Advised by an Algorithm: Learning with Different Informational Resources and Reactions to Heterogeneous Advice Quality*

Jan Biermann†, John J. Horton‡, Johannes Walter§

February 14, 2025

Abstract

In a wide range of settings, decision-makers increasingly rely on algorithmic tools for support. Often, the algorithm serves as an advisor, leaving the final decision to be made by human judgment. In this setting, we focus on two aspects: first, identifying the informational resources that aid individuals in evaluating algorithmic guidance, and second, exploring human reactions to varying qualities of algorithmic advice. To address these questions, we conducted an online experiment involving 1565 participants. In the baseline treatment, subjects repeatedly perform the same estimation task and are provided with algorithmic guidance, all without knowledge of the type of algorithm or feedback after each round. Subsequently, we introduce two interventions aimed at enhancing the quality of human decisions when receiving algorithmic advice. In the first intervention, we explain the way the algorithm functions. We find that while this intervention reduces adherence to algorithmic advice, it does not improve decision-making performance. In the second treatment, we disclose the correct answer to the task after each round. This intervention leads to a reduction in adherence to algorithmic advice and an improvement in human decision-making performance. Furthermore, we investigate the extent to which individuals can adjust their assessment of the algorithm when advice quality fluctuates due to external circumstances. We find some evidence that individuals can assess algorithmic advice thoughtfully, adjusting their adherence depending on the quality of algorithmic recommendations.

Keywords: Human-algorithm decision making, algorithmic advice, learning.
JEL Codes: C91, D79, D80, M21, O30

*Valuable discussions with the following people have greatly improved this paper: Adrian Hillenbrand, Kris Johnson Ferreira, Ben Green, Rudi Kerschbamer, Lydia Mechtenberg, Robert Dur, Dominik Rehse and Marco Schwarz. We also want to thank the participants of TUHH Institute for Digital Economics Seminar 2021, ZEW Digital Economy Seminar 2022, YEM 2022, ASFEE 2022, UHH Collective Decision-Making PhD Seminar 2023 and Innsbruck Spring Summit on (Un)Ethical Behavior in Markets 2023.

†University of Hamburg, jan.biermann@uni-hamburg.de.

‡Massachusetts Institute of Technology & National Bureau of Economic Research, jjhorton@mit.edu.

§ZEW – Leibniz Centre for European Economic Research & Karlsruhe Institute of Technology, johannes.walter@zew.de.

1 Introduction

Human decision makers are increasingly being guided by algorithmic recommendation systems. This trend is evident in various fields, such as healthcare, where physicians use these systems to determine the most suitable treatments for patients, in the judiciary, where judges rely on them for sentencing decisions, and in e-commerce, where pricing managers utilize them to strategically set discounts for products. The focus of our paper is on the ability to assess algorithmic advice, as this skill becomes critical whenever a human holds the final decision-making authority. Our work provides experimental evidence examining which types of informational support help people to learn to evaluate algorithmic advice. First, we study an algorithm with a stable advice quality. Second, we analyze people's assessments under varying algorithmic performance.

To do so, we conduct an online experiment and ask 1565 participants to estimate how many dots are in an image. Our subjects receive algorithmic advice and are free to choose to what extent—if at all—they want to incorporate this advice in their answer. The task is repeated for 16 rounds, and the image has so many dots that counting them is infeasible. One important ongoing debate in the literature on human-AI decision-making focuses on a human's ability to recognize and correct an algorithm when it malfunctions. We therefore study an algorithm which – under certain conditions – considerably underestimates the number of dots.

In the first part of the study, we test two interventions designed to improve a human's ability to optimally incorporate algorithmic advice. First, we provide participants with an explanation of how the algorithm arrives at its recommendation. Second, we reveal the solution to the prediction task after each round. This allows participants to better assess the algorithm's performance as well as their own ability. We also test a combination of both interventions. These interventions speak to two different concepts of how people could learn to assess the algorithm: learning by thought or learning by experience (Myagkov and Plott, 1997). In the former, participants could learn by receiving abstract information about how the algorithm works, while in the latter, they could learn by directly experiencing the consequences of relying on the algorithm. Our study transfers these concepts of learning to our domain of interest, which is decision-making under algorithmic guidance.

The recommendation quality of our algorithm depends on the interaction with the inputs (or environment) in which it operates. In the first part of our paper, we design the inputs (the dot images) in a way that leads to biased recommendations. In the second part, we also consider cases in which the same algorithm makes accurate predictions due to different inputs. Hence, our algorithm is not biased per see; it rather delivers predictions of different qualities depending on the context. This mimics many real-world settings in which it is impossible for the developers of an algorithm to foresee all possible inputs or whether there

will be a distributional shift in the input data. For example, Obermeyer et al. (2019) have shown that a health care algorithm advising physicians on which patient should receive a more expensive treatment was biased for Black patients, but not for White patients.

The focus of the second part of the paper is on evaluating individuals' ability to account for contextual influences on advice quality. Hence, we expose subjects to a varying performance of the algorithm due to changing inputs. Given that that algorithms are not equally suited for every environment, adjusting adherence to the algorithm based on the contextual circumstances becomes critical to achieving good outcomes.

Regarding our first intervention, we find that providing an explanation of the algorithm decreases participants' algorithm adherence, but it does not improve guessing performance. We also find evidence that it has an adverse effect on the performance of some participants. Informing participants of the true number of dots at the end of the round also makes participants follow the algorithm less, but in this case, it improves participants' guessing performance. Our analysis focuses on how the treatments influence outcomes compared to the baseline. We remain cautious interpreting the absolute levels (of adherence to the algorithmic advice and the resulting performance) as these are strongly contingent on the task and context.[1]

Regarding the second part of our experiment, we find evidence that participants adjust their behavior to the changing environment and the resulting varying quality of the algorithmic advice.[2] They put more weight on the advice when the algorithm produces good recommendations. Hence, they are – to some extent – capable of viewing the algorithm in a nuanced way and alternate between relying more on the algorithmic advice and more on their own assessment. In addition, adherence to algorithmic advice in an unfavorable environment does not depend on whether subjects encounter the algorithm only in that context or also in a favorable setting.

There is already a sizable and rapidly growing literature exploring trust in algorithmic predictions and the willingness to follow them. While it has been shown that AI is capable of outperforming humans in various tasks (Lai et al., 2021; Kleinberg et al., 2018), there have been prominent cases illustrating that relying on dysfunctional algorithmic advice can have detrimental consequences (Angwin et al., 2016). The ability to recognize biased algorithmic advice becomes critical.[3]

[1] Existing work has e.g. shown that individuals' willingness to trust algorithms depends on whether the task is (perceived as) objective (Castelo, Bos, and Lehmann, 2019) and whether the decision has a moral component (Bigman and Gray, 2018).

[2] In this part of the study, we provide our participants with both types of decision-making support from the first part of the study: explanation and feedback.

[3] Several existing studies have compared how decision makers react to advice depending on whether it originates from an algorithm or from a human counterpart and conclude that decision makers' responses vary significantly (Önkal et al., 2009; Dietvorst, Simmons, and Massey, 2015; Prahl and Van Swol, 2017; Dietvorst, Simmons, and Massey, 2018; Logg, Minson, and Moore, 2019; Prahl and Van Swol, 2021; Sele and Chugunova,

Existing field experiments have provided evidence that people don't optimally incorporate AI advice in their decision making and don't effectively update their beliefs about AI performance (Glaeser et al., 2021; Agarwal et al., 2023). This illustrates that humans need further assistance when incorporating algorithmic advice in their decision-making.

In the field of computer science, several studies have empirically explored such assistance interventions and how to calibrate trust in algorithms such as stating accuracy of the algorithm (Yin et al., 2019) or providing confidence scores and local explanations (Zhang, Liao, and Bellamy, 2020; Alufaisan et al., 2021). It has also been shown that a slower response time of an algorithm can enhance human evaluation of its performance (Park et al., 2019). Most closely related to our work from this angle is Green and Chen (2019) who report experimental results of how people react to different aids, including two interventions that are similar to our treatments. However, Green and Chen (2019) do not explore heterogeneous performance, do not discuss different ways of learning and study a quite different experimental set-up and task (pretrial release and financial lending) than we do.

Related to our investigation of how people react to varying algorithmic performance, several studies have looked at how people react when they see the algorithm make mistakes. Dietvorst, Simmons, and Massey (2015) have shown that people tend to abandon algorithmic advice after seeing algorithms err. Several subsequent studies have explored different nuances of this seminal finding (e.g. Prahl and Van Swol, 2017; Dietvorst, Simmons, and Massey, 2018; Jung and Seiter, 2021; Zhang and Gosline, 2022; Reich, Kaju, and Maglio, 2023).

A number of studies from the field of decision-making under risk are related insofar as they explore the two types of learning which we examine in our work through the interventions. Myagkov and Plott (1997) are the first to distinguish between learning by thought only (i.e. without feedback) and learning by thought and experience (i.e. with feedback). They argue that irrationalities should decrease over time but only when subjects experience the consequences of their choices. The difference between these two has been empirically analyzed by several researchers. Kuilen and Wakker (2006) and Kuilen (2009) find that learning by thought and experience leads to more rational behavior while learning by thought only does not. Conversely, Hey (2001), Birnbaum and Schmidt (2015), and Nicholls, Romm, and Zimper (2015) report that learning by thought alone is sufficient to promote more rational behavior over time. In sum, there is mixed empirical evidence regarding the sufficiency of learning by thought. Building on these findings, we apply this framework to our domain of interest, which is the ability to evaluate algorithmic advice. While the importance of feedback and directly experiencing the consequences has been discussed in other domains, providing feedback has not received a lot of attention from policymakers in the context of

2022). Examining the differences of human and algorithmic advice is therefore not the focus of our work. We rather acknowledge a fundamental difference as being established and ask the question: Given that the advice comes from an algorithm, which tools help decision makers to better assess this advice?

AI regulation. For example, while article 14 in the EU AI Act (European Commission, 2021) discusses several measures to ensure human oversight (e.g. appropriate training and explanation), it does neither implicitly nor explicitly mention feedback.

We contribute to the existing literature in two major ways. First, we test interventions designed to improve people's assessment of algorithmic advice. Importantly, our interventions are informed by the existing work from decision-making under risk. We introduce the concept of learning through thought vs. learning through experience to the literature on empirical human-AI decision-making. By doing so, we bring more nuance to the debate on how people learn about algorithms.

Second, we investigate the extent to which individuals can skillfully adjust their evaluations of the algorithm in response to varying levels of advice quality. We create a scenario in which algorithmic performance fluctuates based on external factors (i.e. we analyze a stable algorithm in a changing setting resulting in a varying performance). Our contribution is to illustrate that subjects do not inevitably abandon the algorithm after seeing it make poor predictions. If the low-quality advice is caused by the algorithm being unsuitable for a certain setting, humans will continue to use it under different circumstances.

Further, our study is innovative with regard to the experimental task we employ. To the best of our knowledge, we are the first to investigate algorithm-supported decision-making through the dot-guessing task. This task is attractive for several reasons. It is accessible to laypeople without specialized expertise. The general skill of counting dots is innate to all humans (except given certain medial conditions). Further, employing this task, enables us to vary the quality of algorithmic advice by altering external conditions (i.e. the distribution of dots) while keeping the functionality of the algorithm unchanged.

The remainder of this paper proceeds as follows: We subdivide the main part of the paper into two parts. Section 2 investigates the effects of providing informational resources (part 1). Section 3 examines our treatments containing varying performance (part 2). Both parts contain subsection explaining the experimental design and presenting the results. Section 4 discusses the findings from both parts. Section 5 concludes.

2 Part I: Aids to Better Assess Algorithms

2.1 Experimental Design

2.1.1 Experimental Task and Algorithm

The central component of our experiment is the dot-guessing task. Subjects see images showing a large number of blue dots and are asked to guess how many dots they think each of the images contains. Examples of dot images are shown in figure 2. The number of dots in the images is chosen randomly and varies between 942 and 3084 dots. Participants have

60 seconds to make their guesses, making it infeasible to count the dots in the image.[4]

Every round of the experiment consists of three stages: In the first stage, participants see the image for the first time and submit their guesses. In the second stage, subjects see the same image again and additionally receive an algorithmic prediction of the number of dots, before submitting a new guess. We call the two entries the subjects make initial guess $guess_i$ and revised guess $guess_r$, both of which are incentivized. In the third stage, participants can see their guesses $guess_i$ and $guess_r$ and some additional information depending on the treatment. They do not take any action at this stage. Every subject plays 16 rounds, each round including a new image and a new recommendation. In every round, all participants see the same image (i.e. the same number of dots) and receive the same recommendation. We employ a between-subject design and randomize our participants on an individual level.

The algorithm we employ samples three subareas from the given image, calculates the average of dots within these subareas, and extrapolates this average to the entire surface of the image. We introduce a bias to the predictions of the algorithm in the following way: We restrict the sampling of the algorithm to the outer edges of the image and in addition we choose a triangular distribution with a denser center and fewer dots at the edges. As a result, the algorithm systematically underestimates the number of dots in the entire image.[5] See figure 2 for further illustrations.

2.1.2 Description of Treatments

The treatments vary along two dimensions: explaining how the algorithm arrives at its prediction and revealing the true answer after the revised guess has been recorded. The design of our interventions speaks to whether humans learn by thought only or by thought and experience (Myagkov and Plott, 1997) and is in this respect related to several empirical studies which apply this concept to test expected utility-theory (Hey, 2001; Kuilen and Wakker, 2006; Kuilen, 2009; Birnbaum and Schmidt, 2015; Nicholls, Romm, and Zimper, 2015).

In the treatments Explanation and Explanation&Feedback, participants receive an explanation of how the algorithm works. This includes a visual and a textual component. Participants see the image of dots overlaid with squares indicating the subareas the algorithm samples from (cf. figure 2). We also inform them in writing that the algorithm counts the number of dots in the visualized squares and then predicts the total number of dots in the entire area based on this sampling. Hence, we do not explicitly state to participants that the predictions are biased, but we deem that our explanation provides the necessary information

[4]Our task is rooted in the tradition of Galton (1907). His research has produced the "wisdom of the crowd" finding and involved a contest in which people guessed the weight of a butchered ox.

[5]Images with uniform distributions will be employed in the second part of the experiment resulting in unbiased algorithmic predictions.

Figure 1: Visualization experimental design part I

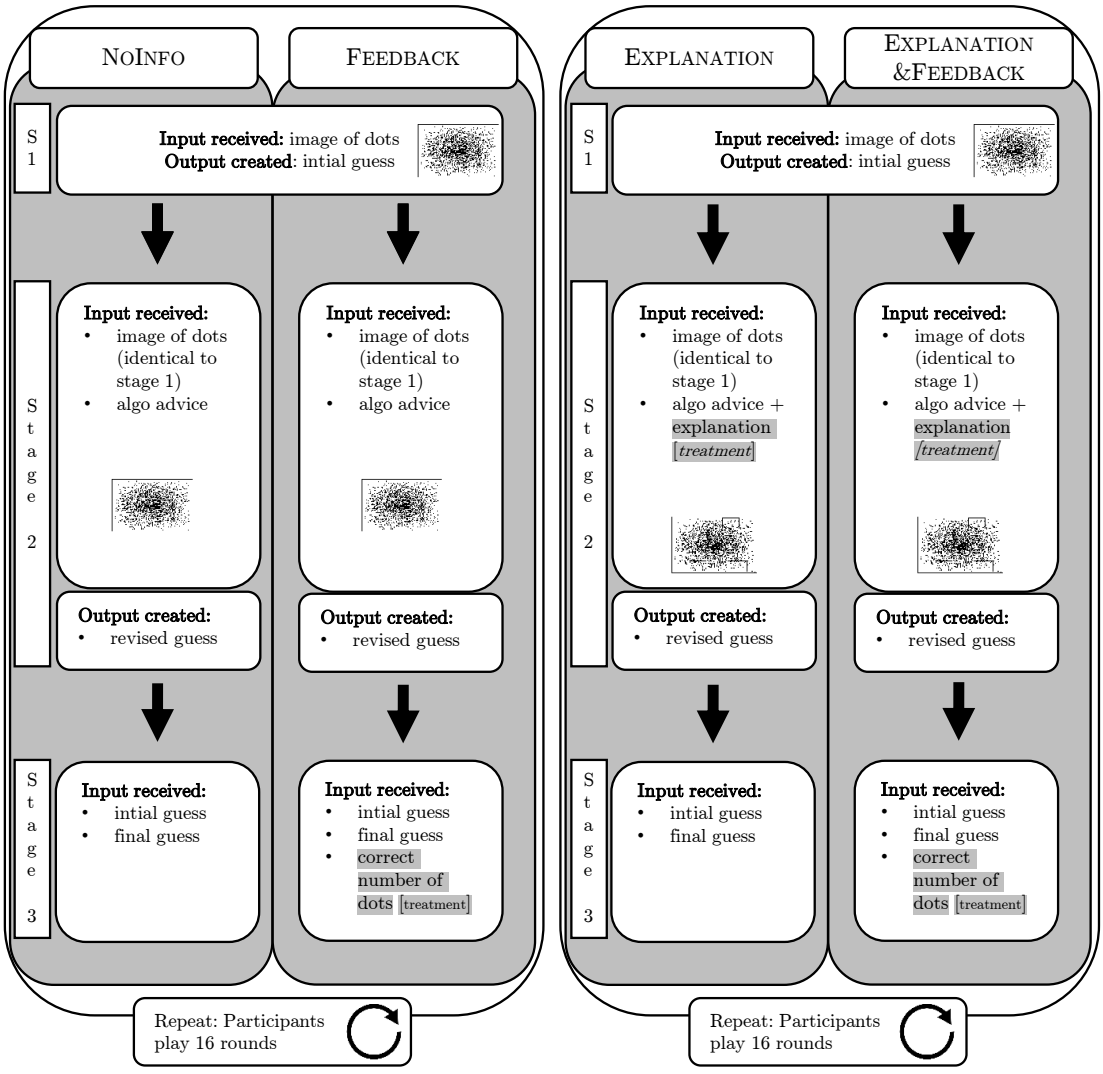
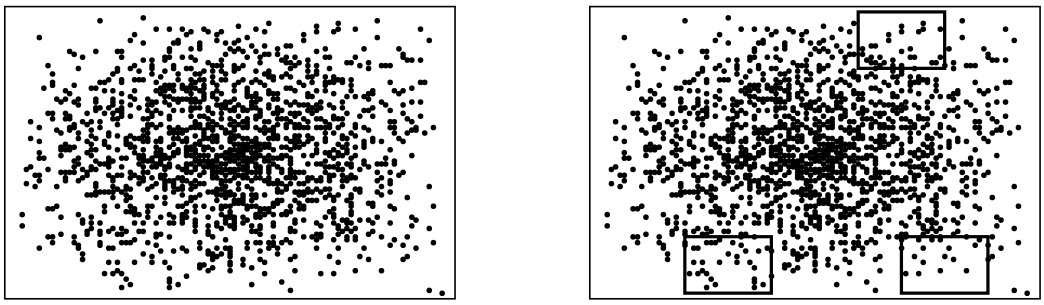


Figure 2: Functioning dot guessing algorithm



(a) Example dot image in treatments in treatments without explanation

(b) Example dot image in treatments with explanation

Notes: The algorithm arrives at its prediction by first randomly sampling three subareas from the edges of each image and counting the number of dots within each subarea. It then calculates the average number of dots over areas and projects this average to the entire image. Importantly, the algorithm always samples the three subareas at the edges of a given image, never from the center. Through the combination of triangular dot distribution and sampling being restricted to the edges, we introduce a bias in the algorithm prediction. The image in panel (a) is an example of a dot image that all participants see in the first stage of the experiment. Panel (b) shows the same image but this time overlaid with the rectangular subareas from which the algorithm samples dots. Only participants in the treatments with explanation see this image from panel (b) in the second stage, complementing the verbal algorithm explanation.

to comprehend that the algorithm vastly underestimates the number of dots.[6] See appendix for experimental interfaces containing the exact wording.

Furthermore, the treatments involving feedback, participants receive the information about the correct number of dots in the image at the end of every round. In treatments without feedback, participants never find out the correct answer to the task. Seeing the solution provides an opportunity to assess the performance of the algorithm in a specific round. It also opens the chance to learn about one's own performance.

To sum up, the two dimensions result in the four treatments NoInfo, Explanation, Feedback, Explanation&Feedback. These treatments enable us to analyze which type of information empowers humans to effectively asses the quality of algorithmic advice and how they influence task performance. The experimental design is visualized in figure 1.

[6]This information is provided in the second stage of each round, i.e. when participants state their revised guess.

2.1.3 Payment Scheme and Experimental Procedure

Our subjects receive a flat fee of \$0.9 for completing the study. Guesses are incentivized as follows: Participants receive \$0.15 for a perfect guess, and this bonus is diminished by \$0.0002 for each point difference, resulting in a bonus if a participant's guess is within the range of +/- 749 dots from the true answer.[7] The experiment contains 16 rounds and each of the rounds involves two incentivized guesses (initial and revised). Therefore, subjects could earn a maximum of \$4.80 bonus payment in addition to the flat fee.

We conducted our online experiment in December 2021. The experiment was developed using the software oTree (Chen, Schonger, and Wickens, 2016). We recruited our subjects via Amazon's crowd-working platform Mechanical Turk (MTurk). All of them are based in the US, have completed at least 500 tasks on MTurk, and have an approval rate of at least 95%. We conducted five sessions (two sessions with 200 and 3 sessions and 400 participants). Our data contains 1565 observations in total. 1263 of them are relevant for the four treatments in part I and an additional treatment containing 302 observations will be introduced in part II. On average, participants have taken 14 minutes and 18 seconds to complete the study and have earned \$2.33. This translates to a hypothetical hourly wage of \$9.82.

2.2 Data

The main data we elicit from our subjects is their guesses with respect to the number of dots in the image they see. When eliciting these guesses, we do not set an upper bound (e.g. by employing a slider) as such an upper bound would serve as an orientation point for some of our subjects. As a result, participants can enter very high numbers, and in fact, some choose to do so. We, therefore, see large outliers in our distributions. Three approaches are common to address the issue of outliers in the data: top-coding, winsorizing, and taking the natural logarithm. We employ the latter method. Similarly, although some participants state a very low guess for the number of dots, including 0, we do not exclude these guesses at the lower end of the distribution either. This approach has the advantage of not requiring us to exclude any observations from our analysis. Instead, we can show that excluding very low guesses does not change our main results. For a more in depth discussion of this issue see section A in the appendix.

Further, since the number of dots and the algorithmic advice changes every round, examining (the logarithm of) the guesses at their face value would have little meaning. We are rather interested in the relation between guesses and i.) the algorithmic advice or ii.) the correct number of dots. Therefore, we focus on two main outcome variables throughout the

[7] The first visual impact of the images might be disheartening for certain participants. Hence, we deliberately selected a broad range (almost 1500 dots) within which participants receive a bonus. This is done to encourage our subjects to maintain an aspiration for receiving a bonus and to motivate them to strive for accurate estimations.

paper: algorithm adherence, calculated as $|log(algo) - log(guess)|$, and guessing performance, calculated as $|log(truth) - log(guess)|$.

Various parts of the analysis are based on a comparison among treatments in which case we pool all rounds together. We recognize that the guesses of each individual are not independent of each other. We, therefore, preprocess the data by conducting the mean of the values of interest (e.g. distance to algorithmic recommendation) of all 16 rounds for each individual.

2.3 Results

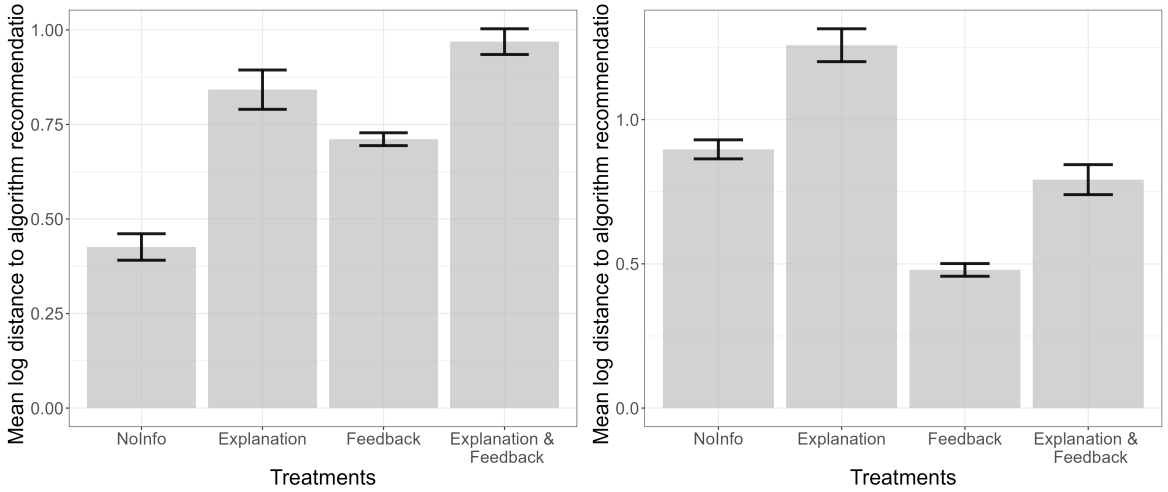
In this section we report results regarding how different aids to better assess the algorithm influence performance and algorithm adherence. More specifically, we are interested in whether revealing the correct answer at the end of the round or an providing explanation of the mechanics of the algorithm can help subjects learn from the interaction with the algorithm. First, we examine the effects on algorithm adherence by examining the distance between the average revised guess and the algorithmic prediction for each treatment. Figure 3a illustrates that feedback reduces algorithm adherence compared to the baseline treatment.[8] The same is true for explanations with an even stronger effect. We observe the strongest reduction of algorithmic adherence in the treatment Explanation&Feedback.

Result 1a: *Explanation reduces algorithm adherence (medium effect).*
Result 1b: *Revealing the truth reduces algorithm adherence (weakest effect).*
Result 1c: *Combining explanation and the revealing truth reduces algorithm adherence (strongest effect).*

Knowing that both interventions reduce algorithm adherence, we now turn to the question of how the treatments influence guessing performance. We hence examine the distance between the revised guess and the correct answer. We present the results in figure 3 and table 3 in the appendix. Based on figure 3, providing the explanation increases the average distance to the true number of dots compared to NoInfo, i.e. the explanation makes participants perform worse.[9] While this result is striking, we remain cautious about over-interpreting this finding. First, we offer a robustness check and find that the negative effect does not disappear, but is starkly reduced (cf, appendix A). Second, our experiment is not designed to uncover the underlying mechanisms of this result.[10] For these two reasons, we err on the side of caution in the main body of the paper in proposing the interpretation that "explanation does not improve performance".

[8] Table 2 in the appendix contains more detailed, numeric information regarding algorithm adherence.
[9] Compare section 2.1 for more information on the content of the explanation.
[10] However, we do provide a discussion of potential drivers of this result in appendix B.

Figure 3: Mean distance to the algorithm and the true number of dots by treatment



(a) Mean distance to the algorithm recommendation by treatment.

(b) Mean distance to the true number of dots by treatment.

Notes: The bar graph in panel (a) illustrates the treatment effects on algorithm adherence (mean distance of the revised guesses to the algorithm recommendation per treatment). The numerical treatment effects on algorithm adherence can be found in table 2 in the appendix. The bar graph in panel (b) illustrates the treatment effects on guessing performance (mean distance of the revised guesses to the true number of dots per treatment). The numerical treatment effects on guessing performance can be found in table 3 in the appendix. The barplots also include the standard errors around the mean. For the outcome distance to the algorithm recommendation, all three treatments are significantly different from the baseline NoInfo treatment on a 1% level based on a independent two-sample t-test. For the outcome distance to the true number of dots, only treatment Explanation&Feedback is *not* significantly different from the baseline on a 1% level. We pre-process the data by taking log values and calculating the mean over all 16 rounds for each individual. For more details on data pre-processing see section 2.2.

In contrast, revealing the truth improves guessing performance. Remarkably, the effects of these two treatments have approximately the same size (and opposite directions). As a result, in treatment Explanation&Feedback, the two effects appear to cancel each other out. Hence, the average performance under Explanation&Feedback is statistically indistinguishable from NoInfo (t-test is not significant on a 5%-level).

Result 2a: *Explanation does not improve performance (and possibly hurts).*
Result 2b: *Feedback improves performance.*
Result 2c: *Combining explanation and feedback does not significantly change performance compared to the baseline treatment.*

In appendix E we also provide an additional analysis of the baseline treatment. It suggests that without any additional aids, most subjects are not capable of assessing the quality of the algorithm. One might therefore consider algorithmic advice to be a credence good. However, we do not put too much emphasis on this analysis because such results are highly dependent on the specific task and setting. Addition discussions and interpretations of the results from this section follows in section 4.

3 Part II: Reactions to Heterogeneous Performance Caused by Varying Circumstances

3.1 Experimental Design

The second part explores how people react to varying performance of the algorithm. To this end, we introduce an additional treatment which we call VaryingQuality. The experimental set-up is largely the same as in the first part: Participants play the dot guessing game for 16 rounds, they submit initial guesses, observe an algorithmic recommendation and then submit their revised guess. The appearance of the interfaces and the incentive structure are identical to the first part. Importantly, in this treatment our participants receive both informational resources: the explanation of the algorithm and the solution of how many dots were in the image at the end of each round as in Explanation&Feedback from the first part. Therefore, we use Explanation&Feedback as a benchmark in this section.

The difference to Explanation&Feedback is that participants in VaryingQuality see only 8 images in which the dots follow a triangular distribution. The other 8 images show dots follow an uniform distribution. Each participant in this treatment sees an image where the dots are triangularly distributed in every odd round and an image with uniformly distributed dots in every even round. The 8 dot images with the triangular distribution are the same as the first 8 images in the four treatments in Explanation&Feedback. Figure

5 shows two exemplary images for each type of distribution.

Changing the distribution has consequences on the performance of the advising algorithm. Since the algorithm always samples from the edges of an image, it draws a biased sample in case of a triangular distribution. In contrast, when the dots are uniformly distributed, the sampled areas are representative for the whole image and this results in precise algorithmic predictions. This setting mimics many real-world scenarios the algorithm's functionality remains consistent; however, the varying settings result in the algorithm's techniques yielding either accurate or inaccurate predictions. The experimental design is visualized in figure 4.

3.2 Results

Do our participants disengage with the algorithm after seeing it perform poorly? Or do they realize when they can rely on the algorithm and when they better ignore it? Figure 6 shows the distance of the revised guesses to the algorithm recommendation averaged over three different sets of rounds: the leftmost bar shows the average over the 8 rounds in which the algorithm's advice was of poor quality. The bar in the center shows this distance averaged over the other 8 rounds in which the algorithm's advice was of high quality. Finally, the rightmost bar shows the average over all 16 rounds in Explanation&Feedback which is our benchmark.

In rounds with high advice quality (uniform distribution), subjects follow the algorithm more closely than rounds with low advice quality (triangular distributions). Our participants appear to be able to appropriately react to fluctuations in advice quality. This might strike as being at odds with previous literature that has shown that people do not forgive algorithmic mistakes (Dietvorst, Simmons, and Massey, 2015). We interpret this discrepancy in the discussion section.

Result 3: *People follow the algorithmic advice more in settings when the algorithm gives high-quality advice compared to setting with low-quality advice.*

Further, there is no statistically significant difference between distance to the algorithm when advice quality is poor in VaryingQuality compared to Explanation&Feedback. In other words, there are no spill-overs of trust from the high-quality advice rounds to the low-quality advice rounds in VaryingQuality. Participants' assessment of the algorithm in poor-advice rounds is not influenced by experiencing the algorithm in a favorable setting.

Result 4: *Seeing the algorithm perform well in a favorable setting does not increase trust in the algorithm in other settings (i.e. there are no positive spill-overs).*

Figure 4: Visualization experimental design part II

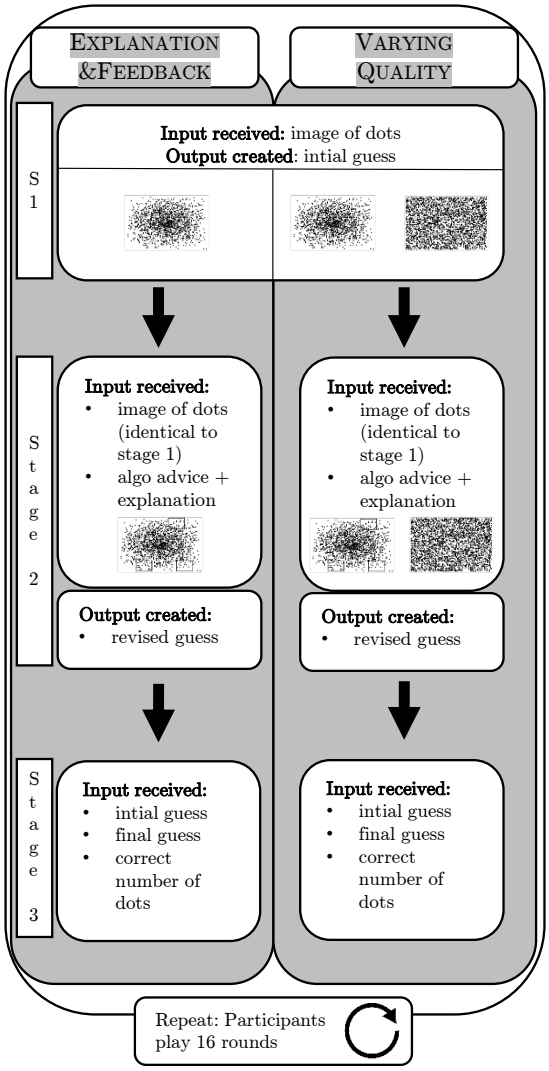
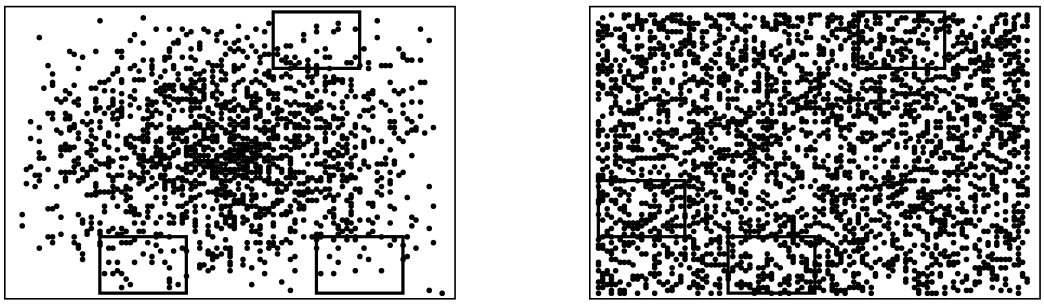


Figure 5: Unifrom and triangular distribution of dots



(a) Triangular dot distribution with boxes

(b) Uniform dot distribution with boxes

Notes: Participants in VaryingQuality see images alternating between triangularly distributed dots as in panel (a) and uniformly distributed dots as in panel (b).

4 Discussion

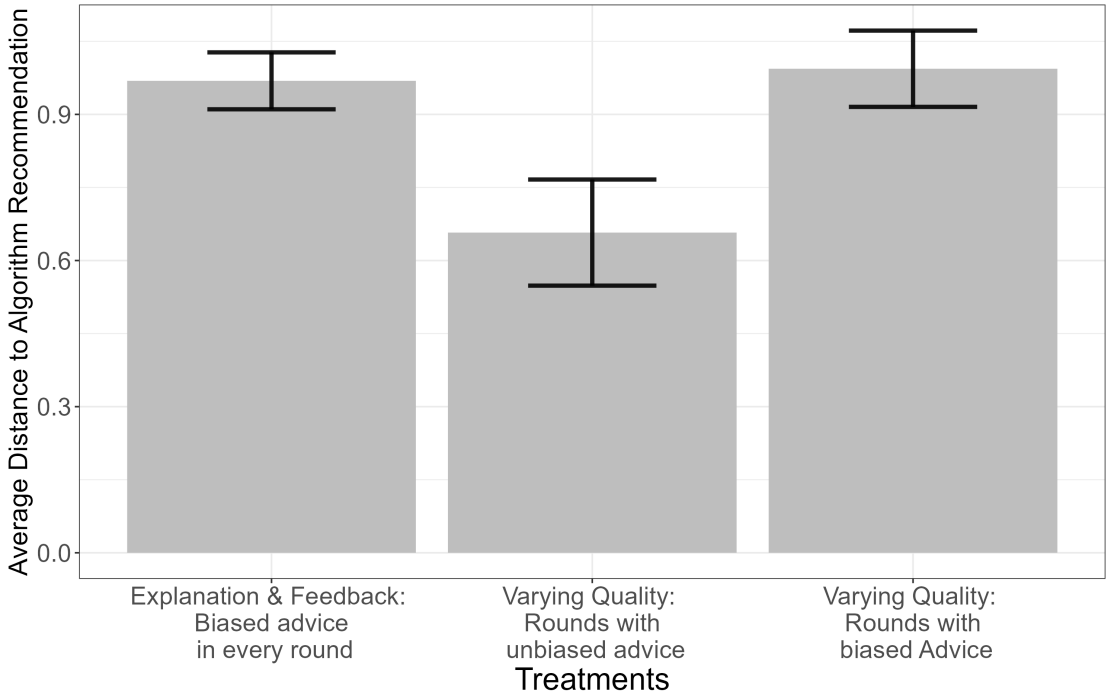
4.1 Aids to Better Assess Algorithms

Following the concept of Myagkov and Plott (1997), we examine whether experience is a necessary component for decision-makers to learn and evaluate the quality of algorithmic support systems. Existing empirical literature from different contexts has provided inconclusive evidence on whether or not learning by thought alone is effective.[11] Addressing this question in the context of algorithms is of vital importance. Consider high economic impact decisions that managers have to make. Learning through thought is always available, as the firm can simulate a series of hypothetical decisions. Learning through experience is not available in some settings, e.g. when a decision occurs only once, the consequences are unclear or only occur after a certain time lag. However, the results from our experiment indicate that learning by thought is not sufficient and experiencing the consequences of (dis-)trusting the algorithm is required in order to improve assessment. Hence, the practical implication of this is that providing timely feedback can help improve decisions and it is worthwhile to strive to implement such feedback in situations where it is feasible.

Compared to NoInfo providing an explanation increases the participants' average distance to the algorithm. This suggests that the explanations help participants to recognize that the algorithmic predictions are biased (in the case of triangular distributions). At the same time, explanation does not improve performance. This is arguably because – after

[11] Compare Hey (2001); Kuilen and Wakker (2006); Kuilen (2009); Birnbaum and Schmidt (2015); Nicholls, Romm, and Zimper (2015).

Figure 6: Log distance from revised guesses to the algorithm recommendation for Explanation&Feedback and VaryingQuality.



Notes: All three bars show the the average distance of the log revised guesses to the algorithm recommendation. The leftmost bar does so for Explanation&Feedback. The middle bar shows this result for VaryingQuality, but only for the 8 rounds in which participants received *un*biased recommendations. The rightmost bar shows this result for VaryingQuality, but only for the 8 rounds in which participants receive *biased* recommendations).

recognizing that algorithmic predictions are biased – participants still don't have a better prediction at their disposal. Taken together, the effects on algorithm adherence and performance can be seen as evidence of the ability to detect a bias, but not the ability to correct a bias.[12]

The absence of a positive impact of explanation on performance is particularly striking. We therefore present an alternative specification as a robustness check in appendix A by excluding unreasonably small guesses. While the negative effect on performance is smaller in this specification, it does not disappear completely. Nevertheless, given the reduction in effect size, we err on the side of caution and conclude that explanation does not improve performance (and possibly even has an adverse effect).

We discuss potential interpretations for why explanation does not improve performance and why it leads to some unreasonably low guesses in appendix B. Overall, it appears that the explanation we have provided has displeased or confused some participants and did not have a positive effect on the average performances. Such heterogeneous treatment effects can also be expected in many real-life applications: While some decision-makers will benefit from such explanations, others will ignore them in face of their busy workday, and still others might feel frustrated when having to read them. This highlights the importance of providing such descriptions in a way that is appropriate for the target user.

Despite discussing potential drivers of this finding, we do not draw any definitive conclusions about the underlying mechanisms since our experiment was not designed for such analysis. Further investigating the patterns of this adverse effect is an interesting direction for future research.

In Feedback participants move further away from the algorithm, but in this treatment, they also move closer to the true number of dots. One potential mechanism of how subjects arrive at their guess is the following: They start with an *orientation point* and adjust this amount of dots based on their judgment. Examining figures 15 to 18 (in the appendix) provides some suggestive evidence that this is indeed an important mechanism. The distribution of initial guesses in the first round is flat. This illustrates that initially subjects are uncertain with regard to their own guess. In addition, participants strongly react to the algorithmic advice. The specific number of dots recommended by the algorithm gives the participants a potential orientation point to use.

The algorithmic advice is an orientation point that is always available when revising the guess. Importantly, in the treatments where people receive feedback, an alternative orientation point is provided: The true answer from the previous round(s). Therefore revealing the true answer does not only offer the opportunity to better access the quality of the algorith-

[12]In principle, explanation could have led to better performance because participants realized that the algorithmic advice is biased, but can be used as a lower bound for a better estimation. However, we don't see evidence for this.

mic advice. It also provides an alternative orientation point. Further, it can also influence participants through a third channel: Providing feedback about their own abilities.

The concept of the orientation point might be considered particularly important after realizing that figures 15 to 18 suggest that participants are remarkably often close to the true answer in treatments where they receive feedback. It appears that participants are strong in taking the true number of dots as an orientation point, estimating the relative change of dots in the next round, and delivering a good estimate for the current round.

Treatment Explanation&Feedback shows an interesting cumulative result: Since providing an explanation and revealing the truth both increase the distance to the algorithm, their combination does as well. And as these treatments individually have opposed effects on guessing performance, their combination seems to annihilate any effect. Guessing performance remains unchanged compared to the baseline. This illustrates that practitioners must be careful when considering tools to improve decision-making. Generally, one cannot simply assume that the more help for the decision-maker, the better.

Despite the algorithmic predictions being biased (given a triangular distribution), it is an empirical question whether it delivers better predictions than our participants. Answering it is relevant because it determines whether participants would have enhanced their performance by incorporating algorithmic advice into their revised guesses. Generally speaking, one can say that there is a sizable proportion of people who have performed better as well as worse than the algorithm. Consequently, there is notable heterogeneity in whether increased or decreased adherence to the algorithm would have resulted in better revised guesses. On average, participants would have gained from assigning a positive weight to the algorithmic advice in most rounds in NoInfo (algorithmic prediction superior in 14/16 rounds) and Explanation (algorithmic prediction superior in 13/16 rounds). Conversely, this is not observed in Feedback (participant guess superior in 14/16 rounds) and Explanation&Feedback (participant guess superior in 12/16 rounds). We provide the corresponding analyis in appendix C.

Another important aspect is whether our interventions show their effect immediately or some repeated interaction is required for the effect to unfold. Overall, there appears to be no pattern that evolves. The interventions do not seem to require warm-up time. One exception are the first rounds of Explanation. Participants move closer to the algorithm in the first two rounds after receiving the advice. In round three, the average distance remains the same. Starting from round four, participants always move further away from the recommendation (cf. figures 15 to 18). This suggests that participants must see the explanation multiple times before the effect starts unfolding.

4.2 Reactions to Heterogeneous Performance Caused by Varying Circumstances

Our result concerning the varying algorithmic performance may seem inconsistent with Dietvorst, Simmons, and Massey (2015). Our participants exhibit nuanced behavior, tending to adhere more to algorithmic advice in scenarios of good performance and less in instances of poor performance. In contrast, Dietvorst, Simmons, and Massey (2015) presented evidence indicating that subjects tend to abandon the algorithm upon observing errors.[13]

We contend that the distinction between the two setups lies in our participants having the opportunity to comprehend the algorithm's suitability for specific settings. We provide our participants with informational resources, aiding their understanding that the algorithm isn't inherently flawed but varies in performance based on the setting. In Dietvorst, Simmons, and Massey (2015), participants lack the opportunity to understand the causes of the observed mistakes.[14]

In this sense, one of our contributions is to investigate the circumstances under which the pivotal result of non-forgiveness towards algorithms holds, and to provide a context for it. Another way to read the Dietvorst, Simmons, and Massey (2015) result is that they document a negative spillover with regard to trust in algorithms: After encountering low-quality advice, individuals tend to distrust the algorithm in other cases. Our study addresses whether there are spillovers in the opposite direction. We find that after witnessing the algorithm perform well in a favorable setting, people do not necessarily trust it more in other cases.

Bringing together our results concerning informational resources and varying algorithmic performance, one can say that FEEDBACK improves overall performance, while EXPLANATION does not. Hence, learning by thought alone is insufficient to improve performance in our context. Given both informational resources, subjects are capable of distinguishing when to follow the algorithmic advice and when to leave it aside. In other words, provided with our explanation, subjects have a hard time assessing the capacities of the algorithm in relation to their own. However, given explanation and feedback, subjects appear capable of assessing the capacity of the algorithm under varying circumstances.

4.3 Caveats and Potential Extensions

Some important caveats are in order. First, we exercise caution in generalizing our findings to all scenarios involving human-machine interaction. We recognize the contextual specificity

[13] Related insights have been made prior to this in the field of ergonomics research. While the "machines" considered in this field are typically very different (e.g. automated alarm systems), fundamental ways of how humans react might still be informative for the research on algorithms and AI. E.g. Madhavan and Wiegmann (2007) have made the argument that decision-makers expect automated advice to be perfect, whereas human advice can be fallible.

[14] The same holds true for relating our study with Bao et al. (2022) who have shown that people are reluctant to follow advice from algorithms that appears inconsistent.

of certain behaviors and focus on interpreting differences between treatments rather than the absolute levels of our variables of interest. Our hope is that future research will explore how our findings translate to different settings.

Second, our emphasis is on analyzing how algorithmic advice influences behavior outcomes (rather than self-reported measures). Our results hint at the fact that subjects hold (overly-)optimistic beliefs regarding the algorithm's capabilities when we do not provide any information regarding the algorithm. Further, we argue that reliance on algorithmic advice is driven by the interplay of two types of beliefs: beliefs about algorithmic performance and beliefs about one's own capabilities. A potentially fruitful avenue for future research involves eliciting and correlating beliefs with behavioral outcomes to better comprehend the underlying mechanisms.

Third, it would be interesting to conduct an additional treatment that includes feedback after each round but does not provide algorithmic advice. This would allow us to separate the different components of providing feedback. In the current setup, by being informed about the correct answer, subjects can infer information about the algorithmic performance and their own performance (and also receive an orientation point).

Fourth, the current study examines the assessment of varying algorithmic performance while providing both information resources that we consider in our work. Exploring how reactions differ when individuals receive only explanations or only feedback in this setting could provide further insights.

Fifth, our study is not designed to pinpoint the reasons for why providing an explanation does not improve performance. While we have suggested some mechanisms, a more in-depth analysis of this phenomenon would be a valuable extension of our work.

5 Conclusion

We design an experiment in which subjects are asked to guess the number of dots they see in an image while receiving advice from an algorithm. We test a set of interventions - providing an explanation of the algorithm, revealing the solution *ex post*, or both - and ask whether they can improve the ability to assess the algorithm.

Our results suggest that providing an explanation of the algorithm does not improve performance and may possibly even hurt performance. Revealing the truth *ex post* on the other hand does benefit people's performance. We are cautious about attributing this solely to participants learning about the algorithm's quality. In this treatment, subjects also have the chance to learn about their own performance and they receive a benchmark for calibrating future guesses. These results suggest that learning by thought appears to be insufficient to improve the assessment of algorithmic quality. In order to learn, people need to experience the consequences of their decisions and hence feedback is required.

Our study further considers a setting of varying algorithmic performance due to changing circumstances. We find that people do not abandon algorithms when errors occur if they can comprehend the reasons behind the errors. Subjects appear to have some ability to understand the strengths and weaknesses of the algorithm and to use it in cases where it proves beneficial.

These findings suggest several practical recommendations. First, in organizational settings, if concrete feedback about previous decisions can be provided and the decision environment does not fundamentally change, managers should seek to disclose the outcome of past decisions. Second, our study indicates that explanations concerning how an algorithm functions must be provided with caution, as they do not necessarily improve human assessment of algorithm quality. This finding is in line with the results from the literature on algorithm explainability: Its ability to improve decisions also depends on various factors, and explanations are no panacea. Third, there is hope that decision-makers can be trusted to assess imperfect algorithms, provided they receive sufficient informational resources for learning and have the opportunity to comprehend the reasons behind algorithmic errors.

References

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2023. "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology." Tech. rep., NBER Working Paper.

Alufaisan, Yasmeen, Laura R Marusich, Jonathan Z Bakdash, Yan Zhou, and Murat Kantarcioglu. 2021. "Does explainable artificial intelligence improve human decision-making?" In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35. 6618–6626.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks." *ProPublica* URL `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

Baksi, Soham and Pinaki Bose. 2007. "Credence goods, efficient labelling policies, and regulatory enforcement." *Environmental and Resource Economics* 37:411–430.

Balafoutas, Loukas and Rudolf Kerschbamer. 2020. "Credence goods in the literature: What the past fifteen years have taught us about fraud, incentives, and the role of institutions." *Journal of Behavioral and Experimental Finance* 26:100285.

Bao, Yongping, Ludwig Danwitz, Fabian Dvorak, Sebastian Fehrler, Lars Hornuf, Hsuan Yu Lin, and Bettina von Helversen. 2022. "Similarity and Consistency in Algorithm-Guided Exploration." .

Bigman, Yochanan E and Kurt Gray. 2018. "People are averse to machines making moral decisions." *Cognition* 181:21–34.

Birnbaum, Michael H and Ulrich Schmidt. 2015. "The impact of learning by thought on violations of independence and coalescing." *Decision Analysis* 12 (3):144–152.

Castelo, Noah, Maarten W Bos, and Donald R Lehmann. 2019. "Task-dependent algorithm aversion." *Journal of Marketing Research* 56 (5):809–825.

Chen, Daniel L., Martin Schonger, and Chris Wickens. 2016. "oTree—An open-source platform for laboratory, online, and field experiments." *Journal of Behavioral and Experimental Finance* 9:88–97.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. 2015. "Algorithm aversion: people erroneously avoid algorithms after seeing them err." *Journal of Experimental Psychology: General* 144 (1):114.

———. 2018. "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them." *Management Science* 64 (3):1155–1170.

Etilé, Fabrice and Sabrina Teyssier. 2016. "Signaling corporate social responsibility: Third-party certification versus brands." *The Scandinavian Journal of Economics* 118 (3):397–432.

European Commission. 2021. "Proposal for Artificial Intelligence Act." URL `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206`.

Galton, Francis. 1907. "Vox populi." *Nature* 75 (7):450–451.

Glaeser, Edward L, Andrew Hillis, Hyunjin Kim, Scott Duke Kominers, and Michael Luca. 2021. "Decision authority and the returns to algorithms." Tech. rep., Harvard Business School Working Paper.

Green, Ben. 2022. "The flaws of policies requiring human oversight of government algorithms." *Computer Law & Security Review* 45:105681.

Green, Ben and Yiling Chen. 2019. "The principles and limits of algorithm-in-the-loop decision making." *Proceedings of the ACM on Human-Computer Interaction* 3:1–24.

Harbaugh, Rick, John W Maxwell, and Beatrice Roussillon. 2011. "Label confusion: The Groucho effect of uncertain standards." *Management science* 57 (9):1512–1527.

Hey, John D. 2001. "Does repetition improve consistency?" *Experimental economics* 4:5–54.

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

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human decisions and machine predictions." *The quarterly journal of economics* 133 (1):237–293.

Kuilen, Gijs van de. 2009. "Subjective probability weighting and the discovered preference hypothesis." *Theory and decision* 67:1–22.

Kuilen, Gijs van de and Peter P Wakker. 2006. "Learning in the Allais paradox." *Journal of Risk and Uncertainty* 33:155–164.

Lai, Vivian, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. "Towards a science of human-ai decision making: a survey of empirical studies." *arXiv preprint arXiv:2112.11471* .

Logg, Jennifer M., Julia A. Minson, and Don A. Moore. 2019. "Algorithm appreciation: People prefer algorithmic to human judgment." *Organizational Behavior and Human Decision Processes* 151:90–103.

Madhavan, Poornima and Douglas A Wiegmann. 2007. "Similarities and differences between human–human and human–automation trust: an integrative review." *Theoretical Issues in Ergonomics Science* 8 (4):277–301.

Myagkov, Mikhail and Charles R Plott. 1997. "Exchange economies and loss exposure: Experiments exploring prospect theory and competitive equilibria in market environments." *The American Economic Review* :801–828.

Nicholls, Nicky, Aylit Tina Romm, and Alexander Zimper. 2015. "The impact of statistical learning on violations of the sure-thing principle." *Journal of Risk and Uncertainty* 50:97–115.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting racial bias in an algorithm used to manage the health of populations." *Science* 366 (6464):447–453.

Önkal, Dilek, Paul Goodwin, Mary Thomson, Sinan Gönül, and Andrew Pollock. 2009. "The relative influence of advice from human experts and statistical methods on forecast adjustments." *Journal of Behavioral Decision Making* 22 (4):390–409.

Park, Joon Sung, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. "A slow algorithm improves users' assessments of the algorithm's accuracy." *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW):1–15.

Pigors, Mark and Bettina Rockenbach. 2016. "Consumer social responsibility." *Management Science* 62 (11):3123–3137.

Prahl, Andrew and Lyn Van Swol. 2017. "Understanding algorithm aversion: When is advice from automation discounted?" *Journal of Forecasting* 36 (6):691–702.

Prahl, Andrew and Lyn M Van Swol. 2021. "Out with the humans, in with the machines?: investigating the behavioral and psychological effects of replacing human advisors with a machine." *Human-Machine Communication* 2:209–234.

Reich, Taly, Alex Kaju, and Sam J Maglio. 2023. "How to overcome algorithm aversion: Learning from mistakes." *Journal of Consumer Psychology* 33 (2):285–302.

Sele, Daniela and Marina Chugunova. 2022. "Putting a Human in the Loop: Increasing Uptake, but Decreasing Accuracy of Automated Decision-Making." *Max Planck Institute for Innovation & Competition Research Paper* (22-20).

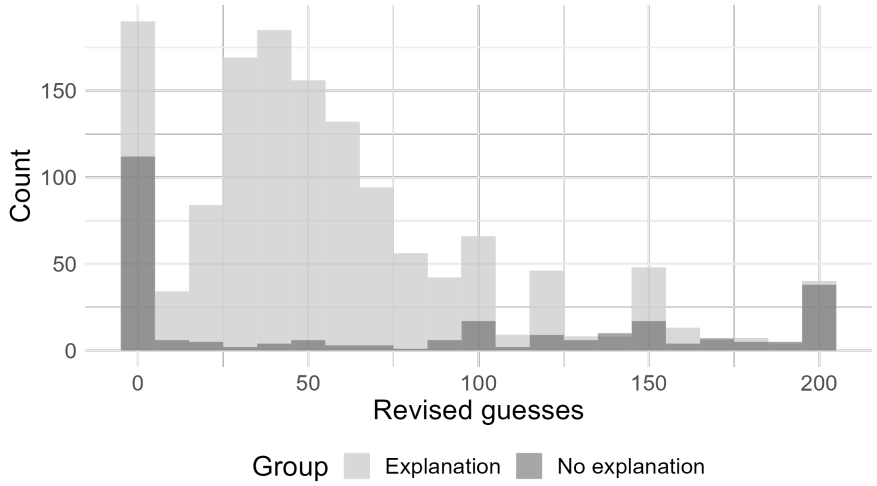


Figure 7: Revised guesses between 0 and 200 pooled for treatments with and without explanation. Raw values (unlogged).

Yin, Ming, Vaughan Wortman, Jennifer Wortman, and Hanna Wallach. 2019. "Understanding the effect of accuracy on trust in machine learning models." In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

Zhang, Yunfeng, Q Vera Liao, and Rachel KE Bellamy. 2020. "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.

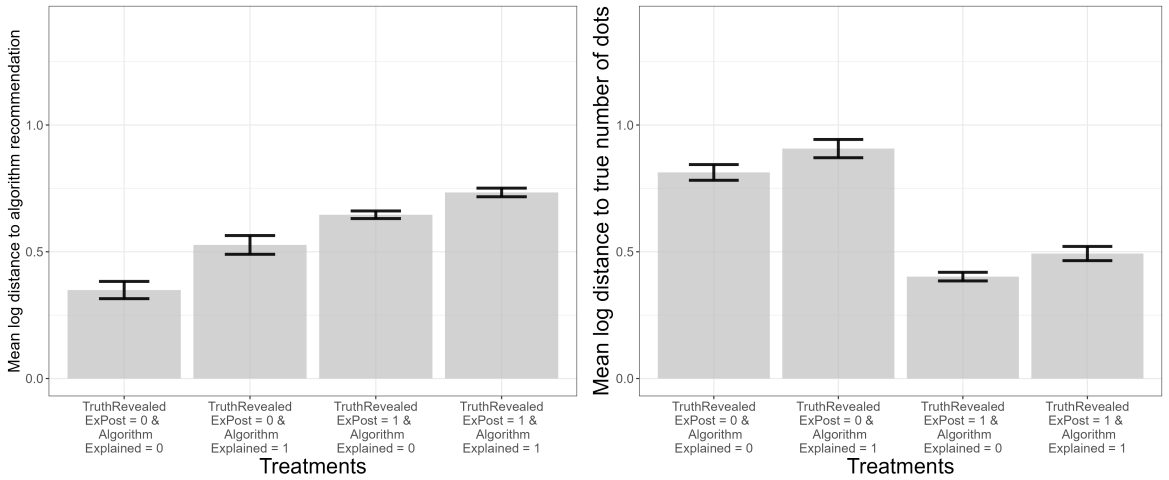
Zhang, Yunhao and Renee Gosline. 2022. "Understanding Algorithm Aversion: When Do People Abandon AI After Seeing It Err?" *Available at SSRN 4299576* .

A Robustness Analysis: Excluding Unreasonably Small Guesses

As discussed in section 2.2, some of our participants state very low guesses, including 0. This is illustrated in figure 7, which plots pooled revised guesses from the two treatments with explanation against the revised guesses from the two treatments without explanation. This figure does *not* include data from the VaryingQuality condition, since that data was generated under alternating algorithm advice quality and can therefore not be compared to the other treatments. To focus on low values, figure 7 shows the revised guess range from 0 to 200.[15]

[15]For the most part of the paper, we analyse the natural logarithm as this is our way to address large outliers. In the following analyses, we are concretely interested in the small values and their interpretation and do not want to address them by taking the logarithm. We, therefore, analyse raw values.

Figure 8: Mean distance to the algorithm and the true number of dots without guesses below 100



(a) Mean distance to the algorithm recommendation by treatment.

(b) Mean distance to the true number of dots by treatment.

Notes: The bar graph in panel (a) illustrates the treatment effects on algorithm adherence (mean distance of the revised guesses to the algorithm recommendation per treatment). The numerical treatment effects on algorithm adherence can be found in table 2 in the appendix. The bar graph in panel (b) illustrates the treatment effects on guessing performance (mean distance of the revised guesses to the true number of dots per treatment). The numerical treatment effects on guessing performance can be found in table 3 in the appendix. The barplots also include the standard errors around the mean. We pre-process the data by taking log values and calculating the mean over all 16 rounds for each individual. For more details see section 2.2.

In figure 7 it is evident that both types of treatments (with and without explanation) exhibit a large number of guesses equal to 0. These 0 guesses are substantially more common in the explanation treatments. Moreover, again only for the explanation treatments, there is a separate smaller hump between 10 and 100 guesses.

How common are these low guesses? As the VaryingQuality condition is not included in this graph, we have 1263 participants (instead of 1565 in the entire sample) who stated revised guesses in 16 rounds. We have therefore $1263 \times 16 = 20,208$ revised guesses. Out of these 1294 values, or 6,4% out of the 20,208, are below the value of 100, where 100 is an arbitrary, but not unreasonable threshold for serious guesses. So the overwhelming majority of revised guesses are in a reasonable range above 100 guesses.

Nonetheless, one can ask if our main results are driven by these low values. In order to answer this question, once can examine figure 8, which shows the same information as our main result figure 3, except that all revised guesses below a value of 100 are excluded from the calculation.

One can see that our main results hold (with one small exception) when revised guesses below 100 are excluded from the analysis:
Result 1a: *Explanation reduces algorithm adherence (weakest effect).*
Result 1b: *Revealing the truth reduces algorithm adherence (medium effect).*
Result 1c: *Combining explanation and revealing truth reduces algorithm adherence (strongest effect).*

Result 2a: *Explanation does not improve performance (and possibly hurts).*
Result 2b: *Feedback improves performance.*
Result 2c: *When providing participants with both explanation and feedback, the net effect is equal to the positive and negative effect of the two individual treatments.*

The one change that appears is the sequence of effects on algorithm adherence. In the initial analysis, revealing the truth has the weakest effect on algorithm adherence. When values below 100 are excluded, revealing the truth has only the second weakest effect.

In the main results, explanation clearly hurts performance. In the robustness check this negative effect is reduced. Still, there is a statistically significant difference as demonstrated by an independent two-sample t-test contrasting the mean outcomes of the baseline and explanation treatments, yielding a p-value of 0.047 ($t = -1.9844$, $df = 602.7$). Nonetheless, given that the robustness check markedly influences the effect size, we adopt a conservative stance and do not claim that explanation degrades performance. The only definitive conclusion is that explanation does not improve performance, and there is a possibility that it may even hurt. The next section explores possible explanations as to why this might be the case.

B Why Does Explanation Hurt Performance?

In the previous section we have seen that very low guesses below 100 account for some part of the negative effect of explanation of performance. Figure 7 illustrates that only the treatments with explanation contain more guesses equal to zero and a hump shaped around 20-100. In this section, we address different hypothesis regarding what causes the 0-guesses and the humps in treatments with explanation.

One possible explanation for the higher number of guesses equal to zero is that some of our participants were displeased or irritated by reading the explanation on how the algorithm works and consequently state a nonsensical guess. Contrary to this hypothesis, we observe that the number of revised guesses equal to zero decreases over the course of the experiment. This can be seen in figure 9, which shows a histogram of revised guesses in the Explanation treatment for the first four rounds.

The trend of fewer zeros over the course of the rounds continues until the end of the

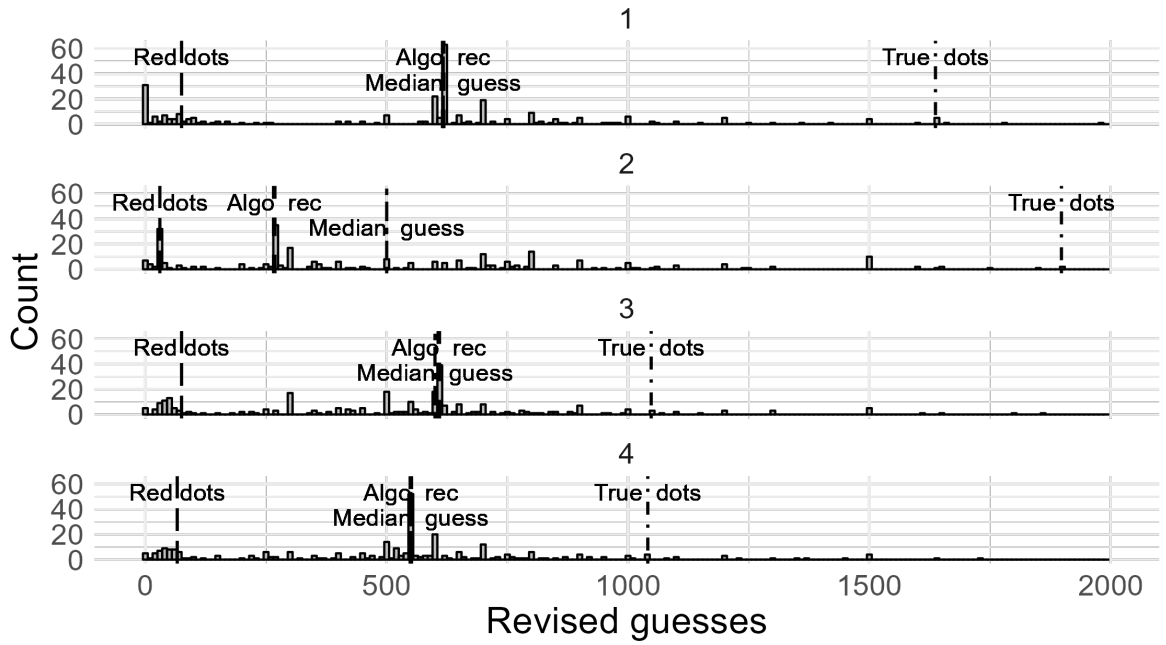


Figure 9: Histogram of revised guesses in the EXPLANATION treatment for the first four rounds.

Notes: "True dots" indicates the true number of dots, "Median guess" indicates the median revised guess, "Algo rec" indicates the algorithm recommendation and "Red dots" indicates the number of dots within the red squares of the explanation which participants in this treatment were exposed to. Raw values (unlogged).

experiment. Therefore, another possible explanation which is more in line with this observation is that some participants did not want to play this task. Note that our participants did not know that each round would contain the dot guessing task. Possibly, some were trying to move quickly on to a task they might find easier and stated a guess that required a low cognitive effort (namely a guess equal to zero).

We now turn to a discussion on what might have caused the hump shaped guesses around values of 20-100. One hypothesis is that participants who saw the explanation of how the algorithm worked interpreted this as an instruction. Instead of a critical assessment of the algorithms quality, some participants might have concluded that the experimenter want them to use the same approach as the algorithm to come up with their guess. Yet, if this were the case, we would expect to see more guesses around the algorithm recommendation - which is not the case - rather than in the range of 20-100.

A related hypothesis is that some participants misinterpreted our instructions and mistakenly thought they should only guess the number of dots *within* the red boxes which they saw as the visual part of the explanation of how the algorithm works. But if this were the case, one would expect the hump shape to be centered above the number of dots within the red boxes. Figure 9 displays the guesses and the number of dots within the red boxes as a dotted dark red line. If this hypothesis was true, we would expect the number of dots in the red boxes to be in the center of the distribution humps. However, what we observe is that, the hump shape is in every round to the left of the dotted line. If this were the driving behavior, participants must have systematically counted fewer dots than there actually were in the red boxes.

Another possibility is that some participants misinterpreted our instructions such that they should guess the *average* number of dots within *one* red box. This would be in line with the observed data, but seems unlikely given that this would require them to read the instructions carefully enough to understand the algorithm, but at the same time to severely misinterpret the goal of the task, while making several incorrect assumptions.

Finally, we discuss the point that there is a statistically significant negative effect of explanation on performance even when low values are excluded. Figure 10 shows the distributions of differences between the revised guesses and the true number of dots (for non-log values), again pooled for the two treatments with and without explanation. From this, it becomes apparent that there is no obvious effect on the distribution that could drive the result. An eye-balling inspection could suggest that the distribution in the explanation treatments is more spread out. This could hint at a heterogeneous effect of explanation: some participants might be adversely effected (due to boredom or other negative unintended externalities), while other participants correctly deduce that the algorithm advice is too low and over-correct for its bias. Yet, a Levene test for homogeneity of variances cannot reject the null of equal variances (F-value of 2.77 and a p-value of 0.095).

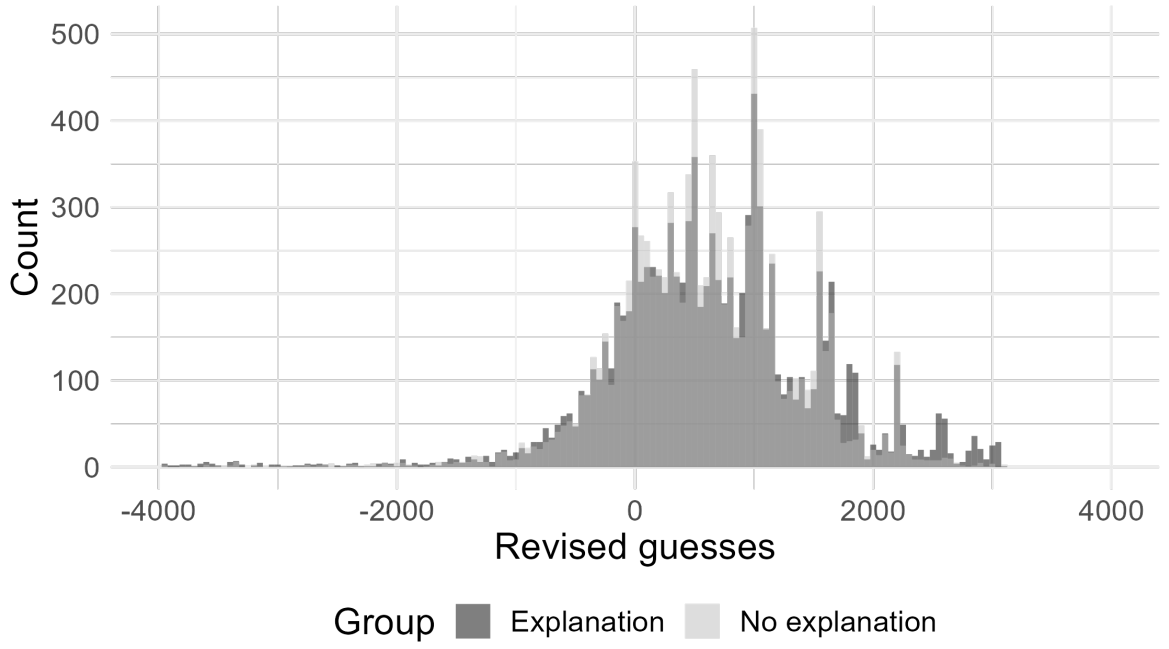


Figure 10: Distribution of distances between revised guesses and true number of dots (non log values).

Overall, we cannot provide a definite answer to the question of why explanation hurts performance. Pinpointing the underlying mechanisms requires a new experimental design and must remain an open question for future research. In this section, we have provided a discussion of the potential underlying mechanisms based on our experimental set-up and the available data.

C Would Participants Have Benefited from Following the Algorithmic Advice?

One question to explore is whether participants would have enhanced their performance by incorporating algorithmic advice into their revised guesses. We address this question for each round individually. The following figure illustrates the distance to the true answer for each round in every treatment. The squares denote the gap between the correct answer and the algorithmic recommendation, while the dots represent the average distance between the true values and the initial guess. Lower values indicate better performance. On average, participants would have gained from assigning a positive weight to the algorithmic advice in most rounds in the baseline treatment (algorithmic prediction superior in 14/16 rounds) and the explanation treatment (algorithmic prediction superior in 13/16 rounds). Conversely, this is not observed in the feedback treatment (participant guess superior in 14/16 rounds)

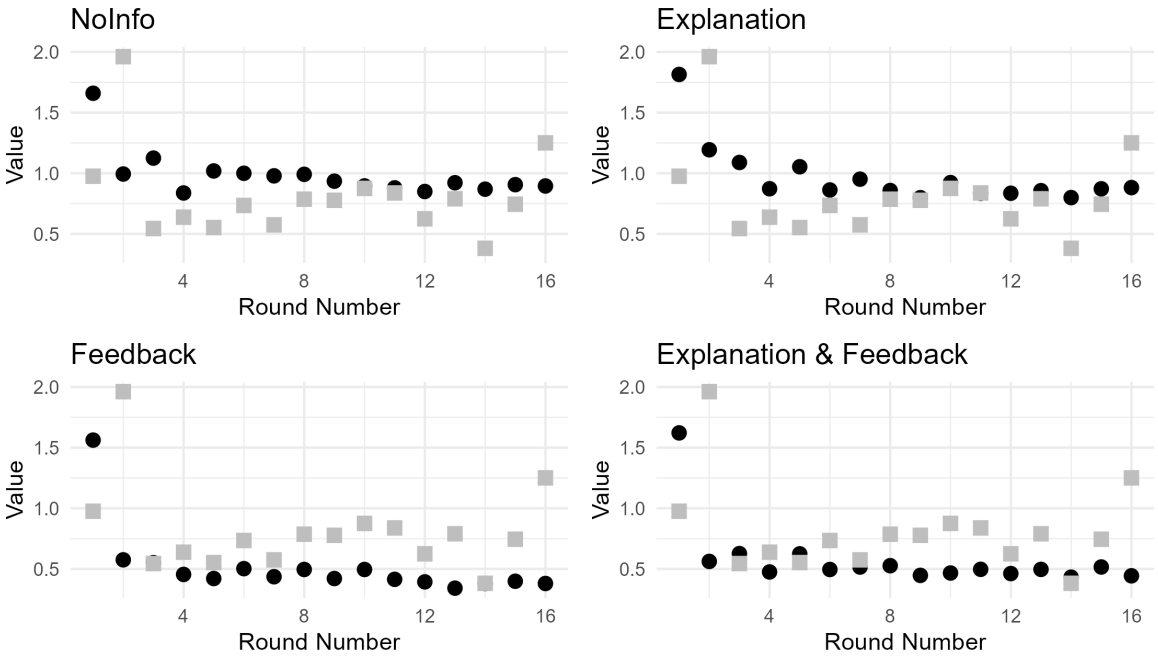


Figure 11: Performance participants vs. algorithm

and the combined treatment (participant guess superior in 12/16 rounds).

D Extreme Behaviors: Complete Adherence to Advice and Complete Disregard of Advice

The depicted figure illustrates the percentage of individuals who fully adhere to the algorithm (i.e., the revised guess matches the algorithmic prediction) and those who disregard the algorithmic advice entirely (i.e., the revised guess aligns with the initial guess). The figure demonstrates a consistent reduction in trust in algorithmic advice across all treatments. Further, treatments including feedback seem to elevate confidence in one's initial guess, likely due to the availability of an orientation point. In the explanation treatment, there is no apparent reason to put greater trust in one's initial guess compared to the baseline treatment. Consequently, in the explanation treatment, subjects exhibit diminished trust in the algorithm, but this doesn't translate into increased confidence in their own assessment.

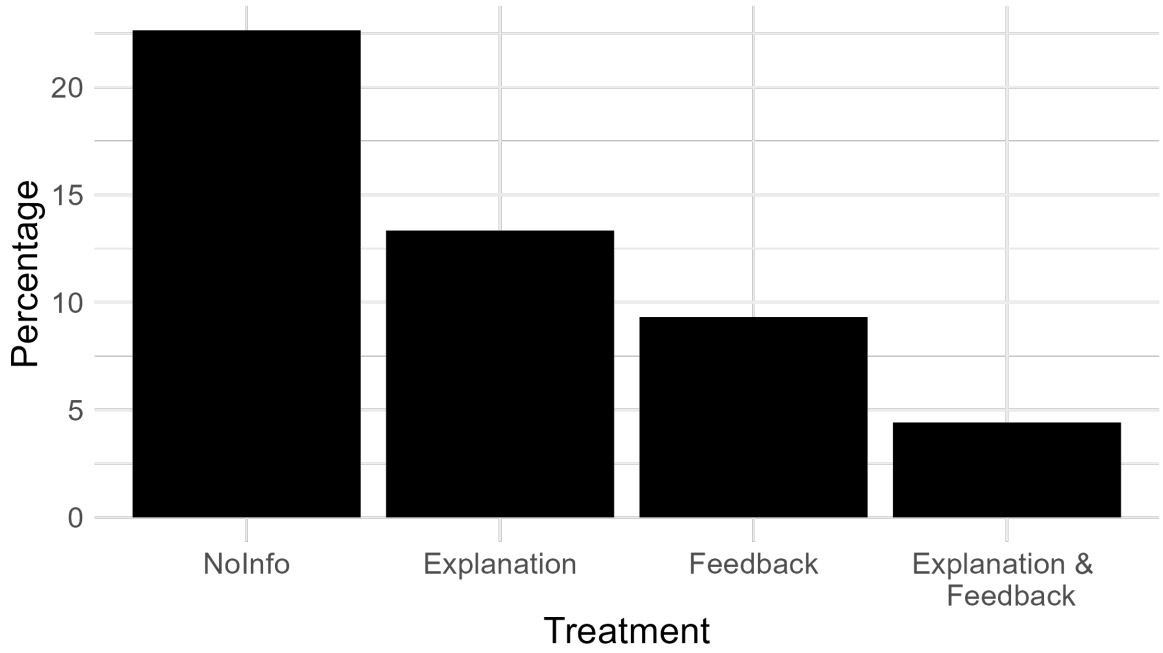


Figure 12: Percentage of complete adherence to advice

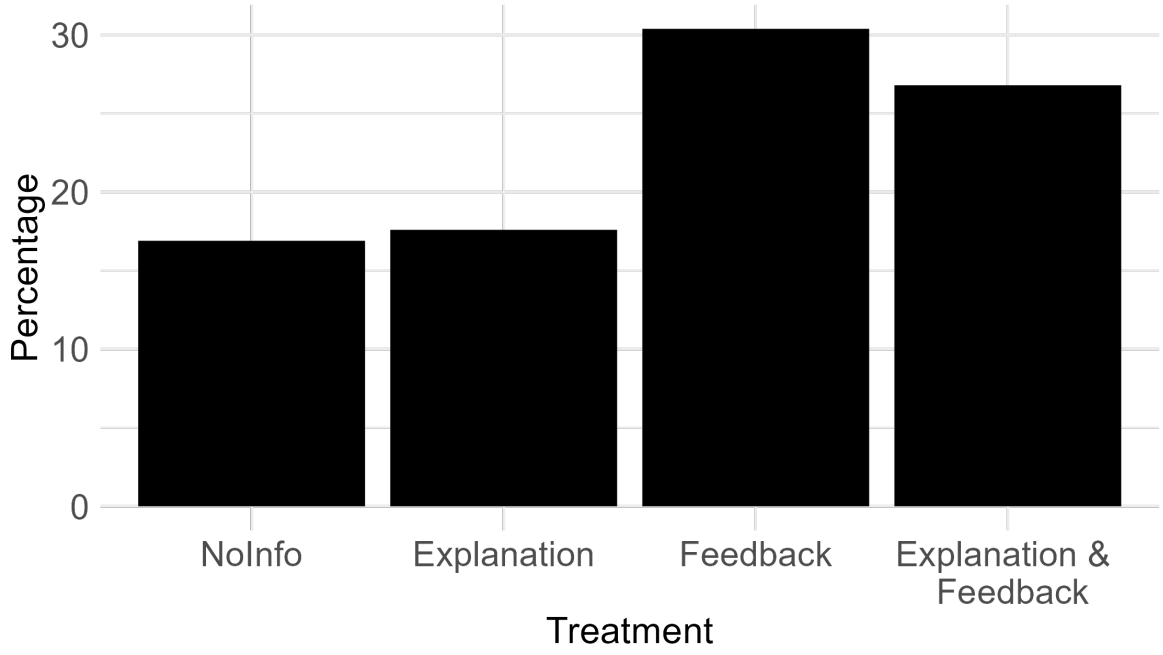


Figure 13: Percentage of complete disregard of advice

E Additional Analysis Baseline Treatment: Algorithmic Advice as a Credence Good

This section offers an additional interpretation of the baseline treatment, arguing that algorithmic advice is often perceived to be a credence good. A common view among policymakers and in the academic literature implicitly assumes that human decision-makers can accurately assess the advice quality after observing the algorithm's recommendation, for instance, by comparing it to their own judgment (e.g. in the Artificial Intelligence Act as proposed by the European Commission, 2021). From an economic point of view, this suggests that algorithm advice is perceived as an experience good, i.e., consumers can accurately assess the quality after consumption of the good. We challenge this assumption and argue that algorithms are often perceived to be *credence goods*. Even after "consumption" of the good – repeated interaction with the algorithm – humans cannot correctly assess its advice quality.

We provide experimental evidence from a reasonable setting that many people cannot correctly assess the quality of algorithmic advice even after "consuming" it. It follows that they perceive algorithmic advice as a credence good.[16]

Our work is hence related to the literature on credence goods that can be categorized into two different strands (Balafoutas and Kerschbamer, 2020). "Classical credence goods" involve asymmetric information between the expert seller and their customer regarding the fit between the characteristics of the products and the customer needs, prominent examples being healthcare or repair services. Our case aligns with the second strand of literature: "label credence goods". Such goods have unobservable attributes that remain undetected after consumption, common examples being organically produced food and fairly traded products (Baksi and Bose, 2007; Harbaugh, Maxwell, and Roussillon, 2011; Pigors and Rockenbach, 2016; Etilé and Teyssier, 2016).

We ask the question of whether people can assess the quality of algorithmic advice (and if they act accordingly) after "consuming" this advice. To answer this question, one can inspect figure 14, which shows the densities of the initial and revised guesses of the first four rounds in treatment NoInfo.

In the first round, the initial guess density is flat: Participants vary vastly in their initial guess. After observing the algorithm recommendation, participants state their revised guess, resulting in the revised guess density. Participants strongly react to the algorithmic advice and many subjects follow the advice closely. This can be directly inferred from the revised density: It is centered above the algorithm recommendation and its variance is greatly re-

[16] For our work, it is important that algorithms have unobservable attributes that remain undetected after consumption. This type of credence goods is often referred to as "label credence goods". The analysis of "classical credence goods", on the other hand, focuses on the asymmetric information between an expert seller and the customer. In this article, we refer to "label credence goods" whenever we employ the term "credence goods".

duced. The second round illustrates that the initial guesses in round two are still influenced by the algorithmic advice from round one: The density of the initial distribution is centered above the previous round's algorithm recommendation. The algorithm recommendation in one round serves as an orientation point for the next initial guess. The revised guess density in round two peaks again above the algorithm recommendation. In rounds three and four one again observes the two effects (1) the revised guesses move closer to the algorithm and (2) their variance is reduced (both compared to the initial guess density in the same round). In fact, this pattern holds for all subsequent 12 rounds (see figures 15 to 18).

Table 1 quantifies these differences and shows the average of the individual log distances to the algorithm and the standard deviation for the initial and revised guesses. It also exhibits p-values from a t-test comparing the initial and revised distance to the algorithm recommendation and Levene's test for homogeneity of variances of the initial and revised densities.

The average distance to the algorithm recommendation is smaller for the revised guesses than for the initial guesses in all 16 rounds (this difference is always significant except for one round). In other words, revised guesses move closer to the algorithm. The standard deviations of the revised guess densities are smaller for the revised guesses in a majority of cases.[17]

If the algorithmic advice exhibited traits of an experience good, we would expect our subjects to learn to optimally incorporate this advice into their decision-making and adjust how strongly they adhere to the algorithm. Note that in each round, participants have the possibility to compare the algorithm recommendation with their own guess, which was elicited in the first stage. Given their own guess as a reference point, participants could realize that the algorithm recommendation is consistently too low. Over time (i.e. over rounds), if participants had this realization repeatedly, and assuming they would also optimally react based on this insight, we would see a shift in the revised guess density away from the algorithm (and potentially closer to the true number of dots). As can be seen from figure 14 and table 1 we see no evidence for this. In figure 14 there is no increase in probability mass in the region above the algorithm recommendation for the revised guesses over rounds. In table 1 average distance of the revised guesses to the algorithm is in every round smaller than the distance for the initial guesses. In other words, the average revised guess always move closer to the algorithm. So the participants never learn to move further away from the algorithm recommendation (or ignore it).

Table 1: Distances to algorithmic recommendation per round: NoInfo

	$\lvert log(algo) - log(guess_i)\rvert$	$\lvert log(algo) - log(guess_r)\rvert$	p-values

[17] Note that many the p-values of these differences are not significant.

Round	mean	sd	mean	sd	t-test	levene
1	1.29	1.78	0.59	1.39	0.00	0.00
2	1.20	0.73	0.88	1.13	0.00	0.00
3	0.78	0.93	0.42	0.74	0.00	0.13
4	0.51	0.67	0.45	1.02	0.28	0.33
5	0.66	0.89	0.35	0.74	0.00	0.77
6	0.67	1.16	0.42	1.00	0.00	0.22
7	0.60	1.06	0.33	0.90	0.00	0.06
8	0.61	0.91	0.37	0.81	0.00	0.21
9	0.58	0.96	0.42	1.03	0.00	0.66
10	0.62	0.99	0.40	0.98	0.00	0.60
11	0.46	0.74	0.31	0.90	0.00	0.69
12	0.47	0.70	0.29	0.68	0.00	0.68
13	0.53	1.08	0.33	0.89	0.00	0.19
14	0.64	0.74	0.42	0.95	0.00	0.82
15	0.52	1.07	0.34	1.01	0.00	0.31
16	0.68	0.89	0.49	0.98	0.00	0.24

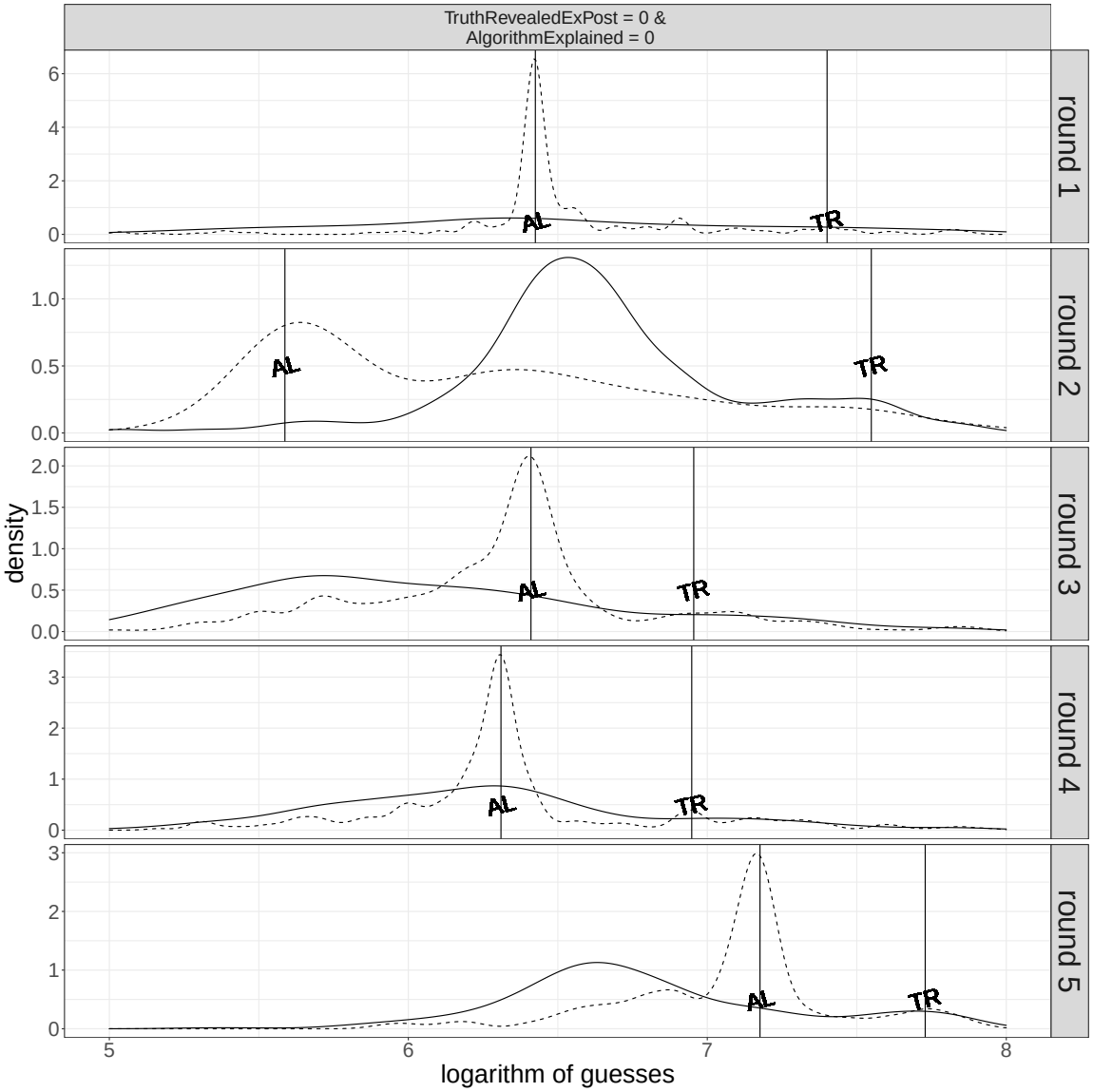
Notes: Table contains the mean and standard deviation of the distance between the guesses and the algorithmic recommendation for every round. This distance is included with respect to the initial and the revised guesses. All values are logs. Table also contains the p-values for the t-test (difference between means) and the Levene-test (differences between variances) to test the difference between the values for initial and revised guesses for every round. Values refer to NoInfo.

Especially participants whose initial guesses fall within this range between the algorithm and the true value should not move closer to the algorithm to maximize their payoff. Yet, our data show that participants move closer to the algorithmic advice. Our results therefore rather indicate that the algorithmic advice exhibits traits of a credence good: Even after "consuming" the advice repeatedly, our participants appear not to be able to assess the quality of the advising algorithm.
Consequently, we document the following result:

Result: *The human decision makers perceive the algorithmic advice as a credence good.*

Participants in the baseline treatment NoInfo repeatedly see biased advice in a task that does not require expert knowledge. Akin to many real-life decision situations, in this treatment, they therefore can compare their own prediction (note that we prime participants by specifically eliciting their initial guesses before seeing the algorithm) with the algorithm

Figure 14: Densities of initial and revised guesses by treatment for the first four rounds



Notes: Initial (black line) and revised (dotted line) guess densities for NoInfo for the first five rounds. Algorithm recommendation (AR) is the leftmost vertical line and the true number of dots (TR) is indicated by the rightmost vertical line. The figure only shows the range of log(guess) from 5 to 8 and therefore does not display the tails of the distributions. The range of the axis showing the density differs between rounds.

prediction. In principle, this could allow any participant to realize that the algorithm is biased and produces dot predictions that are too low. The question is how many participants realize this and if they can correctly adjust for this bias.

From the written feedback that participants could state at the end of the experiment, we know that some participants in the baseline treatment indeed recognize that the algorithm is downward biased (e.g. "I thought the algorithm consistently underestimated the number of dots", "I did not trust the algorithm. It seemed to be generating numbers that were too low."), yet others fail to assess the existence, magnitude, or direction of the bias (e.g. "I quickly began to depend on the algorithm and as the study progressed, got close to guessing what the algorithm predicted", "I figured that it was overestimating"). Ultimately, the question about the nature of the algorithm as a good is an empirical one: How do the majority of participants assess the algorithm? As described in the previous section, most participants follow the algorithm closely and never realize that they could improve their payoff if they move away from a biased recommendation.

Some important caveats are in order. We do not claim that all algorithms are always perceived as credence goods. We merely point out that algorithms can be perceived as credence goods in many applications. Clearly, one important determinant is how the human and algorithm abilities compare. Our results and implications could be considered in scenarios where neither human nor algorithm is obviously better, but where there is ambiguity about performance comparison. Moreover, the nature of our task is rather mathematical and objective, and behavior may vary for more subjective tasks (Castelo, Bos, and Lehmann, 2019). Further, our participants may assume that the algorithm performs better than humans because it would not be employed otherwise. In sum, one might consider subjects' priors that the algorithm is more capable to be justified. In addition, one might criticize that learning is generally not possible without feedback and explanation,[18]

In this section, we have stated the assumption that individual humans can successfully provide oversight for algorithms is flawed and discuss the implications of this. A natural adjustment is, therefore, to put less emphasis on *individual* oversight and instead shift the focus to *collective* oversight. For example, organizations could audit algorithmic advice systems. This could entail checking possible training data, systematically challenging the algorithm or controlled human-subject field experiments before deployment.[19]

Our results have implications for both policymakers and managers. We show that there are situations in which humans are not able to accurately assess an advising algorithm's quality. In the context of regulation, this casts doubt on the effectiveness of individual human decision-makers to recognize biased algorithms and to correct this bias to prevent

[18]In response, we argue that what makes learning possible is the repeated engagement with the task and the fact that subjects provide initial as well as revised guesses.

[19]This idea is similar to Green (2022), which suggests "institutional oversight".

harm. In the context of management, it means that organizations generally can neither rely on individual decision-makers to optimize decisions nor can they rely on feedback from their decision-makers about the quality of an algorithm as a product.

F Tables and Figures

Table 2: Treatment effect on algorithm adherence: Log-distance to algorithm recommendation: Overview

Treatment	n	min	mean	max	std. err.
TruthRevealedExPost = 0 & AlgorithmExplained = 0	324	0.001	0.426	9.514	0.035
TruthRevealedExPost = 0 & AlgorithmExplained = 1	314	0.001	0.842	5.188	0.052
TruthRevealedExPost = 1 & AlgorithmExplained = 0	312	0.001	0.711	1.989	0.017
TruthRevealedExPost = 1 & AlgorithmExplained = 1	313	0.001	0.969	4.271	0.034

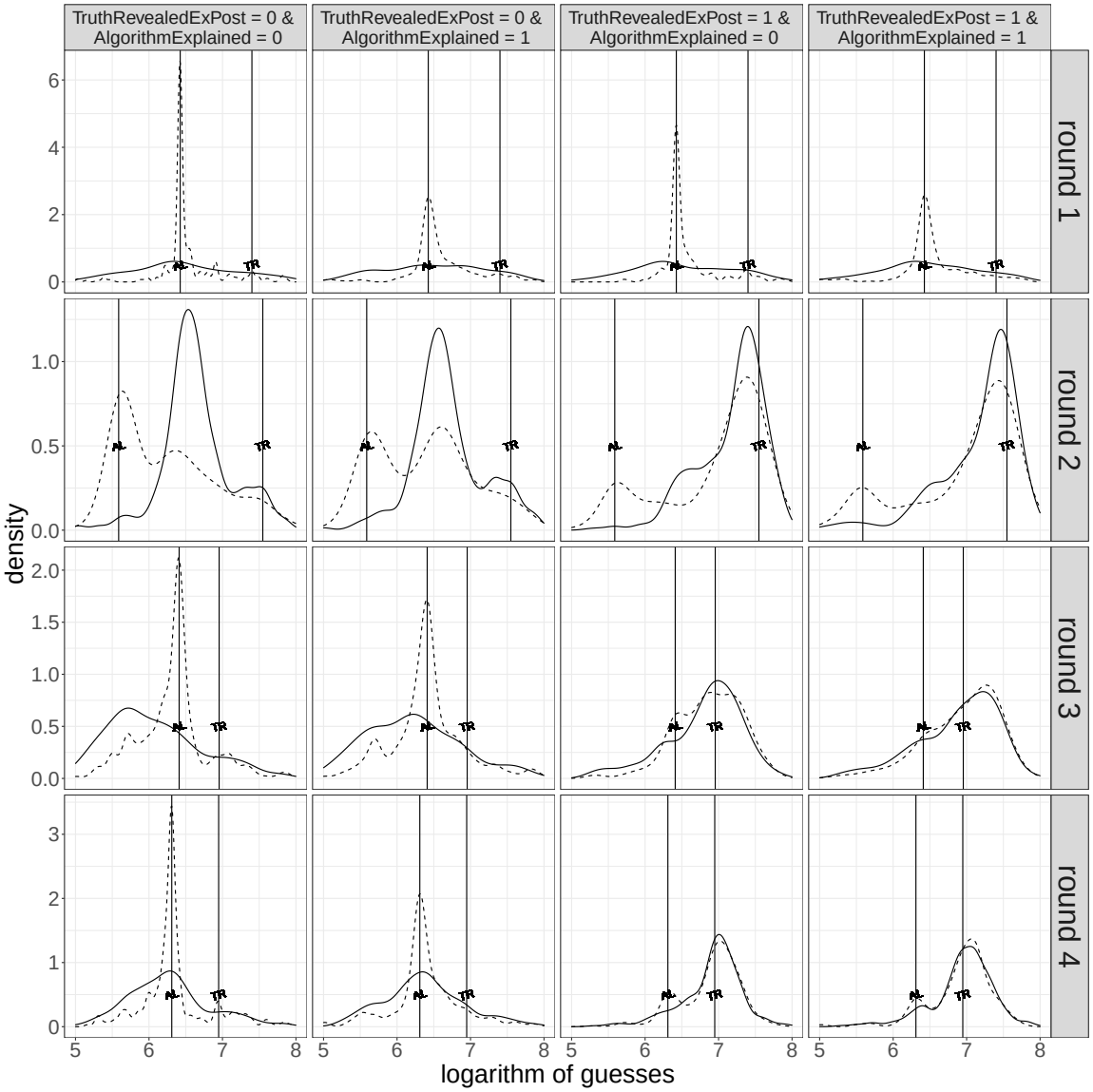
Notes: The values in this table refer to the barplot in panel (a) of figure 3a. "Std. err." refers to the standard error of the mean.

Table 3: Treatment effect on guessing performance: Log-distance to true number of dots: Overview

Treatment	n	min	mean	max	std. err.
TruthRevealedExPost = 0 & AlgorithmExplained = 0	324	0.001	0.897	9.122	0.033
TruthRevealedExPost = 0 & AlgorithmExplained = 1	314	0.001	1.258	4.678	0.057
TruthRevealedExPost = 1 & AlgorithmExplained = 0	312	0.001	0.479	2.025	0.022
TruthRevealedExPost = 1 & AlgorithmExplained = 1	313	0.001	0.792	5.087	0.052

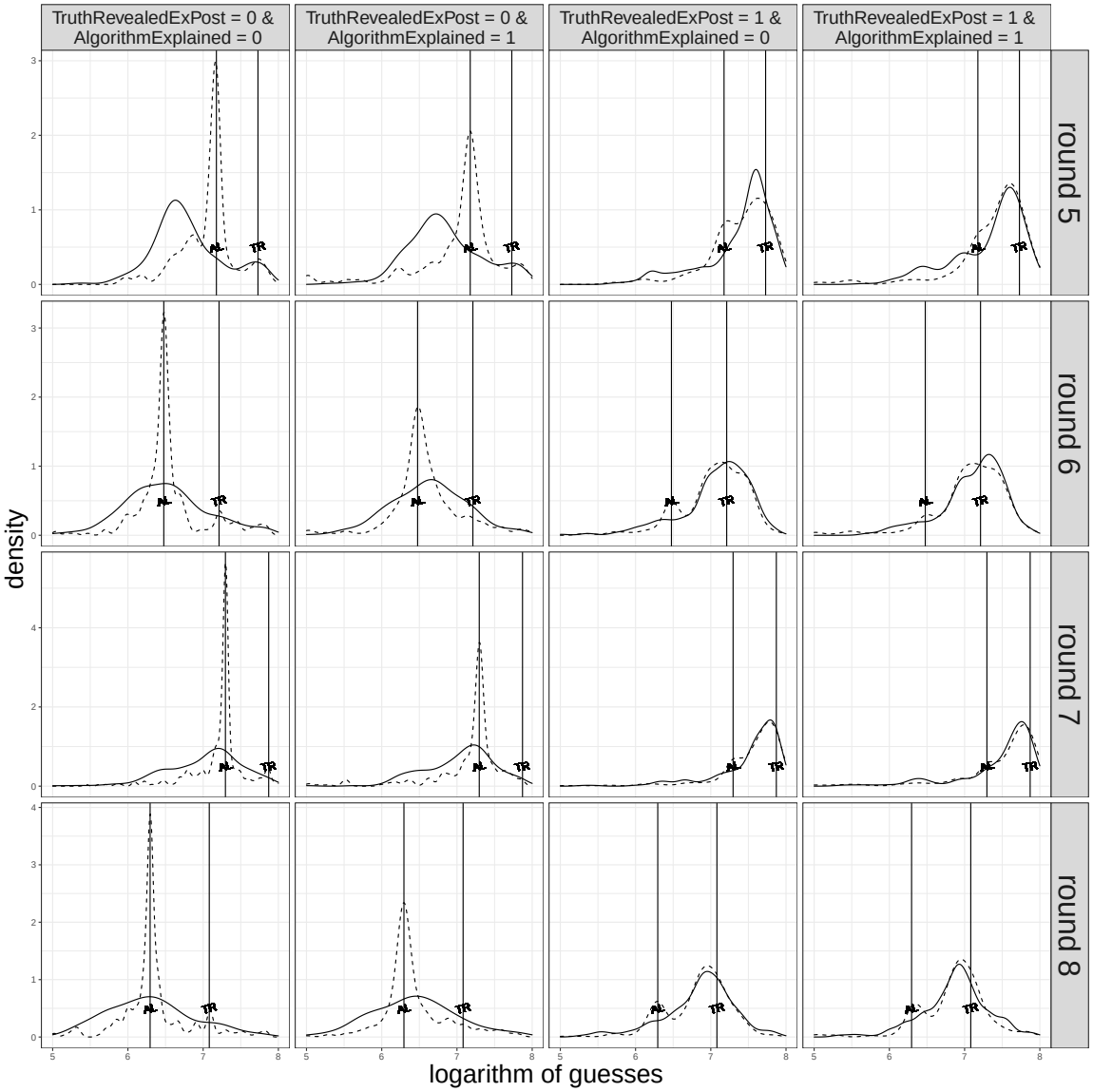
Notes: The values in this table refer to the barplot in panel (b) in figure 3. "Std. err." refers to the standard error of the mean.

Figure 15: Distribution of initial and revised guesses by treatment for rounds 1 to 4



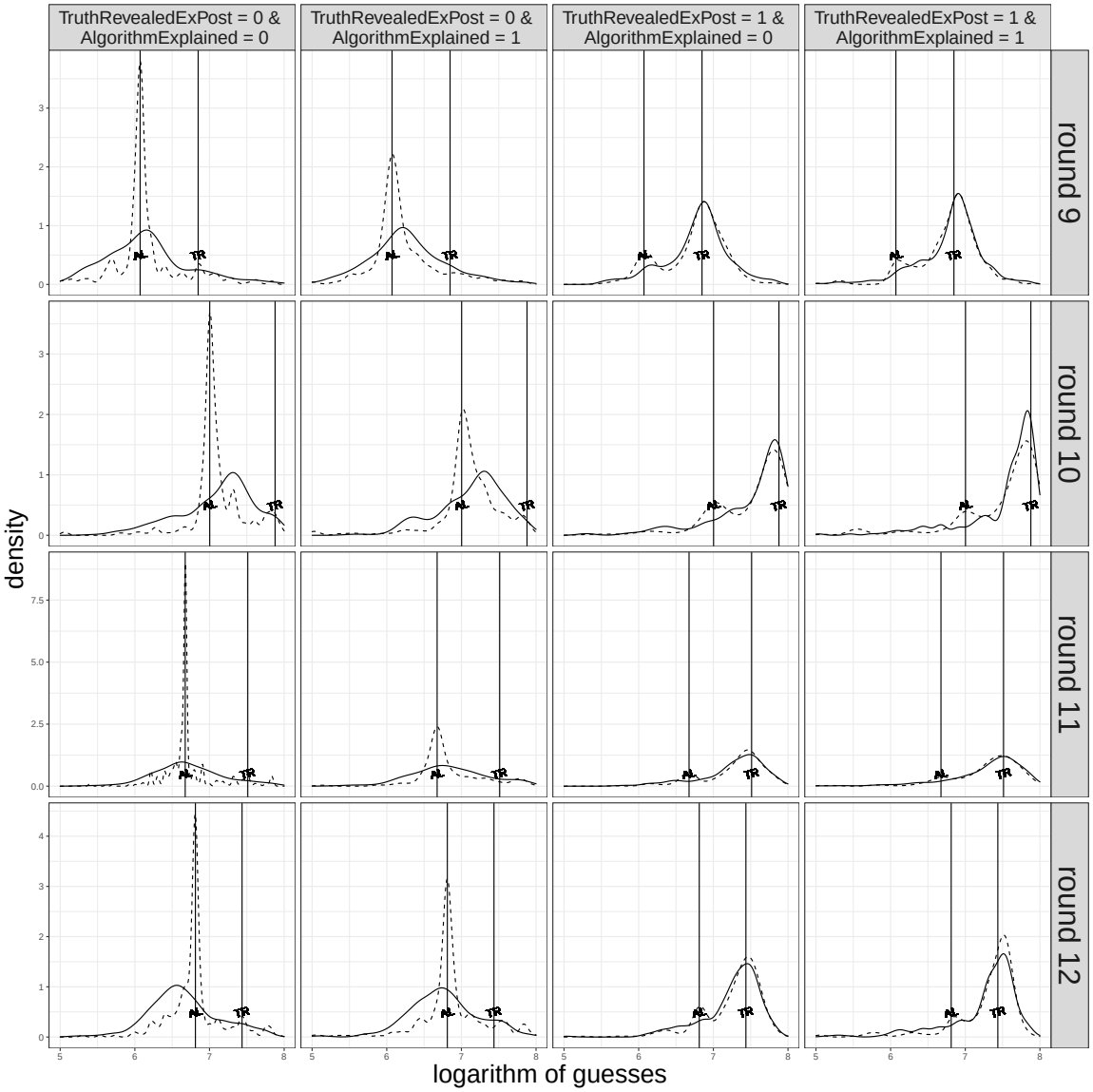
Notes: Initial and revised guess densities for round 1 to 4.

Figure 16: Distribution of initial and revised guesses by treatment for rounds 5 to 8



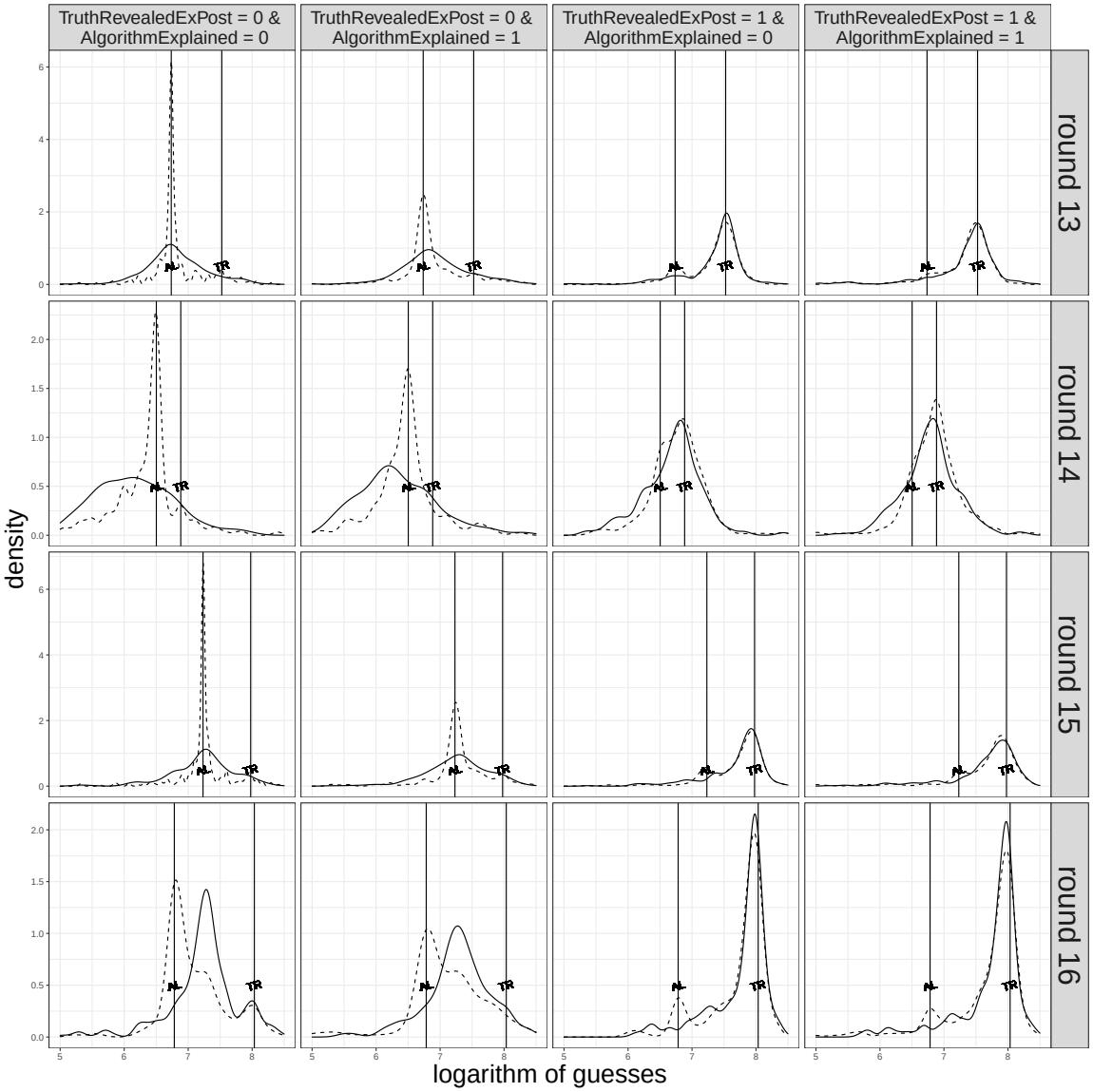
Notes: Initial and revised guess densities for round 5 to 8.

Figure 17: Distribution of initial and revised guesses by treatment for rounds 9 to 12



Notes: Initial and revised guess densities for round 9 to 12.

Figure 18: Distribution of initial and revised guesses by treatment for rounds 13 to 16



Notes: Initial and revised guess densities for round 13 to 16.

Table 4: Distances to algorithmic recommendation per round: EXPLANATION

	$\lvert log(algo) - log(guess_i)\rvert$		$\lvert log(algo) - log(guess_r)\rvert$		p-values	
Round	mean	sd	mean	sd	t-test	levene
1	1.45	1.95	1.26	1.89	0.14	0.58
2	1.44	1.19	1.24	1.16	0.00	0.04
3	0.82	1.12	0.84	1.17	0.77	0.04
4	0.65	0.93	0.79	1.14	0.03	0.00
5	0.75	0.94	0.91	1.46	0.04	0.00
6	0.66	0.99	0.80	1.25	0.06	0.00
7	0.63	0.98	0.76	1.30	0.05	0.00
8	0.62	0.77	0.70	1.02	0.19	0.00
9	0.60	0.69	0.78	1.20	0.00	0.00
10	0.70	1.02	0.79	1.12	0.20	0.00
11	0.58	0.68	0.75	1.13	0.00	0.00
12	0.56	0.82	0.75	1.21	0.00	0.00
13	0.60	0.95	0.73	1.12	0.04	0.00
14	0.63	0.66	0.69	1.02	0.34	0.00
15	0.59	0.87	0.82	1.33	0.00	0.00
16	0.81	0.81	0.86	1.01	0.32	0.00

Notes: The structure of the table is the same as in table 1, but the values refer to EXPLANATION.

Table 5: Distances to algorithmic recommendation per round: FEEDBACK

	$\lvert log(algo) - log(guess_i)\rvert$		$\lvert log(algo) - log(guess_r)\rvert$		p-values	
Round	mean	sd	mean	sd	t-test	levene
1	1.25	1.76	0.44	1.05	0.00	0.00
2	1.71	0.72	1.52	0.94	0.00	0.00
3	0.71	0.86	0.68	0.87	0.58	0.50
4	0.80	0.87	0.72	0.75	0.14	0.73
5	0.52	0.35	0.59	0.99	0.23	0.01
6	0.82	0.89	0.77	0.82	0.20	0.94
7	0.56	0.66	0.51	0.64	0.36	0.99
8	0.76	0.74	0.63	0.52	0.00	0.15
9	0.85	0.75	0.71	0.42	0.00	0.13

10	0.86	0.77	0.75	0.71	0.06	0.93
11	0.75	0.60	0.69	0.51	0.11	0.74
12	0.64	0.70	0.61	0.69	0.47	0.86
13	0.76	0.58	0.70	0.60	0.06	0.58
14	0.42	0.37	0.45	0.69	0.39	0.11
15	0.65	0.51	0.59	0.50	0.00	0.70
16	1.07	0.54	1.01	0.56	0.12	0.61

Notes: The structure of the table is the same as in table 1, but the values refer to FEEDBACK.

Table 6: Distances to algorithmic recommendation per round: EXPLANATION&FEEDBACK

	$\lvert log(algo) - log(guess_i) \rvert$		$\lvert log(algo) - log(guess_r) \rvert$		p-values	
Round	mean	sd	mean	sd	t-test	levene
1	1.23	1.86	1.25	1.89	0.90	0.14
2	1.74	0.71	1.64	0.96	0.09	0.00
3	0.81	0.95	0.92	0.99	0.06	0.19
4	0.84	0.82	0.91	0.90	0.19	0.19
5	0.67	1.02	0.89	1.35	0.00	0.01
6	0.85	0.84	0.95	0.91	0.07	0.12
7	0.64	0.85	0.74	0.91	0.15	0.12
8	0.79	0.87	0.91	1.05	0.06	0.08
9	0.89	0.79	0.96	0.85	0.15	0.16
10	0.89	0.71	1.07	1.26	0.01	0.00
11	0.86	0.79	0.93	0.92	0.16	0.10
12	0.73	0.82	0.85	0.89	0.04	0.16
13	0.88	0.90	0.91	0.88	0.61	0.48
14	0.51	0.67	0.59	0.70	0.08	0.11
15	0.75	0.80	0.84	0.90	0.12	0.05
16	1.10	0.63	1.14	0.68	0.41	0.04

Notes: The structure of the table is the same as in table 1, but the values refer to EXPLANATION&FEEDBACK.