10-9-2023

The Broken Leg of Algorithm Appreciation: An Experimental Study on the Effect of Unobserved Variables on Advice Utilization

Dirk Leffrang
Paderborn University, Germany, dirk.leffrang@upb.de

The Broken Leg of Algorithm Appreciation: An Experimental Study on the Effect of Unobserved Variables on Advice Utilization

Research Paper

Dirk Leffrang[1]

Paderborn University
dirk.leffrang@uni-paderborn.de

Abstract. Despite the widespread use of machine learning algorithms, their effectiveness is limited by a phenomenon known as algorithm aversion. Recent research concluded that unobserved variables can cause algorithm aversion. However, the impact of an unobserved variable on algorithm aversion remains unclear. Previous studies focused on situations where humans had more variables available than algorithms. We extend this research by conducting an online experiment with 94 participants, systematically varying the number of observable variables to the advisor and the advisor type. Surprisingly, our results did not confirm that an unobserved variable had a negative effect on advice-taking. Instead, we found a positive impact in an algorithm appreciation scenario. This study provides new insights into the paradoxical behavior in which people weigh advice more despite having fewer variables, as they correct for the advisor's errors. Practitioners should consider this behavior when designing algorithms and account for user correction behavior.

Keywords: Algorithm aversion, Data, Decision-making, advice-taking, Human-Computer Interaction

1 Introduction

Algorithms are playing an ever-growing role in a wide range of decision-making processes, from the workplace to our personal lives (Maedche et al. 2019). Algorithm-powered systems offer valuable support thanks to their unparalleled ability to process large amounts of data. Not least because of this property, algorithms have been outperforming human experts in terms of accuracy for decades (see Grove et al. 2000). While algorithms have the potential to offer many benefits, sometimes their use is met with skepticism and distrust compared to human advisors. This phenomenon, known as algorithm aversion, has been studied in various contexts (Jussupow et al. 2020). For example, people prefer human advice to algorithmic advice when predicting the success of students (e.g., Dietvorst et al. 2015). However, inconclusive results exist as the opposite observation was made as well, termed algorithm appreciation (Logg et al. 2019).

One possible explanation for algorithm aversion is that algorithms may not take unique circumstances into account (Longoni et al. 2019). The "broken leg" scenario,

18th International Conference on Wirtschaftsinformatik
September 2023, Paderborn, Germany

proposed by Meehl (1954), illustrates how algorithms can overlook unique characteristics. While an algorithm may be good at predicting the probability of an event, it may fail to account for unobserved variables, such as a broken leg. In contrast, a human expert who is aware of the broken leg can use this information to make a more informed decision.

In summary, algorithms' strengths lie in processing large amounts of data. Weaknesses arise when the algorithm suffers from unobserved variable bias. Based on these two considerations, the broken leg scenario appears to be unfair to the algorithm as the human expert had more observed variables. Therefore, we pose the research question: *What is the impact of an unobserved variable on algorithm appreciation?*

We address our research question by systematically varying (1) human versus algorithm advisors and (2) unobserved versus observed variables. We conducted an online experiment with 94 participants. To measure the weight participants gave to the received advice, we used the Judge-Advisor System (JAS) framework, which we will explain later in the paper (Bonaccio & Dalal 2006, Logg et al. 2019). Surprisingly, our findings contradicted the idea that an unobserved variable would lead to less weight on advice. Participants gave more weight to the advice. Varying advisor types, we found algorithm appreciation. We found no significant interaction between these effects.

This study extends previous research by providing empirical evidence that algorithms lacking an unobserved variable can lead to less advice-taking compared to a human advisor who observes that variable. The findings highlight how people weigh advice more heavily even when they have an unobserved variable. This paradoxical behavior may be due to people correcting for the advisor's errors. To address this paradoxical behavior, practitioners may need to design algorithms that account for their user's correction behavior or provide more transparent explanations.

2 Related Work & Hypothesis Development

2.1 Related Work

The Judge-Advisor System (JAS) is a framework for assessing how much people integrate advice from external sources into their own decision-making. A judge – whether a decision maker in a real-life scenario or a participant in an experiment – faces a prediction task and makes an initial prediction. Afterward, she may consult an advisor, for example, an algorithm or a human expert. The judge forms her final prediction by combining the initial prediction with the advisor's prediction, which is discounted and weighted depending on factors such as the advisor's characteristics (Bonaccio & Dalal 2006).

There is an ongoing debate on whether people generally have an aversion to or appreciation for algorithms. One of the factors related to algorithm aversion is the decision context. Research has observed that algorithm aversion tends to be lower in more objective tasks (Castelo et al. 2019). People view actions that benefit others as more significant for their own personal development than actions that are motivated by financial gain. As a result, they place a greater value on exhibiting empathy and independence when making decisions in situations where others are the primary beneficiaries (Heßler et al. 2022). In our study, we focus on objective tasks to minimize any unintended correlations with the decision maker's personal characteristics.

2.2 Hypothesis Development

Similar to current research on algorithm aversion, research in psychology has examined the effectiveness of human clinical judgment versus algorithmic approaches. In this context, Meehl's "broken leg" scenario is a thought experiment highlighting the limitations of relying solely on algorithmic approaches to make a prediction, such as predicting the probability of a person going to the cinema (Meehl 1954). The human expert had more information about the broken leg than the algorithm in this scenario.

However, algorithmic approaches generally outperformed humans by 10% in terms of accuracy, according to a meta-review of Grove et al. (2000). The only systematic factor that can give humans an advantage over algorithmic approaches is having access to more data, as highlighted by the authors. Building on these insights from psychological research, we wanted to reproduce the anecdotal scenario of Meehl (1954):

Hypothesis 1: In an objective prediction task, people will use the advice of an algorithm less when it suffers from an unobserved variable than the advice of a human expert who observes it.

Meehl (1954) varies two factors simultaneously. Table 1 shows two situations that the scenario does not address. First, it is unclear whether an unobserved variable such as the broken leg would hinder advice-taking from the algorithm as much as it benefited the human expert. Second, it remains unclear whether advice from a human expert would be preferable if she suffered from the same unobserved variable.

Table 1. Potential outcomes in the broken leg scenario of Meehl (1954).

	Unobserved variable	Observed variable
Algorithm	yes	?
Human	?	yes

People exhibited algorithm aversion when the algorithmic advisor was imperfect (Dietvorst et al. 2015). Although subsequent studies failed to establish a general pattern of aversion, Prahl & Van Swol (2017) did find evidence of algorithm aversion after receiving bad advice. However, after non-significant differences directly after bad advice, Leffrang et al. (2023) reported increasing algorithm appreciation over time. In another study, participants only exhibited aversion if they had previously seen the advisor's prediction performance (Berger et al. 2021).

Overall, research found algorithm appreciation in multiple contexts (e.g., Dijkstra et al. 1998, Logg et al. 2019). Human reliance on automation can also lead to a bias known as automation bias (Goddard et al. 2012). In light of these findings, algorithm aversion appears to be more prevalent in temporal or subjective settings. Therefore, we propose:

Hypothesis 2: In an objective prediction task, people will use the advice of an algorithm more than the advice of a human expert.

In a prior study, participants could choose four of five variables as input data for an algorithm. The participants' choice of input variables did not significantly affect

algorithm aversion (Jung & Seiter 2021). However, participants had no choice about the number of observable variables and could not choose for the human condition. Incomplete information can lead to suboptimal decision-making, as demonstrated in Akerlof (1970)'s study on market signaling. In the broken leg scenario, the author considered human experts superior to formal methods when they had more information (Meehl 1954). Given that more variables theoretically enable better predictions, we propose independent of the advisor type:

Hypothesis 3: In an objective prediction task, people will use the advice of an advisor less when the advisor suffers from an unobserved variable than the advice of an advisor who observes it.

Einhorn (1986) discusses the implications of deterministic and probabilistic worldviews for prediction. In a deterministic world, everything is causal and every event can be perfectly predicted. In contrast, a probabilistic world contains random errors that can prevent perfect predictions. Humans have a tendency to reinterpret the past in a way that reflects a deterministic worldview. However, if individuals rely on this reinterpretation to predict the future, they may encounter inconsistencies between their predictions and reality due to the probabilistic nature of events. Models behind algorithms are simplified representations of reality as they are limited by their input data. Instead of assuming perfect predictability, algorithms aim to minimize errors by taking a probabilistic approach, according to the author.

Algorithm aversion arises from people's tendencies to overlook the probabilistic nature of predictions and overestimate the human ability to make deterministic predictions based on intuition and experience (Highhouse 2008). For example, people may resist medical algorithms out of fear that the algorithms might overlook their unique characteristics (Longoni et al. 2019). Human advisors may receive more credibility than algorithms when an unobserved variable is present because the advisors can compensate for the missing information using their intuition and experience. Since its invention, machine learning has built on the idea of enhancing its performance through data and experience in an inductive manner (Mitchell 1997). Empirically, demonstrating the algorithm's ability to improve its prediction performance could mitigate algorithm aversion (Berger et al. 2021). Based on the idea that algorithmic performance improves with more variables, we hypothesize:

Hypothesis 4: The effect of advisor type on advice-taking is greater for an advisor when the advisor suffers from an unobserved variable compared to an advisor who observes it.

3 Research design

To test our hypotheses, we conducted a reward-based online experiment. We preregistered the experiment before we ran it[1]. While incentivization may not always

[1] Note that we reworded the research question and hypotheses to provide a better reading flow, but did not change their meaning `https://aspredicted.org/QZS_8R9`

improve performance, previous studies suggest it can enhance performance in prediction tasks (Camerer & Hogarth 1999). This is why incentivized experiments are frequently utilized for algorithm aversion investigations (see Burton et al. 2020).

Due to uncertainty about the training data for the algorithm, we opted not to use the broken leg scenario. Instead, we chose to use a typical objective task that has been used in previous research, namely the staffing of a call center (e.g., Berger et al. 2021).

3.1 Procedure

Participants accessed the instructions for the experiment through an online survey tool (LimeSurvey). They had to imagine themselves as a manager in a call center, assigned the responsibility of managing staff. The call center had recently gained a new client, and as a result, it was necessary to make regular estimates of the number of incoming calls. This was crucial for the call center to make informed decisions about staffing. Figure 1 shows the produre of the experiment. The task assigned to participants was to predict the number of calls for a specific day by following these steps:

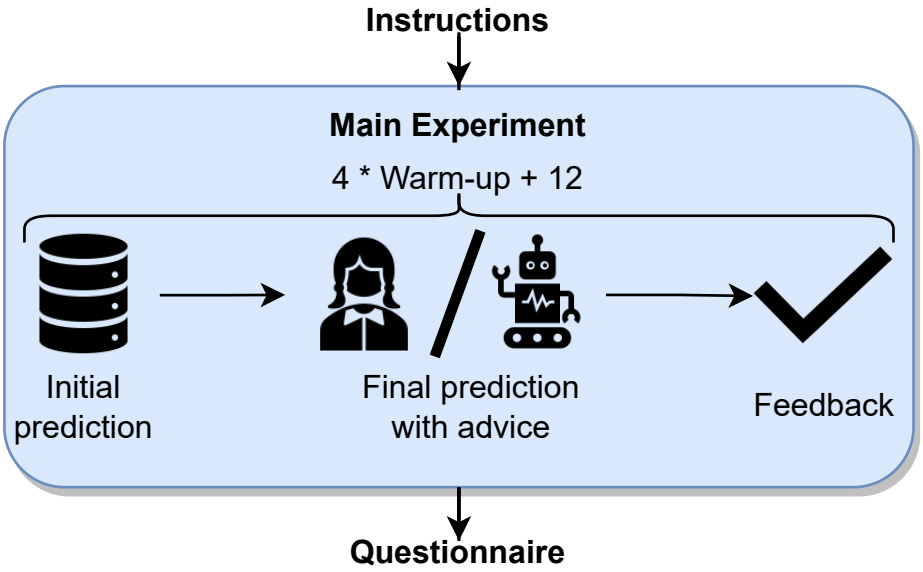


Figure 1. Procedure of the experiment.

1. We presented them with data for a particular day and asked them to make a prediction for that day's incoming calls.
2. After submitting their prediction, we displayed an advisor's estimate for the same day using past data. They then had the possibility to revise their initial forecast.
3. Finally, we informed them of the actual number of phone calls.

As per our instructions, the number of calls on a specific day is primarily influenced by six variables:

– Quarter of the year (Q1 to Q4)
– Day of the month (1 to 31)
– Day of the week (Monday to Friday)
– Running of a promotion campaign (yes or no)
– Recent sales (above or below average)
– Recent website traffic (above or below average)

Berger et al. (2021) described the generation process behind the data in detail. Multiplying the factors for these six variables with a random influence determined the number of incoming calls. As in the experiment of Berger et al. (2021), we did not provide exact effect sizes of the variables. Instead, we showed visualizations displaying the effect sizes in comparison to an average day. For instance, the call center received an average amount of calls when there was no promotion and 20% more calls when there was a promotion. Participants also learned that the average daily call volume is 5000. We excluded the first four tasks as warm-up exercises from the subsequent data analysis.

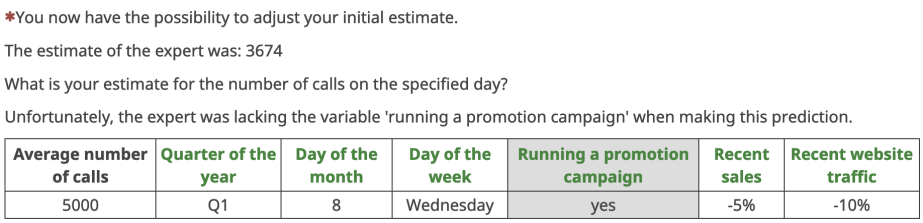
3.2 Conditions & Participants

We employed a 2x2 within-subjects design with four conditions, varying in two factors:

1. Human versus algorithm
2. Unobserved versus observed variable

The only distinction in the first factor was the origin of the advice provided. We informed the treatment group that the advice came from "an algorithm", while the control group received advice from "an expert". However, the actual advice given to both groups was the same.

Importantly, we informed participants in the unobserved variable condition that the advisor did not have the variable "promotion" when making predictions. Figure 2 shows an example of the stimuli in the experiment.


*You now have the possibility to adjust your initial estimate.

The estimate of the expert was: 3674

What is your estimate for the number of calls on the specified day?

Unfortunately, the expert was lacking the variable 'running a promotion campaign' when making this prediction.

Average number of calls	Quarter of the year	Day of the month	Day of the week	Running a promotion campaign	Recent sales	Recent website traffic
5000	Q1	8	Wednesday	yes	-5%	-10%

Figure 2. Screenshot of the human condition with unobserved variable.

The experiment took place between February 2nd and February 13th. We offered an online experiment to third-year bachelor's students in a business informatics course at a medium-sized, IT-focused university in Europe. Participants got a bonus point for their final exam when they passed all attention checks and had a mean absolute percentage error (MAPE) below 0.1. They learned about these restrictions before the experiment and

participation was voluntary. We excluded data from participants who failed to answer all of our questions, failed attention checks, consistently provided the same answers, or completed the tasks too quickly (i.e., less than 90 seconds for all tasks) or too slowly (i.e., more than two standard deviations above the mean participation time). We determined the necessary sample size for detecting a medium effect using G*Power (Faul et al. 2007), and the result was a minimum sample size of approximately 50. The test was a repeated measure, within factors test with Cohen's f = 0.25 (Cohen 1988), indicating a medium effect size (α = 0.1, power: 0.9, 1 group, 2 measurements).

3.3 Model Specification

In our experiment, we utilized the JAS paradigm and measured the Weight of Advice (WOA) as the dependent variable for each subject-task combination. WOA indicates the degree to which participants adopt either human or algorithmic advice (see Bonaccio & Dalal 2006):

$$WOA = \frac{|final\ prediction - initial\ prediction|}{|advisor's\ prediction - initial\ prediction|} \tag{1}$$

A value of zero indicates no adoption of the advice, while a value of one corresponds to full advice adoption. Based on earlier studies on algorithm appreciation (e.g., Logg et al. 2019), we performed Winsorization on any WOA values that were above 1 or below 0.

To analyze our data, we employed a linear mixed-effects model with varying intercepts for condition, participant, and task. Our dependent variable was WOA, and we defined each of the twelve prediction days as tasks. A combination of participant and task defined a task. We randomized the order of conditions for each participant. A dummy variable indicated whether the participant was in the algorithm (Algorithm = 1) or human (Algorithm = 0) condition. Another dummy variable indicated whether the advisor had the promotion as an unobserved variable (Unobserved variable = 1). It resulted in the following generic model specification:

$$Y_i = \beta + \gamma_{algorithm[i]} + \delta_{unobserved\ variable[i]} + \zeta_{participant[i]} + \eta_{task[i]} + \epsilon_i$$

4 Results

4.1 Model-Free Evidence

We obtained predictions from 145 people. After applying the exclusion criteria and dropping the warm-up tasks as mentioned above, 94 participants remained with 1128 predictions. Figure 3 displays the WOA by condition. There are no obvious differences between the algorithmic and human conditions. Surprisingly, the distributions indicate higher levels of WOA for the conditions where there was an unobserved variable. It appears that participants weighted the advice more if the advisor suffered from an unobserved variable.

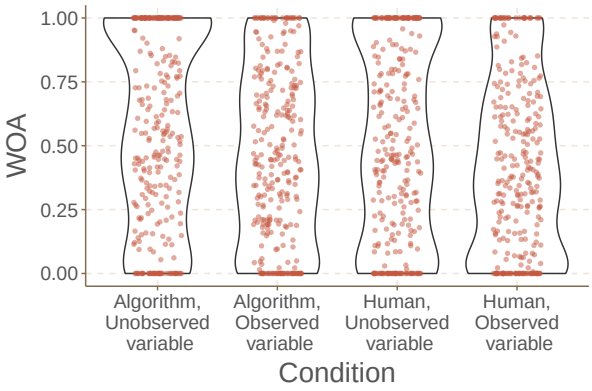


Figure 3. Violin plots for WOA values by condition.

Especially in the unobserved variable conditions, WOA took the value one which indicates potential overshooting due to the Winsorization. Overshooting happens for example when the participant's initial prediction was 80 and the advice was 100, but the final decision is 120. In this case, WOA takes a value greater than 1. However, Winsorization limits WOA to the value of 1. Figure 4 shows the fraction of overshooting by condition and by WOA values being bigger than 1. Notably, when the advice was greater than the pre-advice prediction of the participant in the algorithm, unobserved variable condition, 34.9% of the final predictions overshooted the advice. For the human, unobserved variable condition, 24.6% of the final predictions overshooted the advice. In 67.1% of all predictions, the participants' initial estimate was smaller than the advisor's predictions.

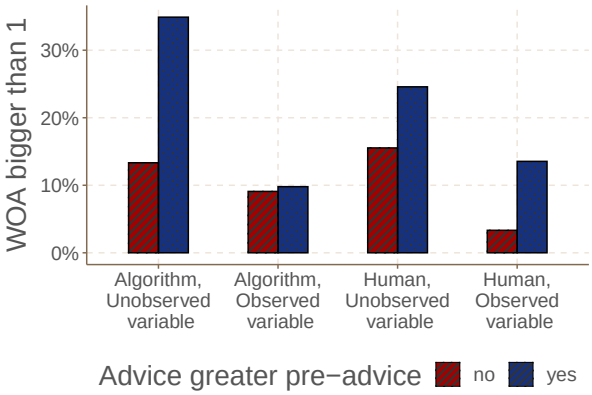


Figure 4. Overshooting plot for WOA values by condition.

4.2 Estimation Results

Our first hypothesis claimed that in an objective prediction task, people will use the advice of an algorithm that lacks an unobserved variable less than the advice of a human expert who observes it. In order to test this hypothesis, we generated a subset for the two conditions presented in Table 1 in the Hypothesis Development (indicated by the label "yes"). We introduced a new variable "Broken leg algorithm", which took the value one if algorithm = 1 and unobserved variable = 1. The variable took the value zero if both were zero.

To test our hypothesis, we regressed WOA on the experimental condition described above, with participant and task differences taken into account as controls. We omitted observations that had undefined WOA differences (due to division by zero). Table 2 model (1) shows the results regarding our first hypothesis.

Table 2. WOA regressed on conditions.

	Dependent variable:		
	WOA		
	(1)	(2)	(3)
Constant	0.422^{***}	0.420^{***}	0.421^{***}
	(0.035)	(0.032)	(0.033)
Broken leg algorithm	0.136^{***}		
	(0.024)		
Algorithm		0.047^{***}	0.046^{*}
		(0.017)	(0.025)
Unobserved variable		0.089^{***}	0.088^{***}
		(0.017)	(0.025)
Algorithm: Unobserved variable			0.002
			(0.036)
Participant?	✓	✓	✓
Task?	✓	✓	✓
Observations	564	1,128	1,128
Log Likelihood	-167.373	-297.528	-299.936
Akaike Inf. Crit.	344.746	607.055	613.872
Bayesian Inf. Crit.	366.422	637.225	649.070
Note:			*p<0.1; **p<0.05; ***p<0.01

Surprisingly, the findings indicate that we cannot support our first hypothesis. On the contrary, the results suggested that for algorithms with an unobserved variable, there was about 13.6%-point more weight of advice compared to human advisors who observe

that variable. This result was statistically significant ($p < 0.01$) and appeared to be economically meaningful with regard to an average WOA of 0.559 across all conditions.

Our second hypothesis stated that people will use the advice of an algorithm more than the advice of a human expert. Just like for the previous hypothesis, we controlled for participant and task differences in model (2). The results implied that algorithms increase the weight of advice by 4.7%-points in comparison to humans ($p < 0.01$). Hence, we find support for our second hypothesis.

Our third hypothesis stated that people will use the advice of an advisor that lacks an unobserved variable less than the advice of an advisor that observes it. In contrast to our hypothesis, the coefficient in model (2) was positive and significant ($p < 0.01$). The results indicated that an unobserved variable increases the weight of advice by 8.9%-points. Therefore, we cannot support our initial hypothesis.

Our fourth hypothesis stated that the effect of advisor type on advice-taking is greater for an advisor who lacks an unobserved variable compared to an advisor who observes it. To test this hypothesis, we created model (3) which also examines the interaction effect of the algorithm and an unobserved variable. Our results did not provide sufficient evidence for an interaction effect, as they did not reveal significant differences regarding the interaction term (Algorithm:Unobserved variable, $p > 0.10$). Thus, we do not find support for our last hypothesis.

The models had an intra-class correlation (ICC) of 0.27, 0.30, and 0.30 for the Participant variable. For the Task variable, the ICC was 0.06, 0.04, and 0.04. The ICC for Task (less than 0.05) suggests that it could be excluded. Doing so only lead to minimal changes to the coefficients of interest, and the p-value for the coefficient of algorithm in model (3) was significant at the 5% level.

5 Discussion

In this paper, we examined the impact of an unobserved variable on algorithm aversion. We conducted an incentivized online experiment in a call center setting to address our research question. Surprisingly, we were unable to generalize the insights of the broken leg thought experiment to empirical advice-taking in a repeated measures design. Our results suggest an underlying effect of unobserved variables leading to more advice-taking.

We can confirm algorithm appreciation for an objective prediction task. Prior research argued that algorithm aversion is more likely for subjective tasks (Castelo et al. 2019). Algorithm appreciation is not necessarily a cognitive bias as algorithms have been outperforming humans for several decades (e.g., Grove et al. 2000). However, we gave imperfect advice and did not observe algorithm aversion as in prior studies (e.g., Dietvorst et al. 2015).

In a similar setting, researchers observed different levels of algorithm aversion and appreciation depending on the familiarity with the algorithm (Berger et al. 2021). In contrast to this study, we measured WOA at different time points. There can be temporal fluctuations in algorithm appreciation depending on the prediction performance (Prahl & Van Swol 2017). Therefore, we controlled for task-level differences between the time

points. In the long run, we can confirm a general algorithm appreciation for objective tasks (Leffrang et al. 2023).

We cannot confirm that people use the advice of an advisor less when the advisor suffered from an unobserved variable. On the contrary, our results imply that people weighted the advice stronger if the advisor suffered from an unobserved variable. We did not observe a significant interaction effect of advisor type and unobserved variables. The ability to choose input variables for an algorithm had no significant impact on advice-taking in prior research (Jung & Seiter 2021). We extend this observation by identifying the number of observable variables as a relevant factor.

One possible explanation for our results could be that people corrected for the mistake the advisor made because of the unobserved variable. As highlighted in Figure 4 in the model-free evidence, there was more overshooting in the unobserved variable condition compared to the observed variable condition when the advice was greater than the pre-advice prediction. Participants knew that promotions increase the number of calls by approximately 20%. This is particularly noteworthy considering that 67.1% of the initial predictions fell below the advisors' predictions. Instead of discounting the quality of the advice, people corrected the advisor's mistake. Paradoxically, correcting for an unobserved variable could lead to a greater weight being assigned to the advice compared to when the variable is observed.

To illustrate the logical nature of this paradoxical behavior, let's consider a simplified example. Imagine a scenario where a judge initially predicts the number of calls to be 80, while an advisor predicts 100. If participants were to assign equal weight to both predictions (see Logg et al. 2019), the final prediction would be 90, resulting in a WOA of 0.5. However, if the judge presumed that the advisor is unaware of a variable that would increase the prediction by 20%, the corrected prediction from the advisor would be 120. In such a scenario, giving equal weight to the judge's initial prediction and the corrected advisor's prediction would result in a final prediction of 100. As the advisor's actual estimate remains at 100, WOA increases from 0.5 to 1. Consequently, knowledge about unobserved variables can amplify the weighting of advice.

In Mehl's thought experiment about estimating the probability of going to a cinema, a judge would have been able to correct the algorithm's error if she had the information that the leg was broken. Since we gave the same numerical advice in our experiment across all four conditions, such a correction was unnecessary.

Observing a previously unobserved variable but leaving the performance unchanged might explain differences to Mehl's thought experiment in the case of a repeated measure design. In Mehl's application case, the algorithm predicted that the person will go to the cinema. However, the human expert predicted that the person will not go to the movies because of the broken leg. If the person goes to the movies despite the broken leg, then it will be rational to trust the algorithm next week, since the additional information from the human expert did not add any value.

We cannot confirm an interaction between the type of advisor and an unobserved variable. For our fourth hypothesis, we hypothesized that observing a previously unobserved variable would lead to a more deterministic worldview (see Einhorn 1986, Highhouse 2008). Future research can investigate to what extent multiple variables lead to such an effect.

This study adds to the current discussion on algorithm aversion versus appreciation by highlighting the importance of data. In light of the ongoing discourse surrounding the impact of AI systems operating as black boxes (Maedche et al. 2019), our research emphasizes data as an essential property of the decision-making process. Input data into algorithms can have an impact on decision-making even though the exact function of the algorithms is not explained (see Wachter et al. 2018). However, more variables do not necessarily lead to more advice being taken. People can correct mistakes they think the advisor is making.

Practitioners need to consider these dynamics when designing effective advice-giving systems. Incorporating human imperfections into recommender systems can increase overall productivity (Sun et al. 2021). An unobserved variable, while keeping the performance constant can increase advice-taking. With this knowledge, practitioners can prevent human behavior from turning a feature into a bug.

6 Limitations & Outlook

It is important to note that our findings were based on a single experimental study. Although our power analysis suggested that our sample size is sufficient, we acknowledge the need for future research. To address this limitation, future research should conduct additional experiments and expand our sample size in order to enhance the robustness and generalizability of our conclusions.

We did not alter performance to vary only one factor between the conditions. In a real-world setting, the performance in the unobserved variable conditions will probably be lower if the advisor does not perform overfitting. Also participants knew the approximate effect size of the unobserved variable, which is not always possible. Future work can vary the performance of the advisors and the effect size as additional degrees of freedom.

This research focused on objective tasks. More subjective contexts may trigger desires in the forms of empathy and autonomy (Heßler et al. 2022). Researchers can investigate whether there is an interaction between unobserved variables and application context.

Within-subjects designs bear the risk of carry-over effects. We chose this design to make it easier for participants to notice the difference between the unobserved and observed variable conditions. Future work can examine the experimental setup in a between-subjects design.

Current research frequently utilizes WOA as the dependent variable. However, potential analyses should take into account the limitations of WOA, such as overshooting, as indicated by Bonaccio & Dalal (2006). We utilized Winsorization to deal with overshooting. Future research can examine other measures of advice-taking.

We plan to investigate other properties of data such as the amount of data points and adding multiple variables to advisors. If we understand algorithms and their interaction with data better, we can also improve human-computer interaction.

References

Akerlof, G. A. (1970), 'The market for "lemons": Quality uncertainty and the market mechanism', *The Quarterly Journal of Economics* **84**(3), 488–500.

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

Bonaccio, S. & Dalal, R. S. (2006), 'Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences', *Organizational Behavior and Human Decision Processes* **101**(2), 127–151.

Burton, J. W., Tina, M.-k. S. & Jensen, B. (2020), 'A systematic review of algorithm aversion in augmented decision making', *Journal of Behavioral Decision Making* **33**(2), 220–239.

Camerer, C. F. & Hogarth, R. M. (1999), 'The effects of financial incentives in experiments: A review and capital-labor-production framework', *Journal of Risk and Uncertainty* **19**(1-3), 7–42.

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

Cohen, J. (1988), *Statistical power analysis for the behavioral sciences*, 2nd ed. edn, Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

Dietvorst, B. J., Simmons, J. P. & Massey, C. (2015), 'Algorithm aversion: People erroneously avoid algorithms after seeing them err', *Journal of Experimental Psychology: General* **144**(1), 114–126.

Dijkstra, J. J., Liebrand, W. B. & Timminga, E. (1998), 'Persuasiveness of expert systems', *Behaviour and Information Technology* **17**(3), 155–163.

Einhorn, H. J. (1986), 'Accepting error to make less error', **50**(3), 387–395.

Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007), 'G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences', *Behavior Research Methods* **39**(2), 175–191.

Goddard, K., Roudsari, A. & Wyatt, J. C. (2012), 'Automation bias: A systematic review of frequency, effect mediators, and mitigators', *Journal of the American Medical Informatics Association* **19**(1), 121–127.

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

Heßler, P. O., Pfeiffer, J. & Hafenbrädl, S. (2022), 'When self-humanization leads to algorithm aversion: What users want from decision support systems on prosocial microlending platforms', *Business and Information Systems Engineering* **64**(3), 275–292.

Highhouse, S. (2008), 'Stubborn reliance on intuition and subjectivity in employee selection', *Industrial and Organizational Psychology* **1**(3), 333–342.

Jung, M. & Seiter, M. (2021), 'Towards a better understanding on mitigating algorithm aversion in forecasting: An experimental study', *Journal of Management Control* **32**(4), 495–516.

Jussupow, E., Benbasat, I. & Heinzl, A. (2020), Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion, *in* 'Proceedings of the 28th European Conference on Information Systems (ECIS)', An Online AIS Conference, pp. 1–16.

Leffrang, D., Bösch, K. & Müller, O. (2023), Do people recover from algorithm aversion? An experimental study of algorithm aversion over time, *in* 'Proceedings of the 56th Hawaii International Conference on System Sciences', Honolulu, USA, pp. 4016–4025.

Logg, J. M., Minson, J. A., Moore, D. A. & States, U. (2019), 'Algorithm appreciation: People prefer algorithmic to human judgment', *Organizational Behavior and Human Decision Processes* **151**(December 2018), 90–103.

Longoni, C., Bonezzi, A. & Morewedge, C. K. (2019), 'Resistance to medical artificial intelligence', *Journal of Consumer Research* **46**(4), 629–650.

Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S. & Söllner, M. (2019), 'AI-based digital assistants: Opportunities, threats, and research perspectives', *Business and Information Systems Engineering* **61**(4), 535–544.

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

Mitchell, T. (1997), *Machine Learning*, McGraw-Hill Science/Engineering/Math, New York.

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

Sun, J., Zhang, D., Hu, H. & Van Mieghem, J. A. (2021), 'Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations', *Management Science* **68**(2), 846–865.

Wachter, S., Mittelstadt, B. & Russell, C. (2018), 'Counterfactual explanations without opening the black box: Automated decisions and the GDPR', *Harvard Journal of Law & Technology* **31**(2), 1–52.