Overcoming Algorithm Aversion with Transparency: Can Transparent Predictions Change User Behavior?

Research Paper

Lasse Bohlen[1], Sven Kruschel[2], Julian Rosenberger[2], Patrick Zschech[1], and Mathias Kraus[2]

[1] Technische Universität Dresden, Dresden, Germany
{lasse.bohlen, patrick.zschech}@tu-dresden.de
[2] Universität Regensburg, Regensburg, Germany
{sven.kruschel, julian.rosenberger, mathias.kraus}@ur.de

Abstract. Previous work has shown that allowing users to adjust a machine learning (ML) model's predictions can reduce aversion to imperfect algorithmic decisions. However, these results were obtained in situations where users had no information about the model's reasoning. Thus, it remains unclear whether interpretable ML models could further reduce algorithm aversion or even render adjustability obsolete. In this paper, we conceptually replicate a well-known study that examines the effect of adjustable predictions on algorithm aversion and extend it by introducing an interpretable ML model that visually reveals its decision logic. Through a pre-registered user study with 280 participants, we investigate how transparency interacts with adjustability in reducing aversion to algorithmic decision-making. Our results replicate the adjustability effect, showing that allowing users to modify algorithmic predictions mitigates aversion. Transparency's impact appears smaller than expected and was not significant for our sample. Furthermore, the effects of transparency and adjustability appear to be more independent than expected.

Keywords: Algorithm Aversion, Adjustability, Transparency, Interpretable Machine Learning, Replication Study

1 Introduction

As machine learning (ML) models are increasingly applied across domains, algorithmic decision-making is becoming more widespread (Janiesch et al. 2021, Berger et al. 2021). Despite those models often demonstrating superior performance compared to human forecasters, individuals frequently exhibit algorithm aversion – a reluctance to rely on algorithmic predictions (Dietvorst et al. 2015, 2018, Jussupow et al. 2024, Mahmud et al. 2023). While algorithm aversion is not strictly tied to performance differences, it is often observed even in cases where algorithms outperform humans. Researchers have identified multiple causes of this aversion (Jussupow et al. 2020), including desire for personal control (Dietvorst et al. 2018), perceived inability to handle unique circumstances (Castelo et al. 2019), and sensitivity to algorithmic errors (Dietvorst et al. 2015, Berger

20th International Conference on Wirtschaftsinformatik
September 2025, Münster, Germany

et al. 2021). A common concern is the black-box nature of many advanced ML models, which generate predictions without revealing their underlying reasoning. According to Burton et al. (2020), achieving coherent decision-making requires aligning the human view with that of the algorithmic model, a process referred to as cognitive compatibility. This alignment is only achievable when the model is transparent enough for its reasoning to be understood. Without cognitive compatibility, algorithmic aids risk clashing with, rather than complementing, human intuition.

This observation has fueled the development of interpretable white-box models, which are fully transparent by design and thus aim to increase user acceptance (Rudin 2019, Kruschel et al. 2025). Interestingly, the empirical evidence on the effect of such models on users remains mixed. Particularly, Poursabzi-Sangdeh et al. (2021) found that study participants were not necessarily more likely to follow the recommendations of a simpler, interpretable model than a complex black-box model. These findings question the practical value of interpretability methods for overcoming algorithm aversion. An alternative approach, suggested by Dietvorst et al. (2018), involves giving users limited control over algorithmic model outputs. In their experiment, participants are allowed to make small adjustments to a model's prediction. This possibility of adjustments significantly increased the willingness to use the algorithmic aid and improved the overall performance of the prediction. However, their study was conducted in a context where participants received no information about the model's decision-making process. This leaves open the question of whether transparency of interpretable models might serve as a substitute for direct control – or whether transparency and adjustability might work synergistically to overcome algorithm aversion.

In our work, we deepen the understanding of the relationship between algorithm aversion, interpretability, and user control. Through a pre-registered user study with 280 participants, we make the following two primary contributions. First, we test the robustness of Dietvorst et al. (2018)'s findings through a conceptual replication of their study using a different prediction task. Second, we extend their paradigm by examining how algorithm transparency through visual explanations of the model's decision logic influences users' willingness to rely on algorithmic predictions, with and without the ability to adjust those predictions.

Our results show that providing participants with the ability to make adjustments to a model's predictions significantly reduces algorithm aversion, closely replicating the findings of Dietvorst et al. (2018). In contrast, transparency alone, implemented in the form of visual explanations of an interpretable model's feature contributions, shows only a modest and statistically insignificant increase in participants' willingness to choose the model. As a result, while participants in our transparency (i.e., *white-box*) condition showed slightly lower error rates, transparency alone did not reliably increase model usage. These findings support the idea that providing users with a tangible sense of control may be more important in overcoming aversion to algorithms than simply revealing the inner workings of the model.

The remainder of this paper is structured as follows. Section 2 provides an overview over related literature on algorithm aversion, interpretable ML, and approaches to overcoming user resistance. Section 3 details our experimental design and analysis. Section 4 presents our findings, and Section 5 discusses their implications for theory and practice.

2 Research Background

2.1 Algorithm Aversion

Algorithm aversion describes people's reluctance to use algorithmic predictions (e.g., Esteva et al. 2017, Brown & Sandholm 2019). Naturally, this phenomenon poses significant challenges in the adoption of algorithmic decision support systems and reasons for this are manifold (Jussupow et al. 2020). In general, people often express less satisfaction with decisions when they learn that they were made by algorithms rather than humans (Longoni et al. 2019). This might be due to users' desire for perfect predictions and their low tolerance for an algorithm's mistakes (Dietvorst et al. 2015, 2018, Wanner et al. 2022). Additionally, subjective or emotion-based tasks contribute to algorithm aversion (Castelo et al. 2019, Germann & Merkle 2023), as people perceive that algorithms fail to consider individual circumstances and characteristics. This aversion is further amplified by ethical concerns, particularly when automated advice influences high-risk decisions (Longoni et al. 2019). Surprisingly, also in cases where algorithmic decisions lead to better outcomes they are often perceived as black-box and thus difficult to understand, ultimately leading to algorithm aversion (Cadario et al. 2021, Filiz et al. 2021).

As a remedy, researchers have developed several strategies to mitigate algorithm aversion. Key approaches include providing choice over the training algorithm (Cheng & Chouldechova 2023), framing algorithms as considering individual characteristics (Longoni et al. 2019) or give users control over algorithmic outputs (Dietvorst et al. 2018). The ability for users to adjust predictions directly addresses the desire for personal control, offering a tangible sense of control and potentially fostering psychological ownership over the final decision (Pierce et al. 2001). Furthermore, this adjustability empowers users with the ability for perceived error correction of the model's suggestions, thereby enhancing their overall sense of control in the decision-making process (Dietvorst et al. 2018). Another promising approach is to increase model transparency. Mahmud et al. (2022) found that the inherent characteristics of prediction models themselves can significantly influence algorithm aversion. Therefore, in this work, we investigate the effect of model transparency on algorithm aversion. Expanding this topic, the following section examines interpretable models as a promising way to empower users to understand an algorithm's decision-making.

2.2 Interpretable Models

In order to obtain transparent ML models, two different approaches are applicable: post-hoc explanation methods or white-box models (Rudin 2019). Post-hoc explanation methods like LIME (Ribeiro et al. 2016) or SHAP (Lundberg & Lee 2017) can explain complex models after they were fit to a specific task. In general, post-hoc explanations provide only approximate insights, adding uncertainty and potentially leaving users' concerns unaddressed (Rudin 2019). In contrast, interpretable models are designed to be transparent in their operation by constraining their complexity. Thus, their decision-making process can be directly examined and understood in detail without additional tools. Linear regression is one of the most basic examples for interpretable models, where each feature's contribution to the prediction is directly quantifiable through its

coefficient. Generalized additive models (GAMs) extend linear models by allowing non-linear relationships between predictors and the response variable, while retaining an additive structure (Kraus et al. 2024). When creating predictions with GAMs the effects of different features remain separate and can be visualized independently, allowing users to understand exactly how each variable influences the prediction (Kruschel et al. 2025). By using an algorithm that can visually communicate its decision logic, we test whether transparency can actually help to reduce algorithm aversion.

2.3 Overcoming Aversion with Interpretable Models

Research suggests that more interpretable models can reduce algorithm aversion because users who understand how predictions are made feel less uncertainty about the algorithm's reasoning, making them more likely to accept and rely on the model (Miller 2019, Binns et al. 2018, Aslan et al. 2024). Unlike human decision-makers, who can explain their reasoning, algorithmic decision aids often provide little to no justification, making users hesitant to rely on them (Kayande et al. 2009). Studies have shown that transparent models can enhance trust and acceptance in algorithmic decision-making (Leichtmann et al. 2023, Wanner et al. 2022). Ideally, transparency not only builds trust but also enables users to more accurately assess when the algorithm is likely to be correct and when its advice should be questioned or modified (Zerilli et al. 2022). Yeomans et al. (2019) demonstrated that algorithm aversion decreases when users can comprehend the model's reasoning. Opening the "black-box" can improve acceptance because transparency helps users understand algorithmic logic (Litterscheidt & Streich 2020, Mahmud et al. 2022), potentially creating cognitive compatibility between human judgment and algorithmic decision logic (Burton et al. 2020). However, transparency has limitations. Work by Poursabzi-Sangdeh et al. (2021) suggests, that interpretability alone does not necessarily mitigate algorithm aversion. If explanations are too complex, they may create only an illusion of understanding rather than genuine insight.

Given the distinct mechanisms through which adjustability enhances perceived control, and transparency aims to improve understanding and trust, their potential interplay in influencing algorithm aversion is critical to understand. For instance, high transparency, by reducing uncertainty and improving understanding, might diminish the perceived need for direct control offered by adjustability (a substitutive effect). Alternatively, enhanced understanding from transparency could empower users to make more effective adjustments, leading to a stronger combined impact (a synergistic effect).

Building on this background, our study proceeds with several key expectations. First, in line with the original work by Dietvorst et al. (2018), we anticipate that providing users the ability to adjust algorithmic predictions reduces algorithm aversion (manifested as increased model usage) and also lead to improved task performance. Secondly, we hypothesize that introducing transparency into the model's decision logic encourages users to rely more on the algorithm, enhancing their task performance.

3 Research Approach

3.1 Prediction Task

We study algorithm aversion in a context that participants can intuitively grasp, selecting the UCI Machine Learning Bike Sharing dataset.[1] This dataset tracks hourly bike rentals based on weather and calendar-related features. In the experiment, participants view these features and predict the number of rentals for a given day and time. From the full dataset, we choose six features that include both continuous (temperature, windspeed, and humidity) as well as categorical (time of day, type of day, and weather situation) variables. We focus on these features because they are easily interpretable and most people intuitively understand how weather and time variables might affect bike rentals.

3.2 Implementation of the Interpretable Model

In Dietvorst et al. (2018), algorithmic predictions came from a simple linear regression model, but the study focused on algorithm aversion under black-box conditions rather than revealing the model's logic to participants. While linear regression provides interpretable coefficients, it may oversimplify complex relationships. Since our analysis requires making the model's decision-making process visually interpretable, we address these limitations by using a GAM (Kruschel et al. 2025). A GAM expresses each feature's contribution with its own potentially non-linear function f_i,

$$\hat{y} = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n). \tag{1}$$

This structure allows more flexible modeling of relationships while maintaining interpretability. Users can examine each f_i independently to see how each feature influences the prediction. We implement our GAM based on IGANN (Kraus et al. 2024). We train IGANN on the bike-sharing dataset, focusing on the six selected features. The final model results in a mean absolute error of around 80, which we communicate to participants so that they are aware of the model's imperfection. Notably, this error is similar in magnitude to the error of the linear model in the original study. The feature plots of the trained model are shown in Figure 1. The structure of the GAM makes it easy for participants to see how each variable contributes to the predicted number of bike rentals, a relationship that is generally intuitive.

3.3 User Study

The goal of our user study is to examine how transparency from interpretable models affects algorithm aversion. Building on Dietvorst et al. (2018), participants complete a prediction task under conditions that systematically vary their control over the model's outputs (i.e., whether they can adjust them) and their insight into the model's logic (i.e., with or without transparent model structure). This setup tests whether the model's transparency amplifies or diminishes the known effect that minimal user control reduces

[1] https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

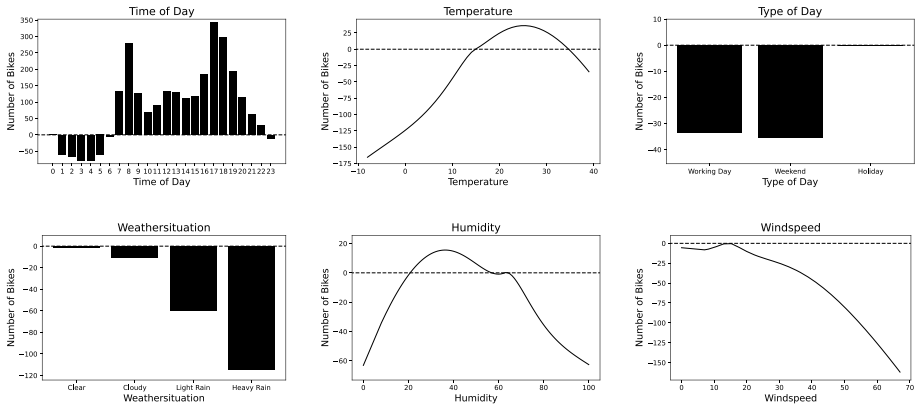


Figure 1. Visualization showing the learned algorithm's decision logic between the individual features and the target (bike demand) of the GAM.

aversion to algorithms. This study was pre-registered[2] and received ethical approval[3] before it was carried out. The remainder of this section covers three key elements. First, the treatment specification explains how participants can adjust the predictions or observe the algorithm's reasoning. Second, the study procedure is outlined and finally, the analysis methods are detailed.

Treatments. We employ a 3 × 2 between-subjects factorial design, through a conceptual replication and extension of the *overcoming algorithm aversion* study of Dietvorst et al. (2018) using the bike rental prediction task. Participants view contextual features and predict hourly rental demands to earn a performance-based bonus. The first treatment factor, adjustability, has three levels. In the *can't-change* condition, participants must accept to rely on the model's prediction or disregard it entirely. In *adjust-by-50*, they can shift the model's prediction by up to 50 bikes or ignore it completely. In *use-freely*, they can change the model's suggestion by any amount. The second treatment factor, transparency, has two levels. In the *white-box* condition, participants see visualizations of the algorithm's decision logic via GAM feature plots. In the *black-box* condition, they get no insight into the model's internal decision logic.

Study Procedure. The study is conducted via an online survey. Participants are introduced to the prediction task, receiving detailed descriptions of the features and learning that demand ranges from 0 to 1,000 bikes, with an average demand of 190 bikes. Next, participants are randomly assigned to one of six conditions in the 3 × 2 experimental design. They are then introduced to the ML model designed to predict bike rental demand. They learn that the model is trained on real data, uses the same features available to them, and has an average error of 80 bikes. Participants in the newly introduced

[2] https://doi.org/10.17605/OSF.IO/RX2TS
[3] https://gfew.de/ethik/8wkDVk4x

white-box condition also receive GAM feature plots visualizing the model's internal structure to help them understand its decision logic (cf. Figure 1). All participants are then informed of the incentive structure: they can earn a bonus of up to £5 based on the accuracy of their predictions, with the bonus decreasing for larger errors. To ensure comprehension, participants are asked to type a sentence summarizing their adjustability condition and the incentive structure. Next, participants in the *can't-change* and *adjust-by-50* conditions must make a one-time binary choice between using the model's predictions or not. They make this choice once and right before they begin making their 20 predictions. Following Dietvorst et al. (2018) procedure, no binary choice is presented in the *use-freely* condition, since participants can fully adjust the model's predictions, making an explicit decision meaningless.

Following this decision, all participants make 20 predictions. In the *use-freely* condition, participants can modify the model's prediction as they wish before submitting their final estimate. In the *adjust-by-50* condition, those who choose to rely on the model can adjust its prediction by up to 50 bikes. In the *can't-change* condition, participants who rely on the model use its prediction directly to calculate their bonus, with no adjustments allowed. In both the *adjust-by-50* and *can't-change* conditions, participants who do not rely on the model make predictions entirely on their own, without seeing the model's output. Figure 2 shows a schematic diagram of the study. The complete online study for all treatments can also be viewed in our online repository.[4]

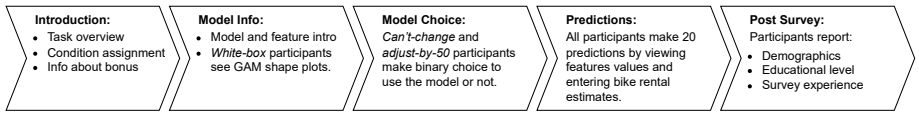


Figure 2. Schematic diagram of the survey procedure

Participants. In total, we recruited 300 participants through Prolific. Key inclusion criteria set on the Prolific platform included an approval rating of 99-100%, English as a first language, and balanced quotas for male and female participants. Participants were recruited from all available countries on the platform. The study lasted an average of 20 minutes and 39 seconds, with participants being paid a base rate and additional bonus payments, resulting in an average of £13 per hour. To ensure data quality, we included two attention checks, resulting in the exclusion of 20 participants. The final sample comprised 280 participants (50.8% female, 47.5% male, one non-binary, and 2 undisclosed). Randomization checks confirmed balanced demographic distribution across treatment groups. Fisher's exact test showed no significant gender differences ($p = 0.498$), and ANOVA revealed no significant age disparities (adjustability: $F(2,\ 280) = 0.721, p = 0.487$; transparency: $F(1,\ 280) = 1.152, p = 0.284$), ensuring comparability across groups. The age and gender distribution for the individual treatment groups is shown in Table 1.

Analysis. We analyze two primary outcomes to examine how transparency and adjustability affect users' behavior in algorithmic predictions: (i) model choice and (ii) task

[4] https://doi.org/10.17605/OSF.IO/RMNFH

Table 1. Descriptive statistics for gender and age across adjustability and transparency treatments

adjustability	transparency	Female	Male	Other	Total	Age (Mean ± SD)
can't-change	*white-box*	27	25	0	52	39.2 ± 12.3
can't-change	*black-box*	19	21	0	40	35.3 ± 11.4
adjust-by-50	*white-box*	22	23	0	45	35.9 ± 10.3
adjust-by-50	*black-box*	18	25	0	43	37.1 ± 13.7
use-freely	*white-box*	24	19	1	44	39.8 ± 14.1
use-freely	*black-box*	27	27	2	56	37.8 ± 13.7
Overall	-	137	140	3	280	37.5 ± 12.6

performance measured as mean absolute error from actual rental demand. We use mean absolute error rather than model deviation because the ML model typically outperforms participants and most deviations increase error rather than improve predictions. However, 68 out of 280 participants managed to outperform the model – remarkably, 67 of them had access to its predictions. This suggests that while the algorithm generally provides strong predictive performance, users can improve upon it under certain conditions. Specifically, access to the model's predictions appears to enable users to identify and correct its shortcomings, leading to better outcomes. In this context, using model deviation as a measure might misinterpret these beneficial adjustments as algorithm aversion, even though they reflected productive engagement with the model. However, additional analyses using the deviation from the model and bonus earned yield similar results and are included in our online appendix.[5]

For the model choice analysis, we used chi-squared tests to assess whether adjustability (*can't-change*, *adjust-by-50*) significantly affected users' decisions to rely on the algorithm. We also analyzed how transparency (*white-box* vs. *black-box*) affected this choice. To analyze the task performance, we performed a two-way ANOVA with transparency and adjustability as factors. This allowed us to identify the potential main effects of each treatment as well as their potential interaction. Before running the ANOVA, we check key assumptions, including normality of residuals and homogeneity of variance. For post-hoc comparisons, we used pairwise t-tests with Bonferroni adjustments to control for multiple comparisons.

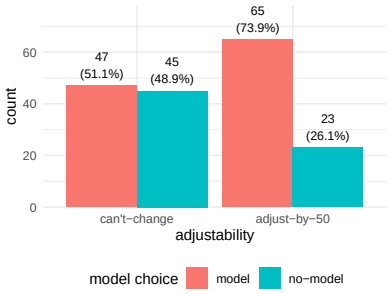
4 Results

4.1 Model Choice

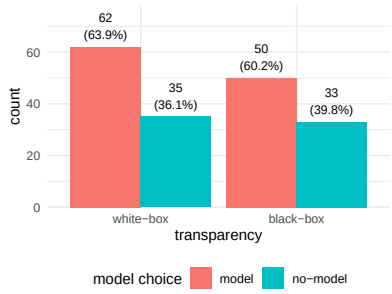
Dietvorst et al. (2018) demonstrate that giving users the option to adjust an algorithm's predictions significantly increases their willingness to rely on that algorithm. Our results replicate this finding clearly. As shown in Figure 3, Participants who had the option to adjust predictions (*adjust-by-50*) chose to rely on the model more frequently (65 participants, 73.9%) compared to those in the *can't-change* condition (47 participants, 51.1%). A chi-square test confirmed this difference was statistically significant ($\chi^2(1,\, N\, =\, 180)\, =\, 8.98, p\, =\, .003$). Thus, our findings support the notion

[5] https://doi.org/10.17605/OSF.IO/RMNFH

that providing users with limited control over algorithmic predictions substantially reduces algorithm aversion. In contrast, transparency alone showed a smaller and non-significant effect on participants' choice. Specifically, in the *white-box* condition, 62 participants (63.9%) chose to rely on the model. In the *black-box* condition, 50 participants (60.2%) chose the model. However, this difference did not reach statistical significance ($\chi^2(1,\ N = 180) = 0.12, p = .724$). Thus, transparency alone was insufficient to meaningfully reduce algorithm aversion without providing users with the ability to adjust predictions. This result differs from our expectation: We assumed that disclosing how the algorithm works would significantly increase adoption, even or especially if users could not adjust its predictions.



(a) Participants' model choice by adjustability. Participants allowed to make adjustments chose the model more (73.9%) than those who couldn't (51.1%), $\chi^2(1, N = 180) = 8.98, p = .003$.



(b) Participants' model choice by transparency. Only minimal difference between *white-box* (63.9%) and *black-box* algorithm recognizable (61.0%), $\chi^2(1, N = 180) = 0.06, p = .724$.

Figure 3. Influence of adjustability (a) and transparency (b) on model choice.

4.2 Task Performance

Table 2. Descriptive statistics for mean task error across adjustability and transparency conditions

		transparency					
		white-box		*black-box*		Overall	
		Mean ± SD	N	Mean ± SD	N	Mean ± SD	N
adjustability	*can't-change*	121.6 ± 51.9	52	137.0 ± 61.9	40	128.3 ± 56.7	92
	adjust-by-50	105.8 ± 42.9	45	113.4 ± 64.9	43	109.5 ± 54.6	88
	use-freely	96.6 ± 20.3	44	102.7 ± 29.3	56	100.0 ± 25.8	100
	Overall	108.8 ± 42.4	141	115.9 ± 53.9	139	112.3 ± 48.2	280

Table 2 shows the mean error for each treatment and as average across all treatment groups. We analyze task performance in terms of mean absolute prediction error, using a two-way ANOVA (cf. Table 3), with adjustability and transparency as independent

factors. Consistent with previous research, the results of the two-way ANOVA indicate a significant main effect from the adjustability treatment ($F(2,\ N = 280) = 9.5$, $p < .001$). However, against our expectations, transparency alone did not significantly affect task performance ($F(1,\ N = 280) = 2.89$, $p = .09$) and the effect size suggests only a small impact in our sample. A possible interaction between both treatment dimensions is not significant and negligible, ($F(2,\ 280) = 0.22$, $p = .807$), indicating the effects of adjustability did not depend on transparency or the other way around. Therefore, we analyze both treatments separately below.

Table 3. Results of a two-way ANOVA reporting the effect of adjustability and transparency on participants' mean task error.

Treatment	df	F	p-value	Effect size η_p^2
adjustability	2	9.500	<0.001	0.065
transparency	1	2.893	0.090	0.010
adjustability × transparency	2	0.255	0.775	0.002

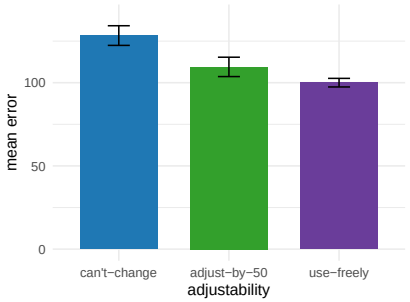
Effect of Adjustability. As the ANOVA shows that there are significant differences between the groups' mean task error in terms of adjustability ($\eta^2 = 0.065$, $p < .001$), we perform a post-hoc test in the form of a paired t-test with Bonferroni correction to further examine this result (cf. Table 4). It reveals that participants in the *can't-change* condition have significantly higher errors (128.3 ± 56.7) than those in the *adjust-by-50* (109.5 ± 54.6, $p = 0.008$) and *use-freely* conditions (100.0 ± 25.8, $p < 0.001$) while the difference between *use-freely* and *adjust-by-50* is not significant ($p = 0.170$) (cf. Figure 4a). These findings align with Dietvorst et al. (2018), as the mere ability to adjust a prediction increases the likelihood of choosing the model, ultimately leading to better performance. However, since participants generally struggle to outperform the predictions from the model, the intensity of possible adjustments has little impact.

Table 4. Results of post-hoc pairwise t-test comparisons for mean task error between different adjustability treatments

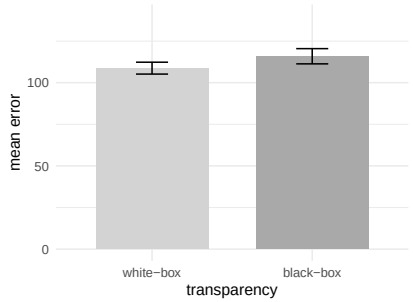
Group 1	Group 2	n_1	n_2	p-value	Adjusted p-value
can't-change	*adjust-by-50*	92	88	0.008	0.024 (*)
can't-change	*use-freely*	92	100	<0.001	<0.001 (***)
adjust-by-50	*use-freely*	88	100	0.170	0.511 (ns)

Effect of Transparency. Although the ANOVA finds no significant differences in mean prediction accuracy between *black-box* and *white-box* participants (cf. Table 3), descriptive analysis reveals a consistent trend: participants in the *white-box* condition exhibit lower average task errors, both overall and within each adjustability subgroup (see Figure 4b and 4c). One possible explanation is that participants who see the model's

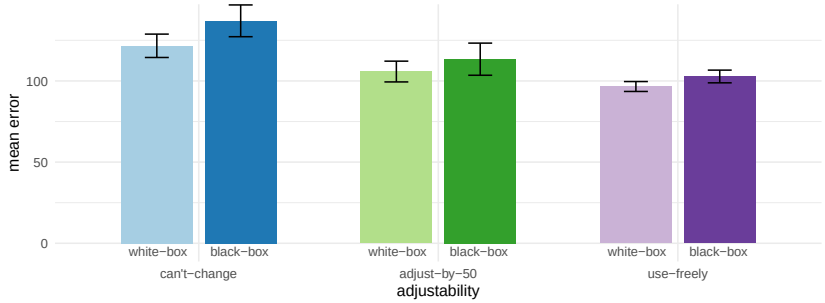
decision logic do not necessarily choose it more often (cf. Section 4.1) but may rely on it more when making adjustments, leading to improved accuracy. Additionally, by interpreting the visual feature plots of the transparent model, they may develop a better understanding of the relationships between predictors and bike demands, enabling them to make more informed predictions even without directly seeing the model's output during the prediction task. To further examine this trend, we conducted a directed post-hoc t-test. The results indicate that mean error in the *white-box* condition ($M = 108.75$) was lower than in the *black-box* condition ($M = 115.90$). However, this difference did not reach statistical significance ($t(261.62) = -1.23, p = 0.110$).



(a) Mean error across adjustability. Errors are significantly higher in the can't-change condition compared to adjust-by-50 ($p = 0.008$) and use-freely ($p < 0.001$).



(b) Mean error across transparency conditions. While white-box participants show average lower errors, the difference is not statistically significant ($p = 0.110$).



(c) Mean error for both transparency and adjustability treatments. The trend suggests slightly lower errors in the white-box condition across all levels of adjustability.

Figure 4. Mean error across experimental conditions. (a) By adjustability treatment, (b) by transparency treatment, and (c) across the combined adjustability and transparency treatments

5 Discussion

In this work, we explore whether revealing a prediction model's decision logic reduces aversion and how transparency and adjustability interact to influence user behavior. Our findings confirm that adjustability significantly reduces algorithm aversion and improves task performance (Dietvorst et al. 2018). In contrast, transparency alone has little impact

on task performance and none on model choice. Although, participants in the *white-box* condition show lower task errors, this trend is not statistically significant. A power analysis showed that to detect such small effects of transparency (partial $\eta^2 \approx 0.01$, Cohen's $f \approx 0.10$), 1,289 participants would be needed for 95% power in the two-way ANOVA. Further, for a significant post hoc t-test, approximately 2,000 participants would be required (Cohen's $d \approx 0.15$, $\alpha = 0.05$, 95% power). A possible explanation for the weak effect of transparency is that simply revealing the model's decision logic once does not guarantee users will engage with or understand it. Moreover, the global nature of the feature plots, while offering general insights into the model's logic, may not have adequately supported task-specific sense-making required by participants when deciding on or adjusting individual predictions. Participants in the *white-box* condition may see the visualizations but not actively process them. Unlike adjustability, which directly alters outcomes, transparency requires cognitive effort, which users may disregard if they see no immediate benefit. Notably, participants view the model's feature plots before knowing they will choose whether to use the algorithm. Even with transparency, people may favor their intuition over algorithmic predictions. Research shows that individuals trust human judgment more, especially in tasks where they feel confident (Yeomans et al. 2019). If participants believe their judgment matches the algorithm's, interpretable visualizations alone may not shift their reluctance toward the algorithm. This suggests that algorithm aversion may be more effectively mitigated by interventions that enhance user adjustability rather than by simply increasing transparency.

Our findings have implications for the design of algorithmic ML-based decision support systems. First, the results indicate that providing users with some degree of control over algorithmic outputs is more effective than transparency alone. Systems designed to assist decision-making should integrate interactive features that allow users to adjust, refine, or personalize algorithmic predictions (Sele & Chugunova 2024). Second, the limited impact of transparency in our study suggests that the way explanations are presented matters. Rather than passively displaying model visualizations, systems may benefit from interactive explanations that encourage users to actively explore how the model makes predictions. Research in interpretable ML increasingly suggests that engagement-driven transparency, where users can manipulate input variables and observe changes in predictions, is more effective than static explanations (Wang et al. 2022).

Our study has limitations that should be acknowledged. First, we do not directly measure participants' engagement with the transparency treatment. While they see visual feature plots of the model's decision logic, we do not assess whether they actively engage with this information. Future research should incorporate measures such as self-reported engagement levels to better measure this engagement. Second, our experiment takes place in a single session. Longitudinal studies could explore whether repeated exposure to interpretable models influences reliance over time. Additionally, our study focuses on a relatively simple prediction task. Future research should therefore examine whether transparency has a stronger effect in complex domains where users inherently struggle to rely on algorithmic reasoning, such as in medical decision-making. Finally, future work could use the Judge-Advisor System paradigm to more precisely quantify how individuals incorporate algorithmic advice by measuring the extent to which participants adjust their initial predictions after viewing the model's output (Bonaccio & Dalal 2006).

References

Aslan, A., Greve, M. & Kolbe, L. (2024), Mitigating discontinuance in medical AI systems: The role of AI explanations, *in* 'Wirtschaftsinformatik 2024 Proceedings'.

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

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J. & Shadbolt, N. (2018), It's reducing a human being to a percentage - Perceptions of justice in algorithmic decisions, *in* 'Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems', pp. 1–14.

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.

Brown, N. & Sandholm, T. (2019), 'Superhuman AI for multiplayer poker', *Science* **365**(6456), 885–890.

Burton, J. W., Stein, M.-K. & Jensen, T. B. (2020), 'A systematic review of algorithm aversion in augmented decision making', *Journal of Behavioral Decision Making* **33**(2), 220–239.

Cadario, R., Longoni, C. & Morewedge, C. K. (2021), 'Understanding, explaining, and utilizing medical artificial intelligence', *Nature Human behaviour* **5**(12), 1636–1642.

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

Cheng, L. & Chouldechova, A. (2023), Overcoming algorithm aversion: A comparison between process and outcome control, *in* 'Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems', pp. 1–27.

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.

Dietvorst, B. J., Simmons, J. P. & Massey, C. (2018), 'Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them', *Management Science* **64**(3), 1155–1170.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M. & Thrun, S. (2017), 'Dermatologist-level classification of skin cancer with deep neural networks', *Nature* **542**(7639), 115–118.

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

Germann, M. & Merkle, C. (2023), 'Algorithm aversion in delegated investing', *Journal of Business Economics* **93**(9), 1691–1727.

Janiesch, C., Zschech, P. & Heinrich, K. (2021), 'Machine learning and deep learning', *Electronic Markets* **31**(3), 685–695.

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)', AIS Virtual Conference.

Jussupow, E., Benbasat, I. & Heinzl, A. (2024), 'An Integrative Perspective on Algorithm Aversion and Appreciation in Decision-Making', *MIS Quarterly* **48**(4), 1575–1590.

Kayande, U., De Bruyn, A., Lilien, G. L., Rangaswamy, A. & Van Bruggen, G. H. (2009), 'How incorporating feedback mechanisms in a DSS affects DSS evaluations', *Information Systems Research* **20**(4), 527–546.

Kraus, M., Tschernutter, D., Weinzierl, S. & Zschech, P. (2024), 'Interpretable generalized additive neural networks', *European Journal of Operational Research* **317**(2), 303–316.

Kruschel, S., Hambauer, N., Weinzierl, S., Zilker, S., Kraus, M. & Zschech, P. (2025), 'Challenging the performance-interpretability trade-off: An evaluation of interpretable machine learning models', *Business & Information Systems Engineering* pp. 1–25.

Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M. & Mara, M. (2023), 'Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task', *Computers in Human Behavior* **139**, 107539.

Litterscheidt, R. & Streich, D. J. (2020), 'Financial education and digital asset management: What's in the black box?', *Journal of Behavioral and Experimental Economics* **87**, 101573.

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

Lundberg, S. M. & Lee, S.-I. (2017), 'A unified approach to interpreting model predictions', *Advances in Neural Information Processing Systems* **30**.

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

Mahmud, H., Islam, A. N. & Mitra, R. K. (2023), 'What drives managers towards algorithm aversion and how to overcome it? Mitigating the impact of innovation resistance through technology readiness', *Technological Forecasting and Social Change* **193**, 122641.

Miller, T. (2019), 'Explanation in artificial intelligence: Insights from the social sciences', *Artificial Intelligence* **267**, 1–38.

Pierce, J. L., Kostova, T. & Dirks, K. T. (2001), 'Toward a Theory of Psychological Ownership in Organizations', *The Academy of Management Review* **26**(2), 298–310.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W. & Wallach, H. (2021), Manipulating and measuring model interpretability, *in* 'Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems', pp. 1–52.

Ribeiro, M. T., Singh, S. & Guestrin, C. (2016), "Why should I trust you?" Explaining the predictions of any classifier, *in* 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', pp. 1135–1144.

Rudin, C. (2019), 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence* **1**(5), 206–215.

Sele, D. & Chugunova, M. (2024), 'Putting a human in the loop: Increasing uptake, but decreasing accuracy of automated decision-making', *PLoS ONE* **19**(2), e0298037.

Wang, Z. J., Kale, A., Nori, H., Stella, P., Nunnally, M. E., Chau, D. H., Vorvoreanu, M., Wortman Vaughan, J. & Caruana, R. (2022), Interpretability, then what? editing machine learning models to reflect human knowledge and values, *in* 'Proceedings

of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining', pp. 4132–4142.

Wanner, J., Herm, L.-V., Heinrich, K. & Janiesch, C. (2022), 'The effect of transparency and trust on intelligent system acceptance: Evidence from a user-based study', *Electronic Markets* **32**(4), 2079–2102.

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

Zerilli, J., Bhatt, U. & Weller, A. (2022), 'How transparency modulates trust in artificial intelligence', *Patterns* **3**(4).