3758/s13428-011-0081-0

Benbasat I, Taylor RN (1978) The impact of cognitive styles on information system design. MIS Q 2(2):43–54. https://doi.org/10.2307/248940

Benlian A, Klumpe J, Hinz O (2020) Mitigating the intrusive effects of smart home assistants by using anthropomorphic design features: a multimethod investigation. Inf Syst J 30(6):1010–1042. https://doi.org/10.1111/isj.12243

Bhattacherjee A, Lin C-P (2015) A unified model of IT continuance: three complementary perspectives and crossover effects. Eur J Inf Syst 24(4):364–373. https://doi.org/10.1057/ejis.2013.36

Burton JW, Stein M-K, Jensen TB (2020) A systematic review of algorithm aversion in augmented decision making. J Behav Decis Mak 33(2):220–239. https://doi.org/10.1002/bdm.2155

Buxmann P, Hess T, Thatcher J (2019) Call for papers, issue 1/2021. Bus Inf Syst Eng 61(4):545–547. https://doi.org/10.1007/s12599-019-00606-2

Castelo N, Bos MW, Lehmann DR (2019) Task-dependent algorithm aversion. J Mark Res 56(5):809–825. https://doi.org/10.1177/0022243719851788

Columbus L (2020) Roundup of machine learning forecasts and market estimates, 2020. https://www.forbes.com/sites/louiscolumbus/2020/01/19/roundup-of-machine-learning-forecasts-and-market-estimates-2020/. Accessed 10 Feb 2020

Dawes RM (1979) The robust beauty of improper linear models in decision making. Am Psychol 34(7):571–582. https://doi.org/10.1037/0003-066X.34.7.571

Dawes RM, Faust D, Meehl PE (1989) Clinical versus actuarial judgment. Science 243(4899):1668–1674. https://doi.org/10.1126/science.2648573

Dellermann D, Ebel P, Söllner M, Leimeister JM (2019) Hybrid intelligence. Bus Inf. Syst Eng 61(5):637–643. https://doi.org/10.1007/s12599-019-00595-2

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: people erroneously avoid algorithms after seeing them err. J Exp Psychol Gen 144(1):114–126. https://doi.org/10.1037/xge0000033

Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them. Manag Sci 64(3):1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Dijkstra JJ (1999) User agreement with incorrect expert system advice. Behav Inf Technol 18(6):399–411. https://doi.org/10.1080/014492999118832

Dijkstra JJ, Liebrand WBG, Timminga E (1998) Persuasiveness of expert systems. Behav Inf Technol 17(3):155–163. https://doi.org/10.1080/014492998119526

Dzindolet MT, Pierce LG, Beck HP, Dawe LA (2002) The perceived utility of human and automated aids in a visual detection task. Hum Factors 44(1):79–94. https://doi.org/10.1518/0018720024494856

Ebadi Jalal M, Hosseini M, Karlsson S (2016) Forecasting incoming call volumes in call centers with recurrent neural networks. J Bus Res 69(11):4811–4814. https://doi.org/10.1016/j.jbusres.2016.04.035

Elkins AC, Dunbar NE, Adame B, Nunamaker JF (2013) Are users threatened by credibility assessment systems? J Manag Inf Syst 29(4):249–262. https://doi.org/10.2753/MIS0742-1222290409

Faraj S, Pachidi S, Sayegh K (2018) Working and organizing in the age of the learning algorithm. Inf Organ 28(1):62–70. https://doi.org/10.1016/j.infoandorg.2018.02.005

Fildes R, Goodwin P, Lawrence M (2006) The design features of forecasting support systems and their effectiveness. Decis Support Syst 42(1):351–361. https://doi.org/10.1016/j.dss.2005.01.003

Flynn LR, Goldsmith RE (1999) A short, reliable measure of subjective knowledge. J Bus Res 46(1):57–66. https://doi.org/10.1016/S0148-2963(98)00057-5

Fukunaga A, Hamilton E, Fama J, Andre D, Matan O, Nourbakhsh I (2002) Staff scheduling for inbound call and customer contact centers. AI Mag 23(4):30–40. https://doi.org/10.1609/aimag.v23i4.1667

Gefen D (2000) E-commerce: the role of familiarity and trust. Omega 28(6):725–737. https://doi.org/10.1016/S0305-0483(00)00021-9

Gefen D, Straub DW (2004) Consumer trust in B2C e-Commerce and the importance of social presence: experiments in e-products and e-services. Omega 32(6):407–424. https://doi.org/10.1016/j.omega.2004.01.006

Goodman JK, Paolacci G (2017) Crowdsourcing consumer research. J Consum Res 44(1):196–210. https://doi.org/10.1093/jcr/ucx047

Green GI, Hughes CT (1986) Effects of decision support systems training and cognitive style on decision process attributes. J Manag Inf Syst 3(2):83–93. https://doi.org/10.1080/07421222.1986.11517764

Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C (2000) Clinical versus mechanical prediction: a meta-analysis. Psychol Assess 12(1):19–30. https://doi.org/10.1037/1040-3590.12.1.19

Gunaratne J, Zalmanson L, Nov O (2018) The persuasive power of algorithmic and crowdsourced advice. J Manag Inf Syst 35(4):1092–1120. https://doi.org/10.1080/07421222.2018.1523534

Hayes AF (2017) Introduction to mediation, moderation, and conditional process analysis: a regression-based approach, 2nd edn. Guilford, New York

Highhouse S (2008) Stubborn reliance on intuition and subjectivity in employee selection. Ind Organ Psychol 1(3):333–342. https://doi.org/10.1111/j.1754-9434.2008.00058.x

Joachims T, Radlinski F (2007) Search engines that learn from implicit feedback. Computer 40(8):34–40. https://doi.org/10.1109/MC.2007.289

Jordan MI, Mitchell TM (2015) Machine learning: trends, perspectives, and prospects. Science 349(6245):255–260. https://doi.org/10.1126/science.aaa8415

Karahanna E, Benbasat I, Bapna R, Rai A (2018) Editor's comments: opportunities and challenges for different types of online experiments. MIS Q 42(4):iii-x

Kim DJ, Ferrin DL, Rao HR (2009) Trust and satisfaction, two stepping stones for successful e-commerce relationships: a longitudinal exploration. Inf Syst Res 20(2):237–257. https://doi.org/10.1287/isre.1080.0188

Kuncel NR, Klieger DM, Connelly BS, Ones DS (2013) Mechanical versus clinical data combination in selection and admissions decisions: a meta-analysis. J Appl Psychol 98(6):1060–1072. https://doi.org/10.1037/a0034156

Leyer M, Schneider S (2019) Me, you or AI? How do we feel about delegation. In: 27th European conference on information systems, Stockholm & Uppsala

Liang T-p (1987) User interface design for decision support systems: a self-adaptive approach. Inf Manag 12(4):181–193. https://doi.org/10.1016/0378-7206(87)90041-3

Liebman E, Saar-Tsechansky M, Stone P (2019) The right music at the right time: adaptive personalized playlists based on sequence modeling. MIS Q 43(3):765–786. https://doi.org/10.25300/MISQ/2019/14750

Lim JS, O'Connor M (1996) Judgmental forecasting with interactive forecasting support systems. Decis Support Syst 16(4):339–357. https://doi.org/10.1016/0167-9236(95)00009-7

Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: people prefer algorithmic to human judgment. Organ Behav Hum Decis Process 151:90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Longoni C, Bonezzi A, Morewedge CK (2019) Resistance to medical artificial intelligence. J Consum Res 46(4):629–650. https://doi.org/10.1093/jcr/ucz013

Mackay JM, Elam JJ (1992) A comparative study of how experts and novices use a decision aid to solve problems in complex knowledge domains. Inf Syst Res 3(2):150–172. https://doi.org/10.1287/isre.3.2.150

Maedche A, Legner C, Benlian A, Berger B, Gimpel H, Hess T, Hinz O, Morana S, Söllner M (2019) AI-based digital assistants. Bus Inf Syst Eng 61(4):535–544. https://doi.org/10.1007/s12599-019-00600-8

Meehl PE (1954) Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. University of Minnesota Press, Minneapolis. https://doi.org/10.1037/11281-000

Mitchell TM (1997) Machine learning. McGraw-Hill, New York

NVIDIA Corporation (2020) Deep learning SDK documentation: Training with mixed precision. https://docs.nvidia.com/deeplearning/sdk/mixed-precision-training/index.html. Accessed 7 Feb 2020

Parisi GI, Kemker R, Part JL, Kanan C, Wermter S (2019) Continual lifelong learning with neural networks: a review. Neural Netw 113:54–71. https://doi.org/10.1016/j.neunet.2019.01.012

Prahl A, Van Swol L (2017) Understanding algorithm aversion: when is advice from automation discounted? J Forecast 36(6):691–702. https://doi.org/10.1002/for.2464

Promberger M, Baron J (2006) Do patients trust computers? J Behav Decis Mak 19(5):455–468. https://doi.org/10.1002/bdm.542

Ransbotham S, Kiron D, Gerbert P, Reeves M (2017) Reshaping business with artificial intelligence. MIT Sloan Manag Rev 59(1):1–17

Russell SJ, Norvig P (2010) Artificial intelligence: A modern approach. Pearson, Upper Saddle River

Saunders J, Syrdal DS, Koay KL, Burke N, Dautenhahn K (2016) "Teach me–show me" – end-user personalization of a smart home and companion robot. IEEE Trans Hum Mach Syst 46(1):27–40. https://doi.org/10.1109/THMS.2015.2445105

Seidel S, Berente N, Lindberg A, Lyytinen K, Nickerson JV (2018) Autonomous tools and design: a triple-loop approach to human-machine learning. Commun ACM 62(1):50–57. https://doi.org/10.1145/3210753

Sniezek JA, Buckley T (1995) Cueing and cognitive conflict in judge-advisor decision making. Organ Behav Hum Decis Proc 62(2):159–174. https://doi.org/10.1006/obhd.1995.1040

Soll JB, Mannes AE (2011) Judgmental aggregation strategies depend on whether the self is involved. Int J Forecast 27(1):81–102. https://doi.org/10.1016/j.ijforecast.2010.05.003

Steelman ZR, Hammer BI, Limayem M (2014) Data collection in the digital age: innovative alternatives to student samples. MIS Q 38(2):355–378. https://doi.org/10.25300/MISQ/2014/38.2.02

vom Brocke J, Maaß W, Buxmann P, Maedche A, Leimeister JM, Pecht G (2018) Future work and enterprise systems. Bus Inf Syst Eng 60(4):357–366. https://doi.org/10.1007/s12599-018-0544-2

Wagner AR, Borenstein J, Howard A (2018) Overtrust in the robotic age. Commun ACM 61(9):22–24. https://doi.org/10.1145/3241365

Wang M, Hua X-S (2011) Active learning in multimedia annotation and retrieval: a survey. ACM Trans Intell Syst Technol 2(2):1–21. https://doi.org/10.1145/1899412.1899414

Watson HJ (2017) Preparing for the cognitive generation of decision support. MIS Q Exec 16(3):153–169

Whitecotton SM (1996) The effects of experience and confidence on decision aid reliance: a causal model. Behav Res Acc 8:194–216

Yeomans M, Shah A, Mullainathan S, Kleinberg J (2019) Making sense of recommendations. J Behav Decis Mak 32(4):403–414. https://doi.org/10.1002/bdm.2118

Zhang T, Agarwal R, Lucas HC (2011) The value of IT-enabled retailer learning: personalized product recommendations and customer store loyalty in electronic markets. MIS Q 35(4):859–881. https://doi.org/10.2307/41409964