

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Some Averse, Some Not: Individual Differences in Algorithm Aversion among Railway Planners

Permalink

<https://escholarship.org/uc/item/0tz788d7>

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Özüdoğru, Talha

Bachmann, Dominik

Barnhoorn, Quintess Eva Simone

et al.

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Some Averse, Some Not: Individual Differences in Algorithm Aversion among Railway Planners

Talha Özüdoğru^{*1}, Dominik Bachmann^{1,2}, Quintess Eva Simone Barnhoorn³, Silja Renooij⁴,
Christian P. Janssen¹ & Leendert van Maanen¹

¹Experimental Psychology, Utrecht University

²Institute for Logic, Language and Computation, University of Amsterdam

³Dutch Railways (NS)

⁴Information and Computing Sciences, Utrecht University

Abstract

Human–AI collaboration is common, yet AI-assisted decision-making processes among experts in cognitively demanding tasks remain understudied. We investigated how railway planners evaluate advice attributed to either an AI algorithm or a human colleague, and whether witnessing poor advice influences advice acceptance. Seventy-nine railway planners completed an abstract scheduling task. Bayesian hierarchical models showed similar acceptance of AI and human advice, both before and after witnessing poor advice, at the group level. Despite similar average acceptance across groups, models that explicitly capture individual differences indicate heterogeneity only in the AI condition. Some participants discounted the advice after witnessing poor advice, whereas others did not show this discounting. No comparable pattern was evident in the human advice condition. Our results suggest that inconsistent findings in the literature may reflect individual differences in how people respond to algorithmic advice. Future research should therefore focus on identifying and explaining these individual differences.

Keywords: AI-assisted decision-making; algorithm aversion; expert decision-making; AI failure; hierarchical modeling; individual differences

Introduction

Imagine a railway planner tasked with creating a schedule for a fleet of trains while balancing maintenance windows, crew availability, and passenger demand. Years of experience inform her judgments, yet the AI-assisted decision support system now proposes an alternative schedule that optimizes resources in ways the planner might not immediately recognize. Faced with this recommendation, should she rely on the AI's advice, her own expertise, or consult another experienced colleague? And in the event that she previously found the system's advice to be poor, should she be even more cautious about trusting its recommendation now?

The railway planner's dilemma is a special case of a broader problem: people rarely incorporate advice without bias toward the advice source (Boonprakong et al., 2025), even when accepting it could improve their decisions (Bailey et al., 2023). This pattern appears across studies: advice can improve judgment and performance on average, yet people often fail to use it optimally (see Bonaccio & Dalal, 2006; Feng & Gao,

2020; Himmelstein, 2022; Longoni et al., 2019; Soll & Larick, 2009; Soll & Mannes, 2011). Advice taking also shows substantial individual variability, with participants differing in whether they ignore advice, partially adjust their confidence, or change their decision entirely in response to it (Pescetelli et al., 2021).

Beyond human–human advice taking, research has increasingly examined how people respond to algorithmic advice. This work is often framed in terms of generalized tendencies, referred to as algorithm aversion versus algorithm appreciation (Dietvorst et al., 2015; Logg et al., 2019). Here, algorithm aversion refers to resisting algorithmic recommendations in favor of human judgment or being less forgiving of algorithmic mistakes than comparable human mistakes, whereas algorithm appreciation refers to the opposite pattern, placing relatively more weight on algorithmic advice than on human judgment. However, the broader literature is inconsistent, with effects ranging from aversion to appreciation to null results; accordingly, multiple studies find no consistent overall preference for advice from humans versus algorithms (Berger et al., 2021; Daschner & Obermaier, 2022; Prahla & Van Swol, 2017; Schaap et al., 2024). Moreover, this body of work has primarily examined algorithmic advice using lay participants and typically reduces decision-making to binary-choice tasks with an identifiable ground truth (for reviews, see Baines et al., 2024; Burton et al., 2020; Kaufmann et al., 2023; Mahmud et al., 2022).

These experimental design choices limit what can be concluded about advice use in real-world relevant settings such as planning and scheduling, which often involve more than two options and lack a singular, objectively correct answer. In such settings, people typically rely on heuristics and local evaluations rather than exhaustive computation (see e.g. Alvarado-Valencia & Barrero, 2014; Chen et al., 2023; Pinedo, 2016). This stands in contrast to experimental binary choice paradigms, where evaluating all information is feasible and a direct comparison between two alternatives is straightforward.

Problem complexity also introduces variation in decision-making strategies. For instance, in planning, experts may use a combination of multidirectional, opportunistic, and incremental approaches (Hayes-Roth & Hayes-Roth, 1979; Van Wezel & Jorna, 2009). Preferred decision-making strategies likely contribute to individual variation in how advice is utilized. Yet decision support systems are primarily deployed

^{*}Corresponding author. Address: Utrecht University, Heidelberglaan 1, P.O. Box 80140, 3508 TC Utrecht, The Netherlands. Email: m.t.ozudogru@uu.nl

to support expert users, and few studies examine algorithmic advice taking in expert populations or tasks requiring expertise (Aickelin et al., 2022). Consequently, existing research on algorithm aversion and appreciation may fall short in generalizing to decision settings where advice must be integrated with adaptive, strategy-dependent reasoning under uncertainty and competing constraints.

At the same time, experts remain underrepresented in the AI-assisted decision-making literature, with only 8.2% of studies in a recent review including participants with domain expertise (Kaufmann et al., 2023). This gap matters because real-world deployments of decision support systems place algorithmic recommendations in the hands of experts, and the limited evidence from expert settings often points to a reduced willingness to defer to algorithmic decision aids (e.g., AI supported decisions, robo advisors, digital agents, statistical advice) (Elkins et al., 2013; Mahmud et al., 2022; Selten et al., 2023). One possible explanation is that increasing expertise may amplify self-reliance and the tendency to discount external input, placing greater weight on one's own initial judgment (Yaniv, 2004). In line with this, domain expertise is often negatively associated with the use of decision support systems, suggesting lower reliance on algorithmic advice among experts than among less knowledgeable users (Bayer et al., 2022; Burton et al., 2020; Elkins et al., 2013; Logg et al., 2019). Similarly, experts appear more willing to accept AI advice when it confirms their existing judgment but discount or reject it when it is incongruent, even if it is objectively accurate (Selten et al., 2023).

In the present study, we studied advice taking by experts using a bin-packing game, a combinatorial optimization problem that abstractly represents part of railway planners' scheduling task (explained in Materials). Using an experimental design, we compared acceptance of advice related to different advisors (AI algorithm and human colleague). Crucially, participants in the experiment are railway planners, for whom combinatorial optimization problems such as bin-packing are part of their day-to-day work (Van Den Broek et al., 2022). Although direct comparisons are not possible because, to our knowledge, this is the first study in this domain to recruit professionals, the small number of studies on other expert decision makers often reports algorithm aversion (see Bankuoru Egala & Liang, 2024; Kaufmann et al., 2023; Logg et al., 2019). Accordingly, we expect to observe algorithm aversion; namely, we hypothesize that the effect of witnessing poor advice on subsequent advice acceptance will depend on the advisor. Specifically, participants who purportedly received advice from humans should show no change in advice acceptance after witnessing poor advice, whereas participants who purportedly received advice from an AI algorithm should be less willing to accept advice after witnessing poor advice. As alluded to above, both problem complexity and varying expertise are believed to introduce variation in decision-making strategies. Therefore, in addition to testing our hypothesis, we conduct exploratory analyses using individual difference models to test whether

responses to algorithmic advice are broadly similar across participants or vary among individuals.

Methods

Participants

We recruited 102 professional railway planners from two railway companies. After excluding those who failed more than two attention checks, our final sample was 79 participants. We did not collect information about participants' age and gender to maintain full anonymity. However, we observed that the vast majority of participants were male. Participants took part voluntarily and participated as part of their regular work. As an extra incentive, participation qualified them for entry into a raffle to win 20 euros. The study was approved by the Ethics Review Board of the Faculty of Social and Behavioral Sciences, Utrecht University (FETC 24-0001).

Materials

Participants performed an online 2-D bin-packing game. The bin-packing game is a combinatorial optimization problem: players fit items (rectangles of different sizes and colors) from a repository (the leftmost rectangular space in Figure 1a) to one or more bins that have a smaller area than the repository.

The goal of the game is to pack all items into as few bins¹ as possible, ensuring that items do not overlap. We initially developed this task in Özüdoğru et al. (2025), as it is an abstract representation of some of the scheduling tasks our railway planner participants perform as part of their job (e.g., finding empty tracks for trains of different lengths).

Participants were asked to judge the quality of proposed solutions for three different bin-packing problem sets (i.e., three different problem repositories). For all three problem sets, a one-bin solution was technically possible. However, this was never shown to the participants. Instead, all proposed "regular" solutions (see "regular" definition below) used two bins, and the majority of items were always placed in the first bin. We evaluate the quality of solutions by considering the percentage of space occupied in the first bin. We will hereafter refer to this percentage of space as a solution's "efficiency".

A total of 75 solutions were presented: 24 regular solutions (e.g., Figure 1b) and 1 poor solution (e.g., Figure 1c) per problem set. We manually constructed all solutions in Özüdoğru et al. (2025) and ensured that mean efficiency and standard deviation were matched before and after the poor solution.

Design and Procedure

Participants were randomly assigned to one of two groups that differed in who, purportedly, generated the proposed solutions: an AI algorithm or a human colleague (Advisor; between-subjects). Going forward, we refer to these proposed solutions as *advice*. Unknown to the participant, both groups received identical advice, but with different labels ("solution

¹The square with the green outline is bin 1, and the square to its right is bin 2. The repository is visible to the participant only during gameplay; later, only the solution parts (bin 1 and bin 2) are shown.

from your colleague” vs. “solution from an AI algorithm”; see Figure 1b).

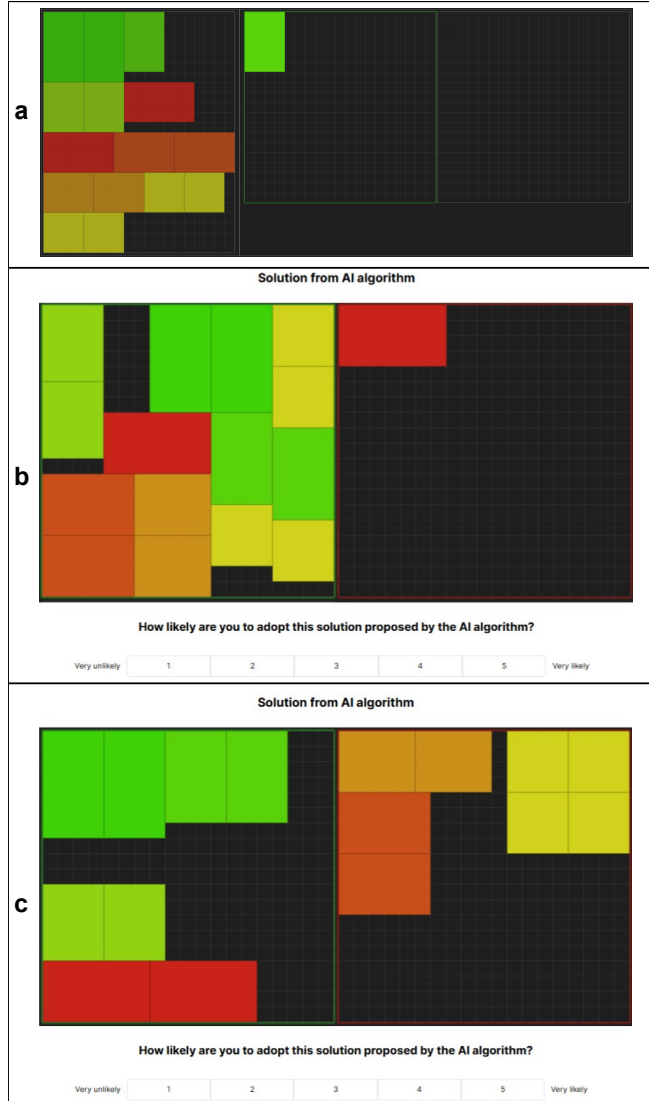


Figure 1: a. Illustration of the bin-packing game; b. an example from experimental trials; c. an example of a poor proposed solution.

After providing informed consent, participants received instructions about the study and were asked to imagine working in a company collaboratively with their assigned (AI or human) advisor. Their role was to deliver the best possible solutions to bin-packing problem sets under time pressure. They could choose to adopt the advisor’s suggestion or reject it in search of a better one, but the latter would cost extra time.

We counterbalanced the starting order of three different bin-packing problem sets using a Latin square design. Before evaluating the advice, participants first attempted to solve a problem themselves within 90 seconds (Figure 1a). They then received and evaluated advice from their assigned advisor across 25 trials per problem set. Each problem set consisted

of 12 regular trials, followed by one trial containing poor advice, and then another 12 regular trials. For the poor advice trials, each item from bin 2 could be moved to bin 1 without mentally rearranging any items in bin 1 (see Figure 1c). For regular trials, on the other hand, integrating items from bin 2 would require mental rearranging (see Figure 1b); better solutions are not obvious. We define witnessing failure as exposure to the poor advice trial; thus, the first 12 regular trials constituted the before witnessing the poor advice block and the last 12 regular trials constituted the after witnessing the poor advice block (i.e., before vs. after witnessing the poor advice). This process was repeated for each problem set. Overall, the experiment followed a 2 (Advisor: Human vs. AI) between-subjects $\times$ 2 (Witnessing Failure: Before vs. After) within-subjects mixed design.

We designed and hosted the experiment on gorilla.sc. The experiment was conducted in Dutch (see <https://osf.io/sfgur> for the original Dutch materials); in this manuscript and all figures, we report English translations. We collected data at the workplace of the railway planners at six different locations throughout the Netherlands, with one session per location. All sessions were conducted in person under the experimenter’s supervision in an office setting. Participants were instructed not to interact with their colleagues; nonetheless, minor protocol deviations (e.g., brief conversations) occurred occasionally.

Measurements

Acceptance of the proposed solution On each trial, we asked participants how likely they were to adopt a proposed solution (e.g., for the AI condition: “How likely are you to adopt this solution proposed by the AI algorithm?”) using a 5-point Likert scale with anchors at “1 – Very Unlikely” and “5 – Very Likely” (see Figure 1b).

Results

Manipulation and Believability Checks

As the present study used the same paradigm as Experiment 1 of Özüdoğru et al. (2025), where 91.7% of participants in the AI advice group believed the advisor label, we did not administer an advisor believability check. The poor advice trial manipulation was effective: mean acceptance across regular trials was $M = 2.81$ ($SD = 1.32$); poor advice trials were rated considerably lower ($M = 1.41$, $SD = 0.82$), with this pattern holding across all three problem sets.

Effects of Advisor Type and Witnessed Failure

To test whether advice acceptance differed by advisor type and by witnessing failure, we analyzed acceptance of advice across trials. We excluded poor-advice trials because they served only to define the before and after blocks, leaving 5,688 observations from 79 participants. We fit a Bayesian multilevel model (brms; Bürkner, 2017) with a Gaussian likelihood, effect-coded predictors (± 0.5), random intercepts for participants

and stimuli, and default priors as specified by the `brms` package. Witnessing failure was coded as trials occurring before versus after the poor-advice trial. Convergence was satisfactory, with $\hat{R} = 1.00$, and effective sample sizes were adequate.

The intercept was estimated at 2.81 on the 5-point scale (95% CI [2.60, 3.02]). Posterior mean estimates for advisor type (−0.01, 95% CI [−0.31, 0.29]), witnessing failure (0.08, 95% CI [−0.20, 0.35]), and their interaction (0.02, 95% CI [−0.08, 0.13]) were all close to zero. The estimates for the main effects and their interaction clustered around zero with wide credible intervals. This suggests that, at the group level, acceptance of advice did not differ by advisor type, nor did witnessing failure interact with advisor type. An equivalent `lmer` model including problem set order confirmed that the Latin square counterbalancing was effective, with order showing no significant effect on acceptance ratings.

Model Comparison

As the inclusion of zero in a 95% CI of the posterior does not offer clear evidence in favor of the absence of an effect, we further evaluated competing models via Bayes factor comparisons. The initial model included two main effects and one interaction; we refer to this model as M1. We subsequently tested simplifications of that model by systematically removing experimental factors (see Table 1).

Table 1: Bayesian Multilevel Model Specifications.

Regression Formula (<code>brms</code> R)
M1: <i>Acceptance</i> ~ <i>Advisor</i> * <i>Failure</i> + <i>RE</i>
M2: <i>Acceptance</i> ~ <i>Advisor</i> + <i>Failure</i> + <i>RE</i>
M3A: <i>Acceptance</i> ~ <i>Failure</i> + <i>RE</i>
M3B: <i>Acceptance</i> ~ <i>Advisor</i> + <i>RE</i>
M4: <i>Acceptance</i> ~ <i>Intercept</i> (1) + <i>RE</i>
Note: RE (Random Effects) = (1 Stimuli) + (1 gr(ID, by = Advisor)), where Stimuli denotes the proposed solution and ID denotes the participant.

Bayes factors quantify the relative evidence for one model compared to another by expressing the ratio of their marginal likelihoods. A Bayes factor close to 1 indicates that the data are about equally likely under both models, whereas values substantially greater or less than 1 indicate stronger evidence for one model over the other.

In this analysis (see Figure 2), Bayes factors favored simpler models that excluded advisor, witnessing failure, and their interaction. More specifically, removing the interaction between advisor and witnessing failure (M2 vs M1) and removing the main effect of advisor (M3A vs M2) provided anecdotal to moderate evidence for the models without these predictors, with the data being 7.05 and 2.57 times more likely under those models, respectively. This suggests that excluding these predictors improves model fit. Evidence favored models excluding advisor and its interaction with witnessing failure. We further compared the two main-effect-only models (M3A: witnessing failure; M3B: advisor) against the intercept-only model (M4). The intercept-only model was favored over both

main-effect models (M4 vs M3A: BF = 5.22; M4 vs M3B: BF = 4.96), with moderate evidence. This suggests that on the group-level, models excluding advisor and witnessing failure were favored, providing evidence against the inclusion of these predictors (see Figure 2).

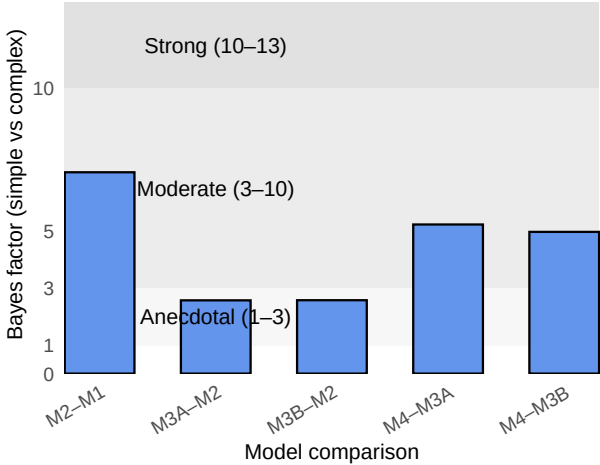


Figure 2: Model Comparison with Bayes Factor.

Modeling Individual Differences

Our initial analysis revealed no group-level differences in advice acceptance between AI and human advisors, nor an effect of witnessing failure. We considered whether this discrepancy might reflect qualitative individual differences that are masked in group-level analyses. To address this, we applied the Bayesian hierarchical model comparison framework developed by Haaf and Rouder (2019) that explicitly tests for the presence of qualitative individual differences.

Model Specification Models were examined separately for the AI advice and human advice groups by modeling within-subject differences in acceptance ratings before versus after the witnessing poor advice trial. For each participant, we defined a latent within subject witnessing failure effect as the difference in expected acceptance before versus after witnessing failure (before minus after). With this coding, positive values indicate aversion (lower acceptance after witnessing failure), whereas values near zero indicate little or no change. Following Haaf and Rouder (2019), we compared five candidate models that reflect different theoretical positions about individual differences in the witnessing failure effect on advice acceptance. The *null model* (M_0) concentrates effects at zero (no systematic change). The *common-effect model* (M_1) assumes a single shared positive effect (aversion) with no between-participant variability. The *positive-effects model* (M_+) allows participants to vary in the magnitude of the witnessing failure effect but constrains all effects to be positive. The *spike-and-slab model* (M_{SS}) captures the “some do and some don’t” account by mixing a spike near zero (little or no change) with a slab of positive effects (aversion) (Haaf & Rouder, 2019).

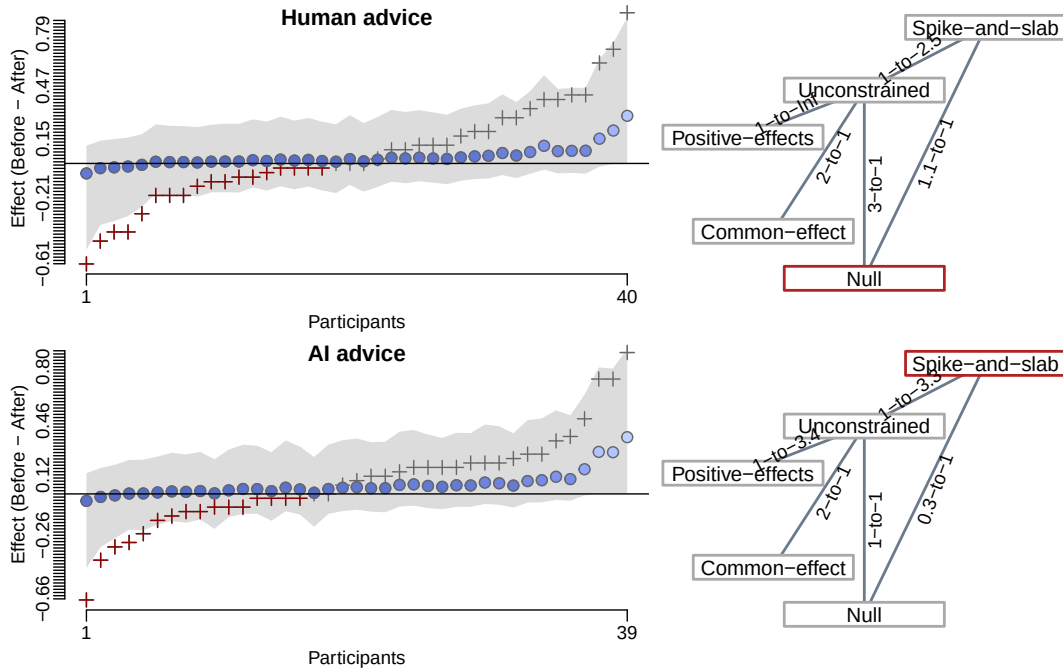


Figure 3: Bayesian hierarchical model estimates and model comparisons for advice acceptance before versus after witnessing poor advice. Left panels show individual observed effects (crosses) and posterior estimates of individual effects (circles; lighter color indicates higher "slab probability") with 95% CI (gray regions). Right panels display Bayes factors comparing five candidate models. Numbers indicate evidence ratios; red borders indicate the best-fitting model.

Finally, the *unconstrained model* (M_u) allows witnessing failure effects to vary continuously without directional constraints, so participants may show aversion, little change, or the opposite pattern (appreciation).

The null, common-effect, positive-effects, and unconstrained models were estimated using MCMC with 10,000 iterations, discarding the first 5,000 as warm-up. The spike-and-slab model required more extensive sampling and was estimated using 100,000 iterations, discarding the first 10,000 as warm-up. Bayes factors were computed using the encompassing approach for the spike-and-slab and positivity-constrained models (posterior-to-prior mass ratio in the constrained region), with remaining model comparisons derived by transitivity (Haaf & Rouder, 2019).

Model Comparisons for Individual Differences Model comparisons indicated that in the Human advice condition, there was moderate evidence for the null model (BF = 3.0 over the unconstrained model), suggesting no systematic change in advice acceptance after the poor advice trial. In contrast, in the AI advice group, model comparisons provided moderate evidence (see Kinney et al., 2021) for the spike-and-slab model (BF = 3.3 over the unconstrained model), which suggests that some participants showed no systematic change in advice acceptance after the poor advice trial (spike) while others showed a positive effect (slab), representing algorithm aversion (see Figure 3).

Discussion

The present study examined whether witnessing poor advice affects advice acceptance, and whether this differs when the advisor is an AI system versus a human colleague, using an abstract scheduling task relevant to the day-to-day work of railway planners. Building on the initial analyses, we modeled individual differences to test whether the impact of witnessing poor advice varies qualitatively across participants within each advisor condition. In the sections that follow, we first contextualize the absence of an advisor effect within the mixed findings of the algorithmic advice taking literature. We then discuss the implications of individual differences in advice taking following witnessing failure, offer a strategy-based explanation, and summarize key methodological, practical, and interpretive limitations.

Our initial analyses showed that, at the group level, advice acceptance was similar across advisors, and witnessing failure did not differently influence acceptance depending on advisor type. In the broader literature spanning novice and expert samples and a wide range of tasks, evidence is mixed, with reports of algorithm aversion (Alvarado-Valencia & Barrero, 2014; Dietvorst et al., 2015; Hertz & Wiese, 2019; Longoni et al., 2019; Yeomans et al., 2019), algorithm appreciation (Araujo et al., 2020; Chacon et al., 2022; Logg et al., 2019; Mesbah et al., 2021), and no reliable differences in acceptance across advisor types (Berger et al., 2021; Daschner & Obermaier, 2022; Prahll & Van Swol, 2017; Schaap et al.,

2024). Such mixed evidence is broadly consistent with existing heterogeneity. However, relative to our hypothesis and to prior studies examining algorithmic advice taking among expert decision makers, our results diverge from the expectation of group-level differences by advisor type. Specifically, we expected witnessing poor advice to reduce acceptance of AI advice more than human advice, which was not observed at the group level.

To examine whether absence of a group-level advisor effect concealed meaningful heterogeneity, we next examined individual differences in how witnessing failure affects advice acceptance from AI versus human advisors, using Bayesian models that explicitly test for qualitative differences between individuals. These analyses revealed a qualitative difference in individual-level heterogeneity across advisor conditions. In the human advice group, the best-fitting model provided moderate evidence for a null effect, suggesting that responses were largely unchanged after witnessing failure. In contrast, in the AI advice group the best-fitting model showed moderate evidence favoring a spike-and-slab model that allowed for qualitative individual differences: some participants exhibited reduced acceptance after witnessing failure, whereas others showed little to no change. This pattern suggests that no clear group-level difference for AI versus human advice may have arisen because participants are a mixture of those who show aversion after failure and those who show little to no change. When the subgroup showing aversion is relatively small or the magnitude of this effect is modest, the resulting average effect may be difficult to detect with group-level analyses.

One plausible explanation for why individual differences emerged primarily in responses to AI advice is that the task permits more than one reasonable way to evaluate solutions, and AI advice may be evaluated more strongly through the lens of these pre-existing strategies. Prior work suggests that, due to the complexity of planning tasks (Hayes-Roth & Hayes-Roth, 1979; Van Wezel & Jorna, 2009), planners use different heuristics, which could be one source of individual variation in how advice is utilized. In scheduling problems, solutions are rarely uniquely correct; instead, they are often assessed relative to constraints, operational preferences, and trade-offs. Under such conditions, the same advice may appear more or less compelling, depending on the evaluator's strategy, and witnessing failure could selectively reduce acceptance among participants who already treat the AI advisor as less reliable or less aligned with their initial strategy. This interpretation is also consistent with the observed variation in witnessing-failure responses within the AI condition.

Strategy-dependent evaluation of advice helps explain why findings in the algorithmic advice acceptance literature differ across studies. If samples differ in the proportion of individuals who penalize algorithm mistakes heavily, group-level comparisons will vary accordingly. Moreover, even when tasks provide objective accuracy feedback, changes in advice acceptance after observing advisor performance may be more uniform when there is less room for strategy-dependent eval-

uation. Differences in task structure and sample composition may therefore shift findings between apparent algorithm aversion, appreciation, and indifferent acceptance across advisors even when underlying individual-level processes vary.

Methodologically, our findings suggest that group-level statistical models can miss theoretically important heterogeneity in advice acceptance and that hierarchical models that allow for the detection of qualitative individual differences can provide a more nuanced account. Practically, designers and organizations implementing algorithmic decision aids for expert users may benefit from (1) anticipating variability in how algorithmic recommendations are evaluated and updated after failures, (2) testing whether certain evaluation strategies are less susceptible to algorithm aversion and gently guiding users toward them where appropriate, and (3) providing interface designs for presenting and interacting with algorithmic advice that support different evaluation preferences rather than relying on one-size-fits-all solutions.

Several limitations constrain the interpretation of our results. Firstly, our study focuses on railway planners and a specific scheduling task, so generalization to other expert domains and task types remains uncertain. Secondly, although the task was designed to reflect key aspects of railway planning, the experimental setting cannot fully reproduce operational constraints (for example, real-time pressure, accountability), reflecting a common trade-off between ecological validity and experimental control. Thirdly, individual differences modeling was conducted separately within the AI and human advice conditions. Yet, an ideal implementation of this approach would use a fully within-subject design to estimate and compare heterogeneity across advisor types within the same participants. Fourthly, our proposed explanation that planners may rely on different heuristics or decision-making strategies is plausible given the task structure, but could not be validated in our present study; future work should measure or manipulate strategy use directly to evaluate this account. Lastly, sample size was constrained by the practical difficulty of recruiting experts, limiting statistical precision; consequently, evidence regarding advisor-related effects was anecdotal to moderate at both the group and individual levels.

In conclusion, railway planners accepted AI and human advice at comparable rates, and we found no clear evidence that witnessing failure differentially affected advice acceptance across advisor conditions at the group level. However, only in the AI advice condition did witnessing poor advice lead to heterogeneous changes in advice acceptance across individuals. Taken together, these results suggest that expert responses to algorithmic advice are better characterized as conditional and participant-specific rather than as a uniform tendency toward algorithm aversion, underscoring the importance of modeling individual differences in AI-supported decision-making contexts.

Acknowledgments

This research was performed within the ICAI ROBUST Rail-Lab, a collaboration between Utrecht University, TU Delft, Dutch Railways (NS), and ProRail, co-funded by the Dutch Research Council (NWO) and the Dutch Ministry of Economic Affairs (EZ).

We thank Joris Snijders and Micha van Os for their assistance with data collection, as well as the railway planners from Dutch Railways (NS) and ProRail who participated in this research study across hubs in Amsterdam, Eindhoven, Rotterdam, Utrecht and Zwolle.

During the preparation of this manuscript, the authors utilized ChatGPT and Grammarly to enhance readability. Following the use of these tools, the authors thoroughly reviewed and revised the content as necessary and take full responsibility for the content of this publication.

References

- Aickelin, U., Maadi, M., & Khorshidi, H. A. (2022). Expert–Machine Collaborative Decision Making: We Need Healthy Competition. *IEEE Intelligent Systems*, 37(5), 28–31. <https://doi.org/10.1109/MIS.2022.3208725>
- Alvarado-Valencia, J. A., & Barrero, L. H. (2014). Reliance, trust and heuristics in judgmental forecasting. *Computers in Human Behavior*, 36, 102–113. <https://doi.org/10.1016/j.chb.2014.03.047>
- Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY*, 35(3), 611–623. <https://doi.org/10.1007/s00146-019-00931-w>
- Bailey, P. E., Leon, T., Ebner, N. C., Moustafa, A. A., & Weidemann, G. (2023). A meta-analysis of the weight of advice in decision-making. *Current psychology (New Brunswick, N.J.)*, 42(28), 24516–24541. <https://doi.org/10.1007/s12144-022-03573-2>
- Baines, J. I., Dalal, R. S., Ponce, L. P., & Tsai, H.-C. (2024). Advice from artificial intelligence: A review and practical implications. *Frontiers in Psychology*, 15. <https://doi.org/10.3389/fpsyg.2024.1390182>
- Bankuoru Egala, S., & Liang, D. (2024). Algorithm aversion to mobile clinical decision support among clinicians: A choice-based conjoint analysis. *European Journal of Information Systems*, 33(6), 1016–1032. <https://doi.org/10.1080/0960085X.2023.2251927>
- Bayer, S., Gimpel, H., & Markgraf, M. (2022). The role of domain expertise in trusting and following explainable AI decision support systems. *Journal of Decision Systems*, 32(1), 110–138. <https://doi.org/10.1080/12460125.2021.1958505>
- 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. <https://doi.org/10.1007/s12599-020-00678-5>
- 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. <https://doi.org/10.1016/j.obhdp.2006.07.001>
- Boonprakong, N., Tag, B., Goncalves, J., & Dingler, T. (2025). How Do HCI Researchers Study Cognitive Biases? A Scoping Review. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–20. <https://doi.org/10.1145/3706598.3713450>
- Bürkner, P.-C. (2017). Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1). <https://doi.org/10.18637/jss.v080.i01>
- 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. <https://doi.org/10.1002/bdm.2155>
- Chacon, A., Kausel, E. E., & Reyes, T. (2022). A longitudinal approach for understanding algorithm use. *Journal of Behavioral Decision Making*, 35(4), e2275. <https://doi.org/10.1002/bdm.2275>
- Chen, V., Liao, Q. V., Wortman Vaughan, J., & Bansal, G. (2023). Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. *Proceedings of the ACM on Human-Computer Interaction*, 7, 1–32. <https://doi.org/10.1145/3610219>
- Daschner, S., & Obermaier, R. (2022). Algorithm aversion? On the influence of advice accuracy on trust in algorithmic advice. *Journal of Decision Systems*, 31, 1–21. <https://doi.org/10.1080/12460125.2022.2070951>
- 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. <https://doi.org/10.1037/xge0000033>
- Elkins, A. C., Dunbar, N. E., Adame, B., & Nunamaker, J. F. (2013). Are Users Threatened by Credibility Assessment Systems? *Journal of Management Information Systems*, 29(4), 249–262. <https://doi.org/10.2753/MIS0742-1222290409>
- Feng, X., & Gao, J. (2020). Is optimal recommendation the best? A laboratory investigation under the newsvendor problem. *Decision Support Systems*, 131, 113251. <https://doi.org/10.1016/j.dss.2020.113251>
- Haaf, J. M., & Rouder, J. N. (2019). Some do and some don't? Accounting for variability of individual difference structures. *Psychonomic Bulletin & Review*, 26(3), 772–789. <https://doi.org/10.3758/s13423-018-1522-x>
- Hayes-Roth, B., & Hayes-Roth, F. (1979). A Cognitive Model of Planning. *Cognitive Science*, 3(4), 275–310. https://doi.org/10.1207/s15516709cog0304_1
- Hertz, N., & Wiese, E. (2019). Good advice is beyond all price, but what if it comes from a machine? *Journal of Experimental Psychology: Applied*, 25(3), 386–395. <https://doi.org/10.1037/xap0000205>

- Himmelstein, M. (2022). Decline, adopt or compromise? A dual hurdle model for advice utilization. *Journal of Mathematical Psychology*, 110, 102695. <https://doi.org/10.1016/j.jmp.2022.102695>
- Kaufmann, E., Chacon, A., Kausel, E. E., Herrera, N., & Reyes, T. (2023). Task-specific algorithm advice acceptance: A review and directions for future research. *Data and Information Management*, 7(3), 100040. <https://doi.org/10.1016/j.dim.2023.100040>
- Kinney, A. R., Middleton, A., & Graham, J. E. (2021). A Bayesian analysis of non-significant rehabilitation findings: Evaluating the evidence in favour of truly absent treatment effects. *Annals of Physical and Rehabilitation Medicine*, 64(4), 101425. <https://doi.org/10.1016/j.rehab.2020.07.008>
- Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, 90–103. <https://doi.org/10.1016/j.obhdp.2018.12.005>
- Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*, 46(4), 629–650. <https://doi.org/10.1093/jcr/ucz013>
- 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. <https://doi.org/10.1016/j.techfore.2021.121390>
- Mesbah, N., Tauchert, C., & Buxmann, P. (2021). Whose Advice Counts More – Man or Machine? An Experimental Investigation of AI-based Advice Utilization. <https://doi.org/10.24251/HICSS.2021.496>
- Özüdoğru, T., Janssen, C. P., Renooij, S., & van Maanen, L. (2025). *AI-Assisted Decision-Making in Scheduling: The Role of Expertise and Witnessing AI Failure [Unpublished manuscript]*.
- Pescetelli, N., Hauperich, A.-K., & Yeung, N. (2021). Confidence, advice seeking and changes of mind in decision making. *Cognition*, 215, 104810. <https://doi.org/10.1016/j.cognition.2021.104810>
- Pinedo, M. L. (2016). *Scheduling*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-26580-3>
- Prahl, A., & Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, 36(6), 691–702. <https://doi.org/10.1002/for.2464>
- Schaap, G., Bosse, T., & Hendriks Vettehen, P. (2024). The ABC of algorithmic aversion: Not agent, but benefits and control determine the acceptance of automated decision-making. *AI & SOCIETY*, 39(4), 1947–1960. <https://doi.org/10.1007/s00146-023-01649-6>
- Selten, F., Robeer, M., & Grimmelhuijsen, S. (2023). ‘Just like I thought’: Street-level bureaucrats trust AI recommendations if they confirm their professional judgment. *Public Administration Review*, 83(2), 263–278. <https://doi.org/10.1111/puar.13602>
- Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others’ opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(3), 780–805. <https://doi.org/10.1037/a0015145>
- Soll, J. B., & Mannes, A. E. (2011). Judgmental aggregation strategies depend on whether the self is involved. *International Journal of Forecasting*, 27(1), 81–102. <https://doi.org/10.1016/j.ijforecast.2010.05.003>
- Van Den Broek, R., Hoogeveen, H., Van Den Akker, M., & Huisman, B. (2022). A Local Search Algorithm for Train Unit Shunting with Service Scheduling. *Transportation Science*, 56(1), 141–161. <https://doi.org/10.1287/trsc.2021.1090>
- Van Wezel, W., & Jorna, R. (2009). Cognition, tasks and planning: Supporting the planning of shunting operations at the Netherlands Railways. *Cognition, Technology & Work*, 11(2), 165–176. <https://doi.org/10.1007/s10111-007-0106-8>
- Yaniv, I. (2004). Receiving other people’s advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, 93(1), 1–13. <https://doi.org/10.1016/j.obhdp.2003.08.002>
- Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, 32(4), 403–414. <https://doi.org/10.1002/bdm.2118>