RATIONALITY & COMPETITION
CRC TRR 190

The Good, the Bad, and the Bayesian: Self-Confidence and Algorithmic Advice

Rongrong Fu (HU Berlin)
Fabian Gerstmeier (TU Berlin)
Michel Tolksdorf (TU Berlin)

Discussion Paper No. 589

October 22, 2026

The Good, the Bad, and the Bayesian: Self-Confidence and Algorithmic Advice

Rongrong Fu[*†], Fabian Gerstmeier [*‡], Michel Tolksdorf[*§]

October 22, 2026

Abstract

Understanding how decision makers incorporate external information is increasingly important as algorithmic predictions become a common source of advice. The literature on advice-taking suggests that people discount others' advice, especially when they feel competent. We experimentally study how self-confidence affects advice utilization and how algorithmic advice shapes confidence in final judgments. In our online experiment, participants make age predictions and assess the prediction accuracy before and after receiving algorithmic advice. We exogenously vary self-confidence through (perceived) task difficulty and vary the (observable) quality of the algorithmic advisor. While participants generally put more weight on high-quality ("good") advice, we find, contrary to the existing advice-taking literature, that they are less influenced by advice when they perceive the task as difficult. By contrasting age and accuracy predictions, we show that advice can influence beliefs without substantially changing point predictions: people who feel proficient at a task barely adjust their predictions toward high-quality advice but become more confident in them. When people feel less proficient, they are more reluctant to adjust their predictions toward low-quality ("bad") advice. Our design shows that prediction and confidence updating respond differently to perceived task difficulty and advice quality, implying that measures based solely on changes in point predictions can misstate the influence of advice. Comparing observed updating with a Bayesian benchmark further reveals systematic underutilization of advice, driven in particular by insufficient confidence updating following high-quality advice in tasks perceived as easy.

Keywords: Algorithmic Advice, Algorithm Appreciation, Confidence, Online Experiment

JEL Codes: C91, D83, D91

[*]Berlin School of Economics.

[†]Humboldt University of Berlin, Spandauerstraße 1, 10099 Berlin, Germany, rongrong.fu@hu-berlin.de. https://www.wiwi.hu-berlin.de/de/administration/fakultaetsverwaltung/Staff/1694881. Corresponding author.

[‡]Technische Universität Berlin, H91, Str. des 17. Juni 135, 10623 Berlin, Germany, gerstmeier@tu-berlin.de. https://berlinschoolofeconomics.de/people/members/fabian-gerstmeier.

[§]Technische Universität Berlin, H91, Str. des 17. Juni 135, 10623 Berlin, Germany, michel.tolksdorf@tu-berlin.de. https://www.tu.berlin/mikro/team/wissenschaftliche-mitarbeiterinnen/dr-michel-tolksdorf.

Financial support by Deutsche Forschungsgemeinschaft through CRC TRR 190 (project number 280092119) is gratefully acknowledged. We thank Dirk Engelmann, Radosveta Ivanova-Stenzel, Jonas Radbruch, Sebastian Schweighofer-Kodritsch, Georg Weizsäcker, participants of the ESA European Meeting 2026 in Barcelona, as well as the participants of the Seminar series by the Berlin Behavioral Economics group for useful comments and advice. This study was pre-registered at AsPredicted.org (https://aspredicted.org/8cj3gd.pdf) and received ethics approval by the Technical University Berlin Faculty of Economics and Management (VII) Ethics Committee (project number 20251016).

1 Introduction

Economic decisions often require combining one's own judgment with information from other sources. A manager combines their assessment of demand with a sales forecast, an investor combines their valuation with an analyst's estimate, and a physician combines clinical judgment with a diagnostic prediction. In each case, the decision maker faces two related questions: How much should the additional information change their prediction, and how certain should they be about their final prediction? The answer to these questions depends in part on the quality of the advice and on whether the quality is correctly assessed. Ultimately, however, advice can improve judgments only if decision makers integrate the information it provides appropriately.

In their seminal paper, Yaniv and Kleinberger (2000) find that individuals discount other people's advice relative to their own judgments, coining the term *egocentric discounting*. Subsequent work has documented this tendency across a wide range of judgment and decision-making settings (Bailey et al., 2023).

The rapid adoption of algorithmic advisors in recent years introduces a new dimension to the advice-taking literature, as highly accurate advice is increasingly available almost instantaneously in both everyday and professional decisions and is already used by many professionals (Bick et al., 2026). This can, in principle, generate efficiency gains, given evidence that algorithms can outperform human professionals, particularly in prediction tasks (Liu et al., 2019; Ardila et al., 2019; McKinney et al., 2020; Mullainathan and Obermeyer, 2022; Liu et al., 2023). Additionally, once trained, algorithms can provide advice at relatively low marginal cost compared to human advisors, making repeated advice seeking feasible even in familiar or recurring situations. Moreover, their performance can be disclosed with greater precision than that of most other advisors, allowing advisees to optimally update their prior beliefs based on the relative precision of their own judgment and the advice. Accordingly, algorithmic advice can generate substantial efficiency gains, provided that it is utilized appropriately.

However, previous work has documented systematic deviations from "appropriate" reliance on algorithmic advice. Dietvorst et al. (2015) coin the term *algorithm aversion* to describe a reluctance to rely on an algorithm even when it outperforms one's own judgment. This finding is consistent with earlier evidence that professional forecasters across different industries do not use or under-

utilize algorithmic advice when making predictions (Fildes and Goodwin, 2007). Similarly, Sanders and Manrodt (2003) report that firms are reluctant to adopt quantitative forecasting models over judgmental approaches, even when quantitative models reduce forecasting errors.

Recent studies suggest that aversion may be partly driven by egocentric discounting and that, in delegation settings, individuals may rely on algorithms more readily than on other humans (Ivanova-Stenzel and Tolksdorf, 2024, 2025). Indeed, Logg et al. (2019) document *algorithm appreciation*, whereby people prefer algorithmic judgment to human judgment. Yet, greater reliance on algorithms need not imply appropriate reliance, as individuals may also overweigh algorithmic advice in ways that reduce the accuracy of their final judgments relative to normative (Bayesian) benchmarks (Logg et al., 2019; Mahmud et al., 2022). Complementing these findings, it has been shown that the uptake of algorithmic advice is shaped by contextual drivers such as performance feedback (You et al., 2022) and incentives (Chacon et al., 2025; Greiner et al., 2026; Normann et al., 2025).

Much of the advice-taking literature quantifies advice utilization by how far decision makers shift their point estimates toward advice. Yet, this captures only one dimension of belief updating. Advice may leave a point estimate largely unchanged while substantially increasing or decreasing confidence in that estimate. Consequently, observing little adjustment of a point estimate does not necessarily imply that advice had little influence on the decision maker's beliefs.

For example, consider a medical professional who wants to assess whether a finding on a medical scan is abnormal. This professional may not change an accurate assessment after receiving (algorithmic) advice, yet may become more confident in that assessment and the subsequent treatment decision. This is particularly consequential when treatment involves (severe) side effects.

Despite its importance, the effects of algorithmic advice on individuals' confidence in their own decisions have received relatively little attention in the literature. Bogert et al. (2021), for example, find that receiving algorithmic advice rather than human advice leads to more updating and greater confidence in a crowd-size estimation task. Conversely, Chong et al. (2022) find that self-confidence rather than confidence in AI is an important determinant of whether individuals accept algorithmic advice. Our design complements this evidence by exogenously varying perceived task difficulty and advice quality, allowing us to causally identify how these factors shape both advice uptake and confidence in final judgments.

We contribute to the emerging literature on the use of algorithmic advice in human decision-making by quantitatively decomposing the effect of algorithmic advice into an *opinion shift* and a *confidence change*, using the methodology introduced by Soll et al. (2022), rather than solely quantifying the effect of algorithmic advice on final predictions. Additionally, we compare the actual updating behavior with the Bayes-optimal benchmark to identify whether and through which channel advice is discounted. Together, these approaches allow us to examine advice uptake and changes in confidence following advice, and to compare both dimensions with optimal updating. To this end, we run an online experiment in which we exogenously vary perceived task difficulty and the observable performance of the algorithmic advisor. Participants had to predict the median age of a group and report their confidence via their expected absolute error before and after receiving advice from the algorithm. By jointly varying these two dimensions, our design also speaks to the broader advice-taking literature on how advice quality and perceived task difficulty shape advice-taking (e.g., Harvey and Fischer, 1997; Gino and Moore, 2007; Bonaccio and Dalal, 2006).

Our findings indicate that focusing only on advice uptake misses important dimensions of how decision makers incorporate external information. In particular, we find that perceived task difficulty does not affect the extent to which individuals shift their predictions toward advice, but does affect their confidence in the final prediction. When individuals perceive a task as easier, receiving high-quality advice increases confidence in their own predictions even when they barely adjust these predictions toward the advice. Thus, advice can influence beliefs without substantially changing point predictions. By exogenously varying perceived task difficulty and advice quality, we show how these two factors shape the extent to which advice operates through opinion shifts and confidence changes. This distinction is particularly relevant in professional settings, where algorithmic advice may influence decision makers even without inducing observable changes in their decisions. Assessing algorithmic advice solely through changes in decisions may therefore overlook an important channel through which such systems affect human decision-making.

Further, comparing observed updating with the Bayes-optimal benchmark, we find that participants systematically underutilize advice, with the shortfall being particularly pronounced for high-quality advice. While opinion updating following high-quality advice is closer to the Bayes-optimal benchmark in easy tasks, participants update their self-confidence insufficiently.

2 Experiment

To examine the relationship between advice uptake and confidence in the final prediction when receiving advice from an algorithm, we conducted an incentivized online experiment on Prolific. Participants predicted the median age of a group of people, represented as a collage of nine face portraits, and estimated the accuracy of their prediction via their expected absolute error, defined as the absolute difference they expected between their age prediction and the true median age (henceforth, MISS).[1] After making their age prediction and associated MISS estimate, participants received advice from an algorithm and could change both.

We implemented a 2×2 between-subject design, varying perceived task difficulty (Easy vs. Hard) and algorithm performance (Weak vs. Strong) as depicted in Figure 1.

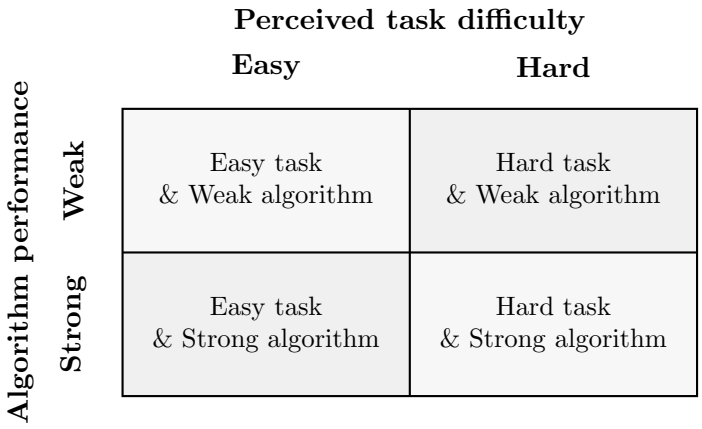


Figure 1: Treatment variations.

Perceived task difficulty was manipulated via the age composition of the collage: in Hard, the collage contained nine middle-aged individuals (homogeneous group); in Easy, it contained four young, one middle-aged, and four old individuals (heterogeneous group).[2] In both difficulty conditions, the gender of the (true) median-aged person was counterbalanced across collages, as it might affect age-prediction accuracy. For ease of readability, we refer to the task that is perceived to be hard as the "hard task" and the task that is perceived to be easy as the "easy task".

[1] All images used in the age-prediction tasks were neutral-expression facial portraits drawn from the FACES database (Ebner et al., 2010).

[2] We pretested the perceived difficulty with a sample of 40 workers on Prolific. In this pre-study, participants had to predict the median age of a group and indicate how difficult they perceived this task to be on a 7-point Likert scale. We find that the more homogeneous the group was, the harder the task was perceived to be (see Section A.1 in the Appendix).

Algorithm performance was varied between Strong and Weak. Participants learned about the algorithm's performance through feedback during the training rounds described below. Importantly, participants were not informed about the type of algorithm generating the predictions (LLM vs. deep learning), such that the manipulation concerned observed algorithm performance rather than information about the underlying algorithm type.

Before the main task, participants completed 12 training rounds in which they predicted the age of a single portrait and reported their associated MISS estimate.[3] The training set consisted of 12 portraits: two male and two female portraits from each of the three age categories (young, middle-aged, and old). The order of the portraits was randomly determined for each participant.

After each training prediction, participants observed the algorithm's actual MISS for the corresponding portrait, allowing them to learn about its prediction accuracy. At the end of the training phase, participants received feedback about both their own realized average MISS and the algorithm's average MISS across the training rounds. The Strong algorithm had an average MISS of 2.83 years (min = 0, max = 5), whereas the Weak algorithm had an average MISS of 6.08 years (min = 3, max = 16).[4]

In the main task before receiving advice, participants (i) identified which person in the group of nine they believed to be the median-aged person, (ii) reported their median-age prediction, and (iii) reported their MISS estimate. An example of this decision screen can be found in Figure A.6 in the Appendix. After observing the algorithm's median-age prediction, participants could revise their median-age prediction and MISS estimate. For an example of this second decision screen, see Figure A.7 in the Appendix.

We quantify advice-taking in two complementary ways. First, in line with the standard approach in the algorithmic advice-taking literature, we compute the Weight on Advice (WOA) as follows:

$$\text{WOA} = \frac{\mu_F - \mu_I}{\mu_A - \mu_I}, \tag{1}$$

where μ_I denotes the human participant's initial prediction, μ_A the algorithmic advice, and μ_F the

[3]Following Soll et al. (2022), we carefully explained the MISS to participants and provided a numerical example to facilitate comprehension.

[4]Predictions from the Strong algorithm were generated using *ChatGPT Edu* with the prompt: *"Can you please estimate the age of the person in the following picture? Please give an exact number. An estimate to the best of your capacity is enough."* For the Weak algorithm, we used the deep learning framework Keras (Chollet et al., 2015) to train a model on the UTKFace database (Zhang et al., 2017).

participant's final prediction. WOA = 0 indicates completely ignoring the advice and WOA = 1 indicates fully adopting it.

Second, going beyond the standard WOA, we compute the Influence of Advice (IOA), which captures changes in both the location and dispersion of participants' subjective distributions (Soll et al., 2022). The IOA incorporates both (a) the shift in the mean prediction and (b) the change in uncertainty as proxied by participants' reported MISS estimate. We compute the IOA as follows:

$$\begin{aligned} \text{IOA} &= 1 - \frac{(\mu_F - \mu_A)^2 + \sigma_F^2}{(\mu_I - \mu_A)^2 + \sigma_I^2} \\ &\iff \underbrace{\frac{(\mu_I - \mu_A)^2 - (\mu_F - \mu_A)^2}{(\mu_I - \mu_A)^2 + \sigma_I^2}}_{\text{OS}} + \underbrace{\frac{\sigma_I^2 - \sigma_F^2}{(\mu_I - \mu_A)^2 + \sigma_I^2}}_{\text{CC}}, \end{aligned} \tag{2}$$

where σ_I^2 and σ_F^2 denote the participant's initial and final prediction-error variances. To estimate σ_I and σ_F, we multiply the estimate of the respective MISS with $\sqrt{\pi/2}$, as Johnson et al. (1995) show that the expected absolute error for a normal distribution is given by $\sigma\sqrt{2/\pi}$. The first component in Equation (2) represents the Opinion Shift (OS) and the second component the Confidence Change (CC). Thus, the IOA quantifies the proportional overall movement of a probability distribution toward the advice, whereas the WOA measures the shift of a point estimate toward the advice. The IOA can therefore be viewed as the distributional counterpart of the WOA.

Bayes-optimal benchmark

In contrast to Soll et al. (2022), participants in our design receive precise feedback on both their own and the algorithm's performance during the training rounds. This allows us to derive Bayes-optimal benchmarks for WOA, IOA, and the two IOA subcomponents based on relative signal precision and assess participants' deviations from optimal updating.

Under the assumption that both signals are unbiased, independent, and normally distributed, a Bayesian decision maker updates the initial prediction using the algorithmic advice according to the relative precision of the two signals, estimated from the training-round performance. Consequently,

the Bayes-optimal weight placed on the advice is given by

$$\text{WOA}^{Bayes} = \frac{1/\sigma_A^2}{1/\sigma_I^2 + 1/\sigma_A^2}, \tag{3}$$

where σ_A^2 denotes the algorithm's prediction error variance in training rounds and σ_I^2 denotes the human participant's prediction error variance.

The Bayes-optimal WOA can equivalently be computed from the squared average MISS observed during the training rounds because, substituting $\sigma = \text{MISS}\sqrt{\pi/2}$, the common proportionality factor cancels when forming the precision weights. The corresponding Bayes-optimal posterior prediction is $\mu_B = (1-\text{WOA}^{Bayes})\cdot\mu_I + \text{WOA}^{Bayes}\cdot\mu_A$. Under independent normal prediction errors, Bayesian updating implies that posterior precision equals the sum of the human and algorithmic precisions. Hence, the posterior prediction-error variance is $\sigma_B^2 = (\sigma_I^2\sigma_A^2)/(\sigma_I^2 + \sigma_A^2)$. Using μ_B and σ_B^2, the Bayes-optimal IOA, as well as the corresponding Opinion Shift and Confidence Change, can be derived as follows:

$$\begin{aligned} \text{IOA}^{Bayes} &= 1 - \frac{(\mu_B - \mu_A)^2 + \sigma_B^2}{(\mu_I - \mu_A)^2 + \sigma_I^2}, \\ &\iff \underbrace{\frac{(\mu_I - \mu_A)^2 - (\mu_B - \mu_A)^2}{(\mu_I - \mu_A)^2 + \sigma_I^2}}_{\text{OS}^{Bayes}} + \underbrace{\frac{\sigma_I^2 - \sigma_B^2}{(\mu_I - \mu_A)^2 + \sigma_I^2}}_{\text{CC}^{Bayes}}. \end{aligned} \tag{4}$$

Experimental procedure

Elicitations of the median age prediction and the MISS estimate were incentivized using a binarized scoring rule (BSR). One training round was randomly selected for payment, and participants were paid based on either their age prediction or MISS estimate, with the payoff-relevant measure determined randomly to prevent hedging. Following Soll and Larrick (2009), in the main task, we incentivized the final prediction and, again, randomly determined whether the age prediction or the MISS estimate was payoff relevant. As suggested by Danz et al. (2022), we informed the participants that it is in their best interest to give their true prediction and provide a description of the BSR only as additional information (accessible via an on-click pop-up window in the instructions).

Participants who did not provide consent or failed any attention or comprehension checks were

excluded from the analysis.[5] Additionally, we implemented the picture-rearranging task introduced by Laske et al. (2024), which participants had to complete before entering the study. In this task, participants were asked to arrange differently colored shapes to reproduce a target figure. This required them to rotate and layer the shapes using left- and right-clicks, thereby screening out simple bots capable of filling out form fields. A screenshot of the task is provided in Figure A.3 in the Appendix. Overall, 1,696 Prolific participants based in the UK started the experiment. 905 were excluded because they did not give consent for their data being used, failed the picture assembly task, failed the comprehension checks, or dropped out for other reasons, yielding a sample of 793 participants.[6] Following our pre-registration, we excluded 35 participants who switched browser tabs during the main prediction task or while updating their predictions after receiving the advice, thereby reducing the risk of data contamination from the use of external AI tools (see Rilla et al., 2025 for a discussion). Furthermore, as pre-registered, we excluded participants whose age predictions in the main task deviated from the mean by three standard deviations or more. This yielded a final sample of 749 participants. On average, participants took around 14.5 minutes (median 12.9 minutes) to complete the experiment and earned on average £3.46 including the fixed payment of £2.

Hypotheses

Following the extant literature on advice-taking (see, e.g., Bonaccio and Dalal, 2006; Bailey et al., 2023), we hypothesize that the quality of advice increases the uptake of advice:

Hypothesis 1. *Holding perceived task difficulty constant, both the WOA and IOA are higher with a strong algorithm than with a weak algorithm.*

Gino and Moore (2007) show that people overweigh advice in difficult tasks and underweigh advice in easy tasks. Similarly, we expect that an increase in perceived task difficulty should increase the uptake of advice. This leads to our second hypothesis:

Hypothesis 2. *Holding algorithm performance constant, both WOA and IOA are higher in the hard task than in the easy task.*

[5] Participants had to solve two comprehension questions about the median and two questions about the MISS correctly before participating. Screenshots of the control questions can be found in Appendix Section A.5.3.

[6] 237 participants failed the comprehension questions regarding the median on two attempts, 74 failed the comprehension questions about the MISS twice. All of these were excluded from participating further in the experiment.

Conversely, participants may adjust their predictions only slightly while becoming less confident in them. Such a response could result in a negative shift (particularly in the case of the IOA).

Assuming Hypotheses 1 and 2 hold, or that either effect dominates, we expect the corollary joint effect of algorithm performance and perceived task difficulty, leading to our third hypothesis:

Hypothesis 3. *WOA and IOA are higher with a strong algorithm in the hard task than with a weak algorithm in the easy task.*

Beyond examining the overall Influence of Advice (IOA), we decompose it into its two components, the opinion shift and the confidence change (Soll et al., 2022). This decomposition provides a more nuanced understanding of how algorithmic advice shapes human decision-making by distinguishing between changes in participants' predictions and changes in their confidence in those predictions.

3 Results

Before turning to the main task performance, we present the participants' performance in the training rounds. Participants reported an average MISS of 4.50 years, while their realized MISS averaged 6.25 years in the weak algorithm condition and 6.39 years in the strong algorithm condition, with no significant difference between the distributions (Kolmogorov–Smirnov test: $D = 0.052$, $p = 0.700$). Figure A.1 in the Appendix shows the corresponding distributions. We find that participants, on average, overestimate their ability to predict the ages in the training rounds. In the strong algorithm condition, no participant outperformed the algorithm, while 50.53% did so in the weak algorithm condition. Thus, the weak algorithm performed comparably to participants, whereas the strong algorithm consistently outperformed them.

In the main task, participants' average MISS estimate is 5.01 years while their realized MISS is 5.49 years on average. This indicates that, on average, participants remain overconfident about the accuracy of their age predictions when moving from the training phase to the main task. Examining the perceived task difficulty manipulation, we find that the average MISS estimate is 4.88 in the easy task and 5.13 in the hard task. Following the literature, we define overconfidence as the actual absolute error minus the expected absolute error. Positive values indicate overconfidence, i.e., participants believe their age prediction is closer to the true age than it actually is, whereas

negative values indicate underconfidence. Participants in the easy task are, on average, overconfident in their age predictions (mean = 1.10), whereas participants in the hard task are, on average, slightly underconfident (mean = −0.14). This difference is statistically significant ($p < 0.001$, two-sided t-test). Figure A.2 in the Appendix shows the corresponding distributions.[7] Thus, participants in the easy task were more confident in the accuracy of their age predictions than those in the hard task, consistent with the intended manipulation of perceived task difficulty.

After receiving algorithmic advice, 354 participants changed their age prediction and 202 changed their MISS estimate. Of the 354 participants who changed their age prediction, 351 updated it toward the advice, while 169 of the 202 participants who changed their MISS estimate reported a smaller MISS after receiving advice. Table 1 breaks down this updating behavior by treatment and shows that participants updated more often when they received advice from the strong algorithm.

	Easy task		Hard task	
	Weak algorithm	Strong algorithm	Weak algorithm	Strong algorithm
Changed age prediction	84	106	71	93
↪ toward advice	83	106	70	92
Changed MISS estimate	32	72	39	59
↪ toward zero	23	62	33	51
No. of observations	191	184	187	187

Table 1: Change of age prediction after advice by perceived task difficulty and algorithm performance.

We calculate the WOA and the IOA as described in Section 2. For 3.5% of the responses, the WOA is undefined because the initial prediction equals the advice, causing the denominator of Equation (1) to be zero.[8] In Table 2, we break down the means and standard deviations of the main measures (WOA and IOA), as well as the components of IOA (opinion shift and confidence change) by task difficulty and algorithm performance.[9] It is apparent that both WOA and IOA increase from weak to strong algorithms for both levels of perceived task difficulty. While this qualitative

[7] The distributions of overconfidence also differ significantly between the perceived task difficulty conditions (Kolmogorov–Smirnov test: $D = 0.122$, $p < 0.001$).

[8] This fraction of undefined WOA is in line with other studies employing the WOA (see, e.g., Soll et al., 2022; Greiner et al., 2026), where around 5% of responses have an undefined WOA.

[9] In line with the literature, we observe an average WOA of 0.3, indicating egocentric advice-taking. Soll and Larrick (2009) show that this average masks a tri-modal distribution, characterized by frequent instances of advice ignoring, a substantial share of equal weighting, and a smaller third mode reflecting full acceptance of advice.

observation applies to the opinion shift and confidence change as well, the opinion shift appears to increase more between the weak and strong algorithm conditions in the hard task than in the easy task, whereas the opposite is true for the confidence change, indicating that perceived task difficulty is a driving factor for the perception of algorithm performance.

	Easy		Hard	
	Weak algorithm	Strong algorithm	Weak algorithm	Strong algorithm
Mean WOA (SD)	0.24 (0.32)	0.42 (0.39)	0.22 (0.33)	0.39 (0.48)
No. of Observations	189	175	180	178
Mean IOA (SD)	0.25 (0.32)	0.39 (0.35)	0.19 (0.31)	0.35 (0.41)
Mean Opinion Shift (SD)	0.20 (0.28)	0.26 (0.29)	0.13 (0.23)	0.26 (0.32)
Mean Confidence Change (SD)	0.04 (0.16)	0.13 (0.28)	0.06 (0.21)	0.08 (0.30)
No. of Observations	191	184	187	187

Table 2: Summary statistics of Weight on Advice (WOA), Influence of Advice (IOA) and IOA components (Opinion Shift and Confidence Change) by perceived task difficulty and algorithm performance.

To test whether these differences are statistically significant, Table 3 reports OLS regression results for the main dependent variables: WOA (Panel A) and IOA (Panel B). In all regressions, we control for gender, age, and performance in training rounds. In specification (1), we estimate the direct effects of perceived task difficulty and algorithm performance. Perceived task difficulty is captured by the dummy variable *Hard task*, which equals 1 for the hard task condition and 0 for the easy task condition. Algorithm performance is captured by the dummy variable *Strong algorithm*, which is equal to 1 in the strong algorithm condition and 0 in the weak algorithm condition. We find a statistically significant effect of *Strong algorithm* on both WOA and IOA, consistent with Hypothesis 1. We do not observe a significant effect of *Hard task* on WOA, whereas we do observe an effect on the IOA, contrary to Hypothesis 2. When including interaction effects in specification (2), these main observations carry through, whereas the interaction *Hard task* $\times$ *Strong algorithm* is not significant in either case, indicating that the effect of getting advice from a strong algorithm does not significantly differ between the two perceived task difficulties.

We utilize post-estimation Wald tests to directly compare specific treatment variations. We find that participants exhibit both higher WOA and IOA when paired with a strong algorithm in the hard task condition than when paired with a weak algorithm in the easy task condition

	(1) w/o interaction	p-value	(2) with interaction	p-value
Panel A: Weight on Advice (WOA)				
Hard task	-0.028	[0.326]	-0.023	[0.468]
	(0.028)		(0.032)	
Strong algorithm	0.166***	[0.000]	0.171***	[0.000]
	(0.029)		(0.036)	
Hard task × Strong algorithm			-0.009	[0.877]
			(0.057)	
Constant	-0.086	[0.289]	-0.088	[0.288]
	(0.081)		(0.083)	
Controls	✓		✓	
Observations	722		722	
Postestimation Wald test:				
H_0: Hard task + Strong algorithm + Hard task × Strong algorithm = 0				[0.001]
H_0: Strong algorithm > Hard task				[0.000]
Panel B: Influence of Advice (IOA)				
Hard task	-0.050**	[0.047]	-0.057*	[0.070]
	(0.025)		(0.031)	
Strong algorithm	0.147***	[0.000]	0.140***	[0.000]
	(0.025)		(0.034)	
Hard task × Strong algorithm			0.013	[0.794]
			(0.051)	
Constant	-0.042	[0.575]	-0.039	[0.605]
	(0.075)		(0.075)	
Controls	✓		✓	
Observations	749		749	
Postestimation Wald test:				
H_0: Hard task + Strong algorithm + Hard task × Strong algorithm = 0				[0.010]
H_0: Strong algorithm < Hard task				[1.000]

Estimation by OLS regression with robust standard errors in parentheses and p-values in square brackets. Controls are gender, age, and performance in training rounds. *, ** and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 3: OLS regression on Weight on Advice (Panel A) and Influence of Advice (Panel B).

($p = 0.001$ for WOA; $p = 0.010$ for IOA). These findings are in line with Hypothesis 3. Notably, the results above are robust to restricting the analysis to treatment combinations in which actual task performance better aligns with perceived task difficulty, i.e., performance in the easy task is better than (or at least as good as) in the hard task. For this robustness check, we exclude the easy task condition with a female median person, as this variation solely drives the lower performance in the easy task. Re-estimating the above specifications using all remaining treatment combinations, we find that our results largely replicate. For details, see Section A.4 in the Appendix.

In addition to the hypothesis tests above, our pre-registration specified a further directional comparison between the strong algorithm in the easy task and the weak algorithm in the hard task. Although we now regard this comparison as exploratory (for the reasons discussed below), we report it here in the interest of transparency. Our intuition was that algorithm performance would dominate perceived task difficulty with respect to the WOA, implying a higher WOA with the strong algorithm in the easy task. In contrast, for the IOA, we conjectured that participants' predictions and the algorithmic advice would, on average, be farther apart with the weak algorithm in the hard task than with the strong algorithm in the easy task. Following the framework of Soll et al. (2022), larger advice distances imply relatively larger opinion shifts, whereas smaller advice distances primarily affect confidence. Thus, if the larger opinion shifts outweighed the larger confidence changes, the IOA would be expected to be higher with the weak algorithm in the hard task than with the strong algorithm in the easy task. Upon revisiting the theoretical framework, we concluded that a directional prediction relies on additional assumptions about the relative advice distances that are not directly implied by the treatment manipulations. We therefore report this pre-registered comparison as an exploratory analysis rather than as a confirmatory hypothesis. We find that the WOA is significantly larger when paired with a strong algorithm in the easy task compared with the weak algorithm in a hard task ($p < 0.001$), while we do not find evidence for the IOA being larger when paired with a weak algorithm in the hard task compared with a strong algorithm in the easy task ($p = 1.000$).

In Table 4, we decompose the effects of perceived task difficulty and algorithm performance on the IOA into two components: opinion shift and confidence change. This decomposition allows us to distinguish between adjustments of participants' predictions toward the advice and changes in their confidence.

Specification (1) of both panels indicates that both the opinion shift and the confidence change are positively affected by algorithm performance. Hence, participants change their prediction more toward the advice of the strong algorithm and become more confident in their prediction after receiving advice from the strong algorithm, indicating that the aforementioned effect on the IOA is driven by both of its components.

Including the interaction effect *Hard task* × *Strong algorithm* in specification (2) reveals that the effect of *Strong algorithm* on the opinion shift is particularly driven by participants who are paired

	(1) w/o interaction	p-value	(2) with interaction	p-value
Panel A: Opinion Shift				
Hard task	-0.033	[0.103]	-0.073***	[0.005]
	(0.020)		(0.026)	
Strong algorithm	0.089***	[0.000]	0.049*	[0.086]
	(0.020)		(0.028)	
Hard task × Strong algorithm			0.081**	[0.046]
			(0.040)	
Constant	-0.084	[0.164]	-0.064	[0.294]
	(0.061)		(0.061)	
Controls	✓		✓	
Observations	749		749	
Postestimation Wald test:				
H_0: Hard task + Hard task × Strong algorithm = 0				[0.801]
Panel B: Confidence Change				
Hard task	-0.017	[0.325]	0.016	[0.403]
	(0.018)		(0.019)	
Strong algorithm	0.058***	[0.001]	0.092***	[0.000]
	(0.018)		(0.024)	
Hard task × Strong algorithm			-0.067*	[0.058]
			(0.036)	
Constant	0.042	[0.378]	0.025	[0.596]
	(0.048)		(0.047)	
Controls	✓		✓	
Observations	749		749	
Postestimation Wald test:				
H_0: Strong algorithm + Hard task × Strong algorithm = 0				[0.363]

Estimation by OLS regression with robust standard errors in parentheses and p-values in square brackets. Controls are gender, age, and performance in training rounds. *, ** and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 4: OLS regression on opinion shift (Panel A) and confidence change (Panel B).

with the strong algorithm in the hard task. This further shows a negative and significant effect of *Hard task*, indicating that the opinion shift is lower for participants who are paired with a weak algorithm in the hard task when controlling for the full interaction of perceived task difficulty and algorithm performance. The post-estimation Wald test indicates that a strong algorithm offsets this negative effect of an increase in perceived task difficulty on the opinion shift ($p = 0.801$), explaining the absence of this effect when not considering the interaction.

In contrast, in the case of the confidence change (Panel B of Table 4), the positive effect of *Strong algorithm* is driven by participants who are paired with a strong algorithm in the easy task, whereas this effect disappears when considering the hard task, as indicated by the negative and

significant interaction effect. Again, the post-estimation Wald test indicates that the interaction effect offsets the direct effect of the strong algorithm in the easy task ($p = 0.363$).

We now summarize our findings with respect to the three hypotheses stated in Section 2, the pre-registered exploratory comparison, and the decomposition of the IOA. Our first result concerns the effect of algorithm performance, holding perceived task difficulty constant:

Result 1. *An increase in algorithm performance increases both the Weight on Advice and Influence of Advice. The latter is driven by an opinion shift for strong algorithms in hard tasks and a change in participants' confidence following advice from strong algorithms in easy tasks.*

Our second result concerns the effect of perceived task difficulty, holding the algorithm performance constant:

Result 2. *We find no effect of increased perceived task difficulty on the Weight on Advice and a negative effect of increased task difficulty on the Influence of Advice. The latter is driven by a negative opinion shift in hard tasks.*

Our third result considers the joint impact of a simultaneous increase in perceived task difficulty and algorithm performance:

Result 3a. *We find that both the Weight on Advice and Influence of Advice are larger in hard tasks with strong algorithms than in easy tasks with weak algorithms.*

We next consider the pre-registered exploratory comparison between the strong algorithm in the easy task and the weak algorithm in the hard task. While the confidence change is the major component when a strong algorithm advises in the easy task, we do not observe that participants follow advice from a weak algorithm in the hard-task condition:

Result 3b. *We confirm that the Weight on Advice is larger with a strong algorithm in an easy task than with a weak algorithm in a hard task. We cannot confirm that the Influence of Advice is larger with a weak algorithm in a hard task than with a strong algorithm in an easy task.*

The decomposition of the IOA provides important new insights. In particular, we observe a differentiated impact of weak and strong algorithms on opinion shift and confidence change for different levels of task difficulty:

Result 4a. *Increased perceived task difficulty decreases the opinion shift (change of prediction toward advice) for weak algorithms but not for strong algorithms.*

Result 4b. *A strong algorithm positively affects the confidence (lower MISS estimate) in the easy task but not in the hard task.*

Comparison to Bayes-optimal benchmark

We next compare observed advice-taking with the Bayes-optimal benchmarks derived in Section 2. This allows us to assess whether and by how much participants deviate from optimal updating, and whether these deviations vary with perceived task difficulty and advice quality, providing novel insights into the extent and nature of egocentric discounting of advice. To do so, we calculate the difference between Bayes-optimal and actual updating for each of our four outcomes. A value of zero indicates Bayes-optimal updating, while positive values indicate under-updating, with larger values reflecting a greater shortfall in updating relative to the Bayesian benchmark. We estimate OLS regressions of the difference measures on task difficulty and algorithm performance, as reported in Table A.2 in the Appendix. Figure 2 presents margins plots based on specification (2) of these regressions, which includes the interaction between task difficulty and algorithm performance.

In line with the literature on advice-taking, we find evidence of egocentric discounting. Participants put less weight on the advice than is Bayes-optimal, with this gap being even larger when the advice comes from a strong algorithm (Figure 2a). This observation is not affected by an increase in perceived task difficulty. Similarly, for IOA, we find underutilization of advice, with the gap from the Bayes-optimal benchmark being larger for advice from a strong algorithm (Figure 2b). Thus, despite having information about both their own performance and the advisor's precision, participants still underutilize the advice, and, in particular, they do not account for the high informational value of high-quality advice.

Turning to the two components of the IOA, OS and CC, reveals a more nuanced pattern than the composite measures. For OS, participants in the easy task under-update most when receiving advice from the weak algorithm, whereas their updating following advice from the strong algorithm is more closely aligned with the Bayes-optimal benchmark. This difference in updating between the weak and strong algorithms does not persist in the hard task (Figure 2c). Lastly, for CC, we again observe a discrepancy between responses to weak and strong algorithmic advice similar to WOA and IOA. Updating falls considerably further below the Bayes-optimal benchmark following advice from the strong algorithm than following advice from the weak algorithm. This difference narrows

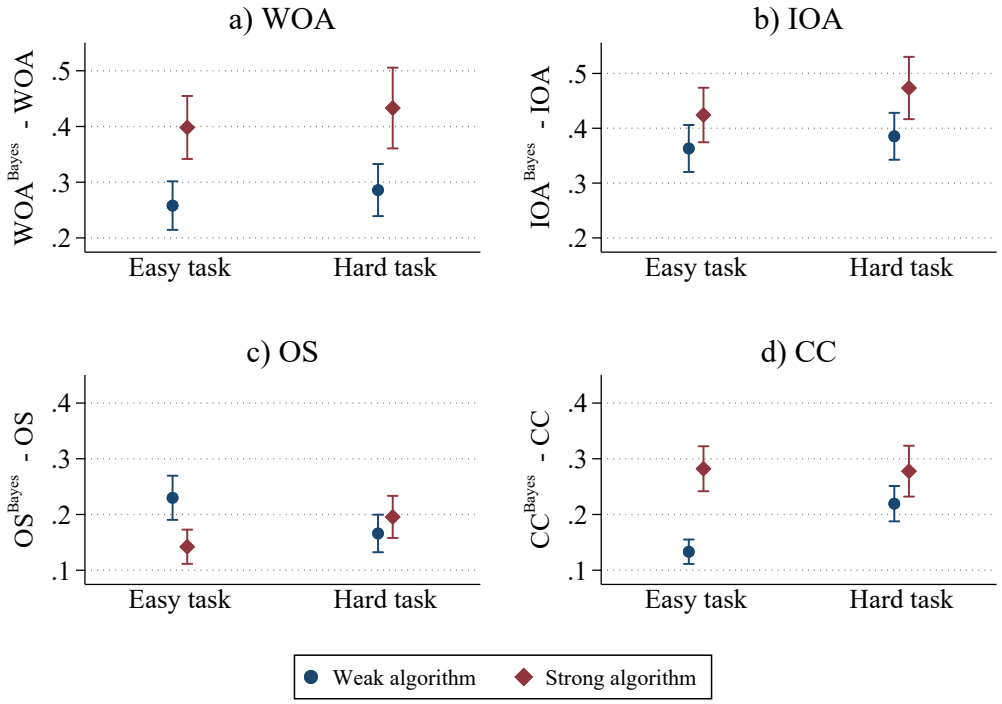


Figure 2: Distance between the Bayes-optimal and observed advice-taking measures by perceived task difficulty and algorithm performance: a) Weight on Advice (WOA), b) Influence of Advice (IOA), c) Opinion Shift (OS), and d) Confidence Change (CC). Points represent regression-adjusted mean differences, controlling for gender, age, and performance in training rounds; whiskers indicate 95% confidence intervals.

but persists in the hard task (Figure 2d).[10]

Taken together, the underutilization of advice from the strong algorithm observed for WOA and IOA does not seem to be solely driven by insufficient adjustment of the point estimate in the easy task. Instead, it appears to reflect insufficient confidence updating. This finding is particularly noteworthy because the widely used WOA measure is based purely on changes in point estimates before and after receiving advice, while the mechanism underlying the observed advice-taking behavior may actually lie in differences in confidence that remain unobserved when confidence is not elicited separately.

[10]Notably, the combination of an easy task and a strong algorithm nevertheless produces the largest absolute confidence change as discussed earlier. These findings are not at odds: participants respond to differences in advice quality in the predicted direction, but not to the Bayes-optimal magnitude.

4 Discussion

We replicate the common finding of the literature that the Weight on Advice (WOA) increases with advice quality.[11] At the same time, we observe egocentric discounting of advice. Implementing the Influence of Advice (IOA) measure developed by Soll et al. (2022), our design allows us to go beyond this observation and obtain a more nuanced understanding of how algorithmic advice is used across different levels of perceived task difficulty and algorithm performance. We find that the IOA is affected by the perceived difficulty of the task, whereas the WOA is not. Decomposing the IOA into its subcomponents, we find that an increase in perceived task difficulty reduces the opinion shift toward advice when advice comes from a weak algorithm, whereas we do not observe this effect when advice comes from a strong algorithm. In contrast, participants become more confident in their prediction after receiving advice from a strong algorithm when the task is perceived as easy, whereas this is not the case when the task is perceived as difficult. Our findings show that focusing solely on the WOA can provide an incomplete picture of how advice affects decision makers, as it may overstate advice-taking when movements toward advice are accompanied by decreases in confidence and understate it when little movement is accompanied by increased confidence. Importantly, our results show that the extent to which these two dimensions diverge depends both on perceived task difficulty and on advisor performance, highlighting a potentially important source of heterogeneity in advice-taking that is not typically observed when either dimension is held constant.

Comparing participants' updating behavior to the normative Bayesian benchmark further reveals that these patterns cannot be solely explained by differences in the relative precision of participants' own judgments and the advice. Participants systematically place less weight on advice than is Bayes-optimal, providing corroborating evidence of egocentric discounting even when they have information about both their own and the advisor's performance. This deviation is particularly pronounced for high-quality advice. Participants fail to fully account for the informational advantage of high-quality advice and therefore underutilize advice that should receive greater weight.

These results are robust to aligning actual task performance with perceived task difficulty. Re-estimating our main specifications after excluding the treatment variation responsible for the discrepancy between actual task performance and the intended difficulty manipulation yields qual-

[11] See Bailey et al. (2023) for a meta-analysis of advice-taking in general and, e.g., Logg et al. (2019); Greiner et al. (2026) for evidence on algorithmic advice.

itatively similar results, indicating that the observed effects are driven by perceived rather than actual task difficulty (see Section A.4 in the Appendix).

Perceived task difficulty and quality of advice reinforce each other in two ways. First, in tasks perceived as difficult, humans rely less on the advice when the algorithm generating the advice is relatively weak, indicating that they overestimate their own predictive ability. Put differently, when in doubt about either party being able to fulfill the task, humans tend to rely more on themselves. This might reflect egocentric discounting and is in line with the observation that participants overweigh positive feedback compared to negative feedback (see, e.g., Möbius et al., 2022). High-quality advice by a strong algorithm negates this effect of perceived task difficulty. Second, in tasks perceived as easy, humans do not rely more on high-quality advice from strong algorithms, which, based on advice uptake alone, would often have been interpreted as an underutilization of (algorithmic) advice. However, we find that humans have more confidence in their predictions when advised by a strong algorithm. Since they perceive the task as relatively easy, they may not see the need to change their prediction, while becoming more certain to have made an accurate prediction.

More broadly, advice can meaningfully affect decision makers even when it does not directly affect their choices, by shaping their confidence in those choices. This is particularly relevant in professional settings in which decision makers are highly proficient in their tasks and increasingly receive advice from algorithmic systems. In such settings, algorithms may serve as quality-assurance tools by inducing decision changes when warranted and reassuring decision makers when their judgments are corroborated by high-quality advice. Consequently, organizations evaluating algorithmic decision-support systems solely by the extent to which they change employees' decisions may underestimate their influence. At the same time, our comparison with the Bayes-optimal benchmark shows that decision makers do not fully incorporate the informational value of advice, particularly when its quality is high. Taken together, our findings show that how advice shapes decisions and confidence depends on perceived task difficulty and advisor performance, and that considering both dimensions relative to a normative benchmark is important for understanding when and how advice is underutilized.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used ChatGPT in order to check for spelling and punctuation mistakes, improve the readability of the manuscript, and make the Python code used to conduct the experiment more efficient. Additionally, the authors used ChatGPT to get age predictions for pictures that are used in the experiment, as outlined above. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G. et al. (2019), 'End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography', *Nature Medicine* **25**(6), 954–961.

Bailey, P. E., Leon, T., Ebner, N. C., Moustafa, A. A. and Weidemann, G. (2023), 'A meta-analysis of the weight of advice in decision-making', *Current Psychology* **42**(28), 24516–24541.

Bick, A., Blandin, A. and Deming, D. J. (2026), 'The rapid adoption of generative AI', *Management Science* **0**(0).

Bogert, E., Schecter, A. and Watson, R. T. (2021), 'Humans rely more on algorithms than social influence as a task becomes more difficult', *Scientific Reports* **11**(1), 8028.

Bonaccio, S. and 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.

Chacon, A., Kausel, E. E., Reyes, T. and Trautmann, S. (2025), 'Preventing algorithm aversion: People are willing to use algorithms with a learning label', *Journal of Business Research* **187**, 115032.

Chollet, F. et al. (2015), 'Keras'.
URL: *https://github.com/fchollet/keras*

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K. and Cagan, J. (2022), 'Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice', *Computers in Human Behavior* **127**, 107018.

Danz, D., Vesterlund, L. and Wilson, A. J. (2022), 'Belief elicitation and behavioral incentive compatibility', *American Economic Review* **112**(9), 2851–2883.

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

Ebner, N. C., Riediger, M. and Lindenberger, U. (2010), 'Faces—a database of facial expressions in young, middle-aged, and older women and men: Development and validation', *Behavior Research Methods* **42**, 351–362.

Fildes, R. and Goodwin, P. (2007), 'Against your better judgment? how organizations can improve their use of management judgment in forecasting', *Interfaces* **37**(6), 570–576.

Gino, F. and Moore, D. A. (2007), 'Effects of task difficulty on use of advice', *Journal of Behavioral Decision Making* **20**(1), 21–35.

Greiner, B., Grünwald, P., Lindner, T., Lintner, G. and Wiernsperger, M. (2026), 'Incentives, framing, and reliance on algorithmic advice: An experimental study', *Management Science* **72**(1), 302–322.

Harvey, N. and Fischer, I. (1997), 'Taking advice: Accepting help, improving judgment, and sharing responsibility', *Organizational Behavior and Human Decision Processes* **70**(2), 117–133.

Ivanova-Stenzel, R. and Tolksdorf, M. (2024), 'Measuring preferences for algorithms—how willing are people to cede control to algorithms?', *Journal of Behavioral and Experimental Economics* **112**, 102270.

Ivanova-Stenzel, R. and Tolksdorf, M. (2025), Delegating in the age of AI: Preferences for decision autonomy, Technical report, CRC TRR 190 Rationality and Competition.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995), *Continuous univariate distributions, volume 2*, Vol. 2, John wiley & sons.

Laske, K., Römer, N. and Schröder, M. (2024), Piece-rate incentives and idea generation – an experimental analysis, CESifo Working Paper No. 11594, CESifo GmbH, Munich.

Liu, C.-W., Yang, M. and Wen, M.-H. (2023), 'Judge me on my losers: Do robo-advisors outperform human investors during the covid-19 financial market crash?', *Production and Operations Management* **32**(10), 3174–3192.

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C. et al. (2019), 'A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis', *The Lancet Digital Health* **1**(6), e271–e297.

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

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

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A. et al. (2020), 'International evaluation of an AI system for breast cancer screening', *Nature* **577**(7788), 89–94.

Möbius, M. M., Niederle, M., Niehaus, P. and Rosenblat, T. S. (2022), 'Managing self-confidence: Theory and experimental evidence', *Management Science* **68**(11), 7793–7817.

Mullainathan, S. and Obermeyer, Z. (2022), 'Diagnosing physician error: A machine learning approach to low-value health care', *The Quarterly Journal of Economics* **137**(2), 679–727.

Normann, H.-T., Rulié, N., Stypa, O. and Werner, T. (2025), 'Delegate pricing decisions to an algorithm? experimental evidence', *arXiv preprint arXiv:2510.27636* .

Rilla, R., Werner, T., Yakura, H., Rahwan, I. and Nussberger, A.-M. (2025), 'Recognising, anticipating, and mitigating LLM pollution of online behavioural research', *arXiv preprint arXiv:2508.01390* . Version 2. **URL:** *https://arxiv.org/abs/2508.01390*

Sanders, N. R. and Manrodt, K. B. (2003), 'The efficacy of using judgmental versus quantitative forecasting methods in practice', *Omega* **31**(6), 511–522.

Soll, J. B. and 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.

Soll, J. B., Palley, A. B. and Rader, C. A. (2022), 'The bad thing about good advice: Understanding when and how advice exacerbates overconfidence', *Management Science* **68**(4), 2949–2969.

Yaniv, I. and Kleinberger, E. (2000), 'Advice taking in decision making: Egocentric discounting and reputation formation', *Organizational Behavior and Human Decision Processes* **83**(2), 260–281.

You, S., Yang, C. L. and Li, X. (2022), 'Algorithmic versus human advice: Does presenting prediction performance matter for algorithm appreciation?', *Journal of Management Information Systems* **39**(2), 336–365.

Zhang, Z., Song, Y. and Qi, H. (2017), Age progression/regression by conditional adversarial autoencoder, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE.

A Appendix

A.1 Pre-Study

In this pre-study, we tested if predicting the median age of groups that are more homogeneous with respect to their age is perceived as harder than predicting the median age of groups that are heterogeneous with respect to their age.[12] We used groups consisting of (i) nine middle-aged people; (ii) seven middle-aged people, one young person, and one old person; (iii) five middle-aged people, two young people, and two old people; and (iv) one middle-aged person, four young people, and four old people. Groups were balanced with respect to either having a male or a female median-aged person. Participants saw these groups of portraits in random order and had to identify the person they perceived as the median-aged person, predict the median age of the group, and indicate the perceived difficulty of the median-age prediction on a 7-point Likert scale. Participants were paid a fixed show-up fee of £2 and could earn an additional £2.5 in the prediction task. We selected one random round and paid the bonus via the binarized scoring rule. We hypothesized that the more homogeneous the group was with respect to age, the more difficult participants would perceive the prediction task to be. We ran the experiment on Prolific in October 2025 with 40 people based in the UK. The experiment took on average 18:28 min (median 16:25 min). Participants earned on average £3.69 including the £2 fixed fee. The average absolute deviation from the true median age was 5.21, which is slightly lower than the average absolute error of 6.75 made by participants in Soll et al. (2022).

We ran a fixed-effects regression of the perceived difficulty on the group's age composition (i.e., the homogeneity of the group). The results of this regression are depicted in Table A.1 Panel A. We find that participants found it significantly harder to predict the median age in groups where the composition was more homogeneous with respect to age. Additionally, we ran a fixed-effects regression of the absolute error of the prediction to check for the true performance. Results of this regression can be found in Table A.1 Panel B. We find that the prediction error decreases with the homogeneity of the group. This means that participants perceived the task as more difficult even though their actual performance increased.

	Panel A: Perceived difficulty	Panel B: Absolute prediction error
Picture Homogeneity	0.30***	-0.72***
	(0.069)	(0.225)
Constant	4.73***	7.02***
	(0.173)	(0.562)
Observations (Subjects)	320 (40)	320 (40)

Standard errors in parentheses. Estimation by fixed-effects regression with clustering at the subject level. Perceived difficulty is measured on a 7-point Likert scale. Picture Homogeneity refers to how many middle-aged people are in the group from which the age prediction is made. It takes the value 1 for one middle-aged person in the group, 2 for five middle-aged people, 3 for seven middle-aged people, and 4 for nine middle-aged people out of nine pictures. *, ** and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.1: Effect of picture homogeneity on perceived difficulty and prediction error.

[12] The pre-study was pre-registered at AsPredicted (https://aspredicted.org/h8em83.pdf) and received ethics approval from the authors' institutional ethics committee.

A.2 Additional figures

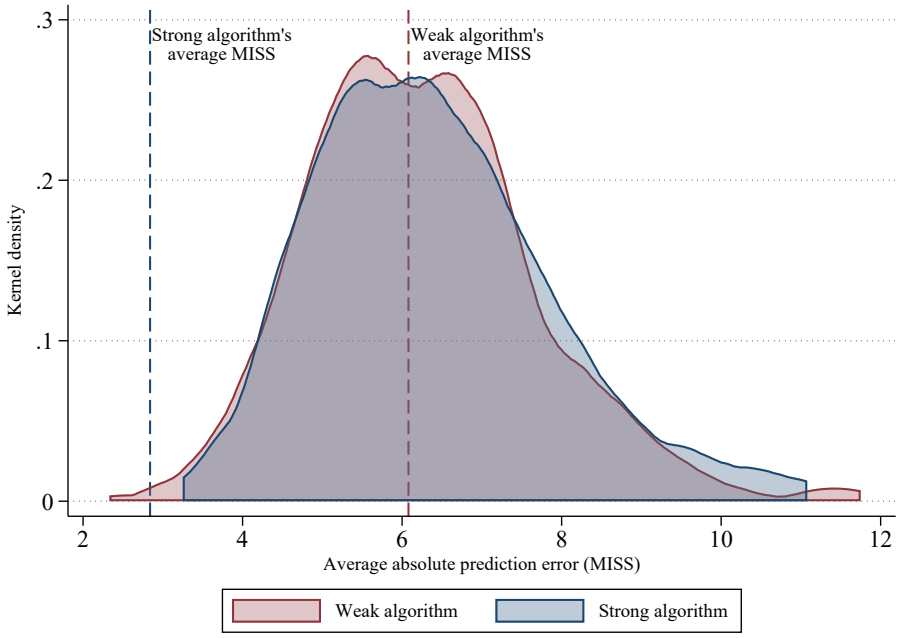


Figure A.1: Distribution of participants' average absolute prediction error (MISS) over the training rounds by algorithm performance.

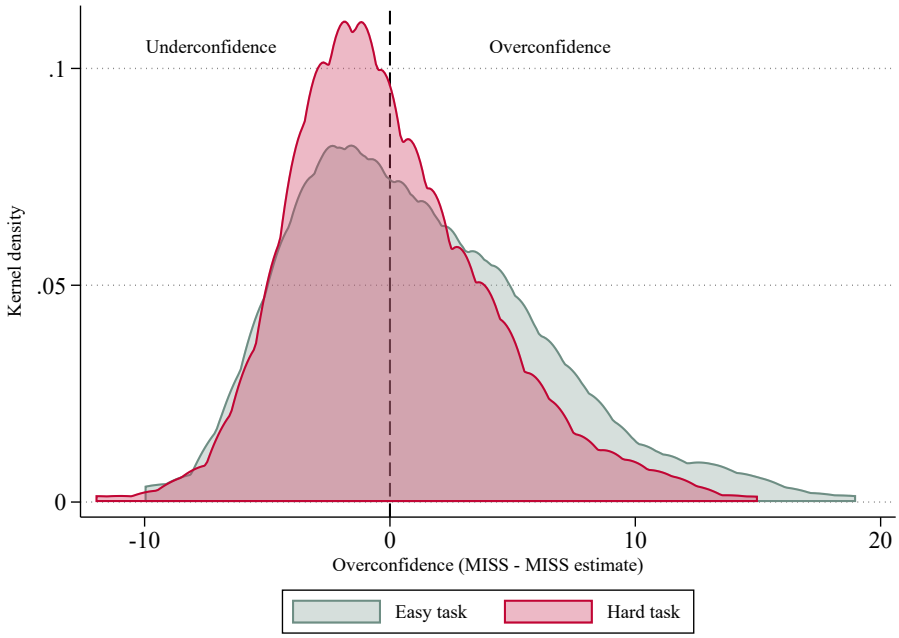


Figure A.2: Distribution of participants' overconfidence (MISS - MISS estimate) in the main task by perceived task difficulty.

A.3 Additional Regressions

	(1) w/o interaction	p-value	(2) with interaction	p-value
Panel A: $\text{WOA}^{Bayes}-\text{WOA}$				
Hard task	0.031	[0.268]	0.028	[0.390]
	(0.028)		(0.032)	
Strong algorithm	0.144***	[0.000]	0.140***	[0.000]
	(0.029)		(0.036)	
Hard task × Strong algorithm			0.007	[0.902]
			(0.057)	
Constant	0.225***	[0.006]	0.227***	[0.007]
	(0.081)		(0.083)	
Controls	✓		✓	
Observations	722		722	
Panel B: $\text{IOA}^{Bayes}-\text{IOA}$				
Hard task	0.036	[0.148]	0.022	[0.469]
	(0.025)		(0.031)	
Strong algorithm	0.075***	[0.003]	0.061*	[0.068]
	(0.025)		(0.033)	
Hard task × Strong algorithm			0.027	[0.585]
			(0.049)	
Constant	0.475***	[0.000]	0.482***	[0.000]
	(0.074)		(0.074)	
Controls	✓		✓	
Observations	749		749	
Panel C: $\text{OS}^{Bayes}-\text{OS}$				
Hard task	-0.006	[0.751]	-0.064**	[0.016]
	(0.018)		(0.026)	
Strong algorithm	-0.029	[0.112]	-0.088***	[0.001]
	(0.018)		(0.026)	
Hard task × Strong algorithm			0.118***	[0.001]
			(0.036)	
Constant	0.229***	[0.000]	0.260***	[0.000]
	(0.057)		(0.057)	
Controls	✓		✓	
Observations	749		749	
Panel D: $\text{CC}^{Bayes}-\text{CC}$				
Hard task	0.041**	[0.024]	0.086***	[0.000]
	(0.018)		(0.020)	
Strong algorithm	0.104***	[0.000]	0.149***	[0.000]
	(0.018)		(0.023)	
Hard task × Strong algorithm			-0.091**	[0.014]
			(0.037)	
Constant	0.245***	[0.000]	0.222***	[0.000]
	(0.052)		(0.051)	
Controls	✓		✓	
Observations	749		749	

Estimation by OLS regression with robust standard errors in parentheses and p-values in square brackets. Controls are gender, age, and performance in training rounds. *, ** and *** denote significance at the 10%, 5%, and 1% levels.

Table A.2: OLS regression on differences between Bayes-optimal and observed advice-taking measures for Weight on Advice (Panel A), Influence of Advice (Panel B), Opinion Shift (Panel C), and Confidence Change (Panel D).

A.4 Robustness Checks

In the main analysis, participants performed better in the perceived hard task than in the easy task, with estimates that were, on average, closer to the true age, as measured by the absolute prediction error (i.e., the absolute deviation between the estimated and true age). Although this performance difference is intentional, since our theoretical framework predicts that perceived difficulty rather than actual performance should influence advice-taking behavior, it remains important to demonstrate that our findings are not merely driven by differences in participants' opportunities to update their beliefs. We therefore conduct additional robustness analyses in which actual performance is either aligned with perceived task difficulty or independent of it.

This adjustment is possible because the observed performance discrepancy is primarily driven by the female version of the easy task. In this treatment arm, participants correctly identified the median individual but provided relatively inaccurate estimates of that individual's age. Excluding this treatment arm results in prediction errors in the hard task that are at least as large as those in the easy task, indicating equal or better performance in the easy task, as shown in Table A.3. Moreover, in two of the three treatment comparisons, predictions are less accurate in the hard task than in the easy task.

	(1) Easy task (male median) vs. Hard task (all)	(2) Easy task (male median) vs. Hard task (male median)	(3) Easy task (male median) vs. Hard task (female median)
Dependent variable: Absolute prediction error			
Hard task	0.939***	0.277	1.594***
	(0.321)	(0.345)	(0.397)
	[0.004]	[0.424]	[0.000]
Constant	4.059***	4.059***	4.059***
	(0.255)	(0.255)	(0.255)
	[0.000]	[0.000]	[0.000]
Observations	560	373	375
R-squared	0.014	0.002	0.042

Estimation by OLS regression with robust standard errors in parentheses and p-values in square brackets. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.3: OLS regressions of pre-advice absolute prediction error on perceived task difficulty across alternative treatment subsamples.

We re-estimate our main specifications after excluding participants who completed the female version of the easy task.[13] In these analyses, the male median easy task sample is compared with (i) the full hard task sample, (ii) the male median hard task sample, and (iii) the female median hard task sample.

As shown in Table A.4, comparing the male median easy task sample with the full hard task sample, the main findings are largely replicated, providing further evidence that perceived difficulty influences advice-taking behavior independently of actual task performance. The same holds for comparisons with the male median hard task sample (see Table A.5) and the female median hard task sample (see Table A.6).

Nevertheless, some minor differences appear. Some coefficients that are significant in the main analysis become insignificant in the robustness checks, possibly due to lower statistical power. Notably, the perceived hard task effect for the influence of advice and confidence change becomes statistically significant in two out of three robustness regressions despite being insignificant in the main analysis.

[13] Note that assignment to the female or male version within each perceived-difficulty condition was randomized.

	(1) w/o interaction	p-value	(2) with interaction	p-value
Panel A: Weight on Advice (WOA)				
Hard task (perceived)	0.002	[0.958]	-0.009	[0.818]
	(0.033)		(0.039)	
Strong algorithm	0.157***	[0.000]	0.143***	[0.005]
	(0.034)		(0.050)	
Hard task × Strong algorithm			0.022	[0.742]
			(0.067)	
Constant	-0.057	[0.563]	-0.049	[0.629]
	(0.099)		(0.101)	
Controls	✓		✓	
Observations	540		540	
Panel B: Influence of Advice (IOA)				
Hard task (perceived)	-0.037	[0.239]	-0.063	[0.107]
	(0.031)		(0.039)	
Strong algorithm	0.137***	[0.000]	0.102**	[0.037]
	(0.030)		(0.049)	
Hard task × Strong algorithm			0.053	[0.388]
			(0.062)	
Constant	-0.025	[0.780]	-0.005	[0.956]
	(0.090)		(0.089)	
Controls	✓		✓	
Observations	560		560	
Panel C: Opinion Shift				
Hard task (perceived)	-0.014	[0.568]	-0.101***	[0.004]
	(0.025)		(0.035)	
Strong algorithm	0.071***	[0.003]	-0.046	[0.241]
	(0.024)		(0.039)	
Hard task × Strong algorithm			0.177***	[0.000]
			(0.049)	
Constant	-0.082	[0.248]	-0.015	[0.837]
	(0.071)		(0.072)	
Controls	✓		✓	
Observations	560		560	
Panel D: Confidence Change				
Hard task (perceived)	-0.022	[0.298]	0.038*	[0.053]
	(0.022)		(0.020)	
Strong algorithm	0.066***	[0.002]	0.148***	[0.000]
	(0.021)		(0.034)	
Hard task × Strong algorithm			-0.123***	[0.004]
			(0.043)	
Constant	0.057	[0.347]	0.010	[0.863]
	(0.060)		(0.057)	
Controls	✓		✓	
Observations	560		560	

Estimation by OLS regression with robust standard errors in parentheses and p-values in square brackets. Controls are gender, age, and performance in training rounds. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.4: OLS regressions of advice-taking measures on perceived task difficulty and algorithm performance: easy task with male median person vs. all hard tasks.

	(1) w/o interaction	p-value	(2) with interaction	p-value
Panel A: Weight on Advice (WOA)				
Hard task	0.006	[0.889]	0.037	[0.440]
	(0.044)		(0.047)	
Strong algorithm	0.115***	[0.009]	0.145***	[0.004]
	(0.044)		(0.050)	
Hard task × Strong algorithm			-0.062	[0.481]
			(0.088)	
Constant	-0.083	[0.502]	-0.100	[0.435]
	(0.123)		(0.128)	
Controls	✓		✓	
Observations	359		359	
Panel B: Influence of Advice (IOA)				
Hard task	-0.047	[0.219]	-0.007	[0.875]
	(0.038)		(0.045)	
Strong algorithm	0.065*	[0.078]	0.105**	[0.032]
	(0.037)		(0.049)	
Hard task × Strong algorithm			-0.080	[0.281]
			(0.075)	
Constant	-0.009	[0.937]	-0.031	[0.777]
	(0.111)		(0.108)	
Controls	✓		✓	
Observations	373		373	
Panel C: Opinion Shift				
Hard task	-0.060**	[0.030]	-0.088**	[0.022]
	(0.027)		(0.038)	
Strong algorithm	-0.013	[0.641]	-0.041	[0.289]
	(0.027)		(0.039)	
Hard task × Strong algorithm			0.058	[0.278]
			(0.054)	
Constant	-0.050	[0.541]	-0.034	[0.679]
	(0.081)		(0.081)	
Controls	✓		✓	
Observations	373		373	
Panel D: Confidence Change				
Hard task	0.013	[0.657]	0.081***	[0.002]
	(0.029)		(0.026)	
Strong algorithm	0.078***	[0.006]	0.147***	[0.000]
	(0.028)		(0.034)	
Hard task × Strong algorithm			-0.139**	[0.017]
			(0.058)	
Constant	0.041	[0.620]	0.003	[0.970]
	(0.082)		(0.078)	
Controls	✓		✓	
Observations	373		373	

Estimation by OLS regression with robust standard errors in parentheses and p-values in square brackets. Controls are gender, age, and performance in training rounds. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.5: OLS regressions of advice-taking measures on perceived task difficulty and algorithm performance: easy task with male median person vs. hard task with male median person.

	(1) w/o interaction	p-value	(2) with interaction	p-value
Panel A: Weight on Advice (WOA)				
Hard task	-0.004	[0.901]	-0.050	[0.254]
	(0.035)		(0.043)	
Strong algorithm	0.191***	[0.000]	0.145***	[0.004]
	(0.036)		(0.050)	
Hard task × Strong algorithm			0.093	[0.191]
			(0.071)	
Constant	-0.186*	[0.091]	-0.155	[0.152]
	(0.110)		(0.108)	
Controls	✓		✓	
Observations	365		365	
Panel B: Influence of Advice (IOA)				
Hard task	-0.028	[0.428]	-0.117***	[0.009]
	(0.035)		(0.045)	
Strong algorithm	0.191***	[0.000]	0.100**	[0.039]
	(0.035)		(0.048)	
Hard task × Strong algorithm			0.181***	[0.009]
			(0.069)	
Constant	-0.170	[0.106]	-0.110	[0.295]
	(0.105)		(0.105)	
Controls	✓		✓	
Observations	375		375	
Panel C: Opinion Shift				
Hard task	0.028	[0.347]	-0.113***	[0.004]
	(0.030)		(0.039)	
Strong algorithm	0.098***	[0.001]	-0.046	[0.245]
	(0.030)		(0.039)	
Hard task × Strong algorithm			0.287***	[0.000]
			(0.058)	
Constant	-0.144*	[0.091]	-0.049	[0.571]
	(0.085)		(0.086)	
Controls	✓		✓	
Observations	375		375	
Panel D: Confidence Change				
Hard task	-0.056***	[0.009]	-0.004	[0.881]
	(0.021)		(0.024)	
Strong algorithm	0.093***	[0.000]	0.146***	[0.000]
	(0.022)		(0.033)	
Hard task × Strong algorithm			-0.106**	[0.013]
			(0.043)	
Constant	-0.026	[0.682]	-0.062	[0.331]
	(0.064)		(0.063)	
Controls	✓		✓	
Observations	375		375	

Estimation by OLS regression with robust standard errors in parentheses and p-values in square brackets. Controls are gender, age, and performance in training rounds. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.6: OLS regressions of advice-taking measures on perceived task difficulty and algorithm performance: easy task with male median person vs. hard task with female median person.

A.5 Experimental Materials

A.5.1 Instructions

Thank you very much for participating in this study.

In this study, you will be shown a series of images and asked to estimate the ages of the people depicted.
Your main task is to determine the **median** age of a group, and you will be paid based on your accuracy. Before making your final prediction, you will see the median age prediction of an algorithm.
The **median** age is the age right in the middle once all the reported ages of the people shown are sorted from youngest to oldest. In other words, at the **median** age, 50% of the people are younger and 50% are older. For example: If the people in an image have the following ages: 50, 10, 30, 100, and 60, the sorted list is: 10, 30, 50, 60, 100. Thus, the **median** age would be 50.
In addition to making your best prediction of the **median** age, you will also be asked to estimate how accurate your prediction is. The accuracy is simply the distance to the truth, which we call the **MISS**. The **MISS** represents the amount by which your guess missed the truth. For example, if the true answer is 1000 the answers 1005 and 995 both have a miss of five.

For example, a person named Robin is estimating the age of another person Terry. Suppose Terry is 60 years old, but Robin does not know this. If Robin guesses 35, then Robin's **MISS** is 25. If Robin guesses 110, then the **MISS** is 50. The lower the **MISS**, the more accurate the guess. Remember, the **MISS** is the amount by which you missed the correct answer, either too high or too low. If you guess exactly right, the **MISS** is zero.

Flow of the Study:
To familiarize yourself with the task, you will complete **12** training rounds in which you estimate a person's age as well as your associated **MISS**.
After each training round, you'll see the algorithm's **MISS** for that photo, i.e., how many years the algorithm's prediction differed from the true age.
In the main task, you will see a group of **9** people. We will ask you to identify the person that you think has the **median** age of the group, the **median** age of the group as well as your **MISS**. Then you will see the median age the algorithm predicted. You are then allowed to change your predictions. Importantly, your payment for this task depends on your final, and thus potentially updated, predictions!

Payment:

In the training rounds, you have the chance to earn a bonus of £0.50 depending on the accuracy of your predictions. We will randomly pick one training round and pay you a bonus for either your age prediction or your **MISS** prediction.

In the main round, you have the chance to earn a bonus of £2 depending on the accuracy of your predictions. Here, it will be randomly determined whether your **median** age prediction or your **MISS** prediction is relevant for payment.

Payment calculation:

The closer your age predictions are to the true **median** age, the greater your chances of earning the bonus. Therefore, to maximize your earnings, you should try to be as **accurate** as possible in all your responses.

Similarly, the closer your reported **MISS** is to the true **MISS**, the greater your chances of earning the bonus. Therefore, it is again in your interest to make the most **accurate prediction** possible.

If you would like to know exactly how your payment is calculated, click here.

A.5.2 Screenshots from the experiment

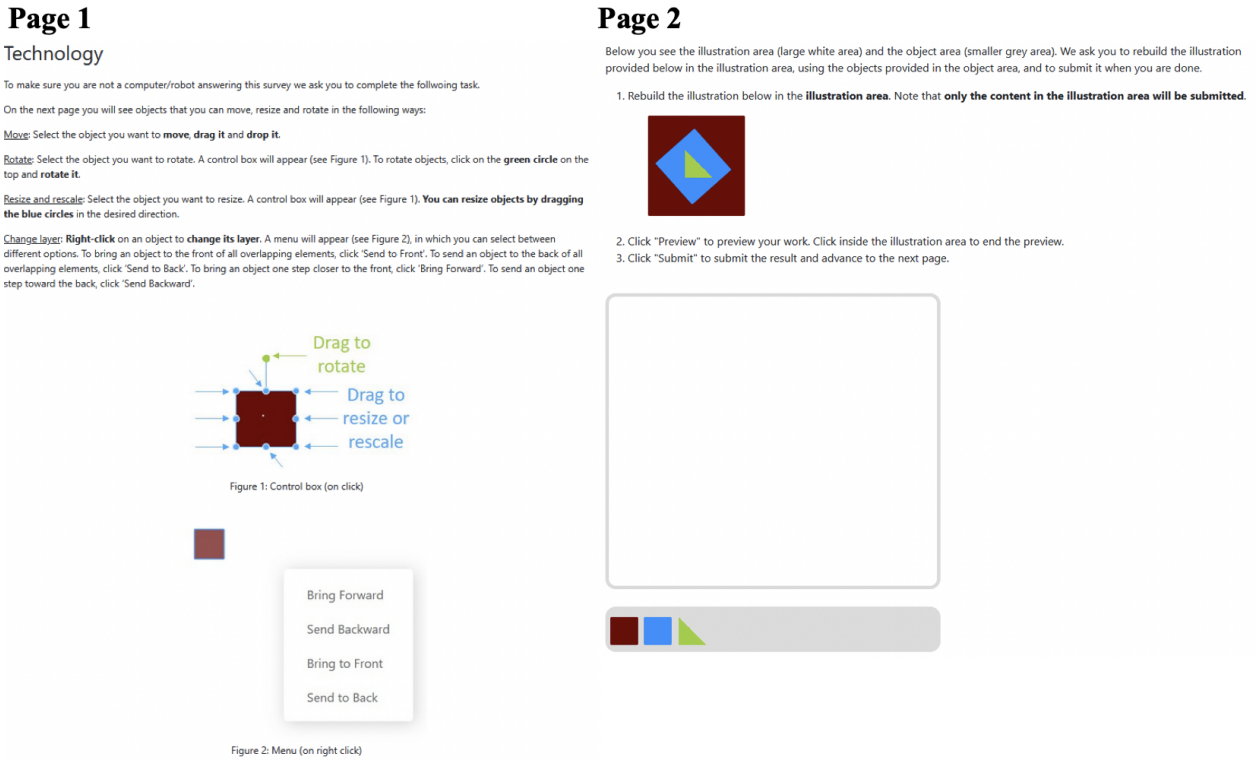


Figure A.3: Screenshots from the picture assembly task

A.5.3 Control questions

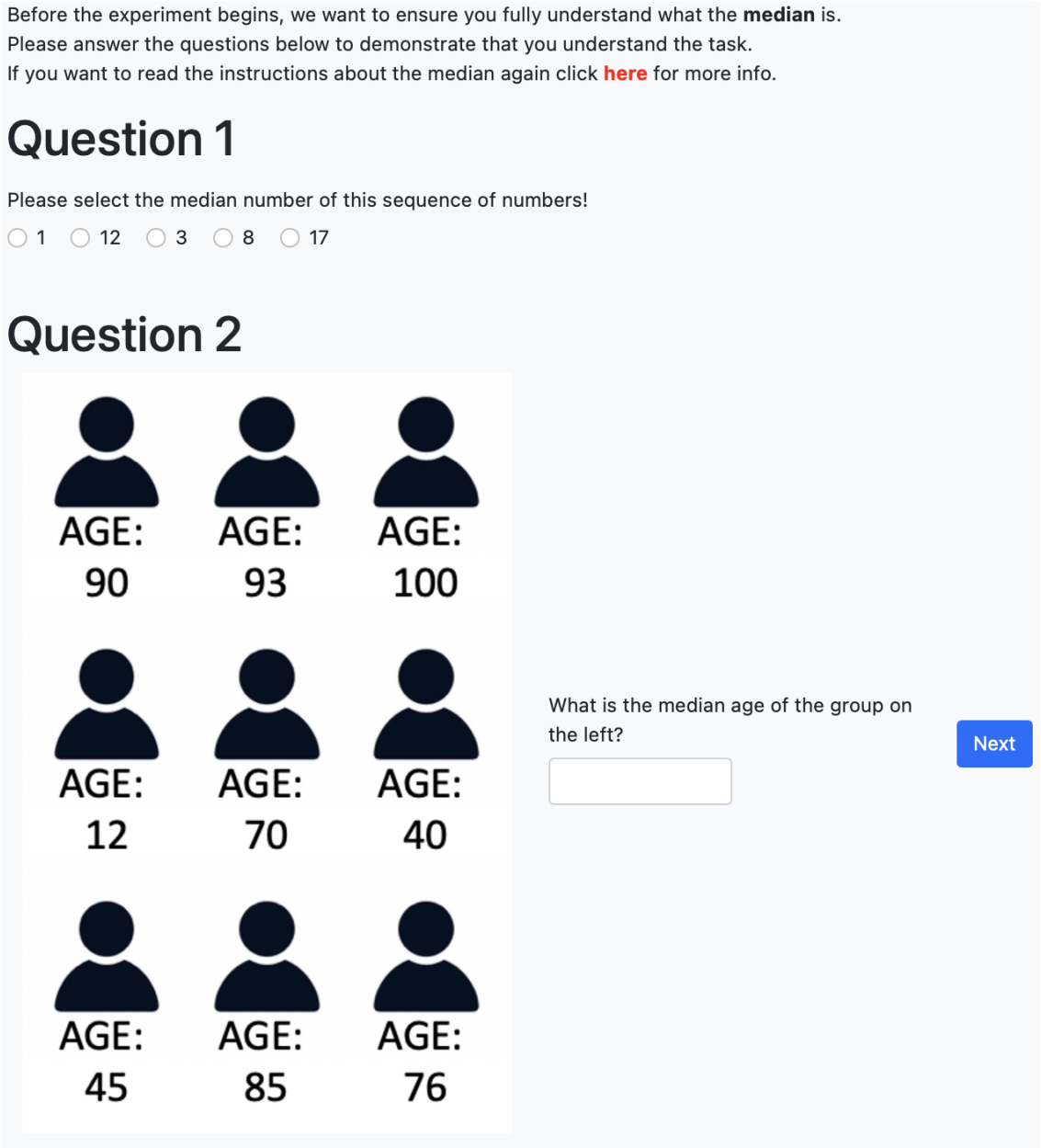


Figure A.4: Screenshot of the median age comprehension question.

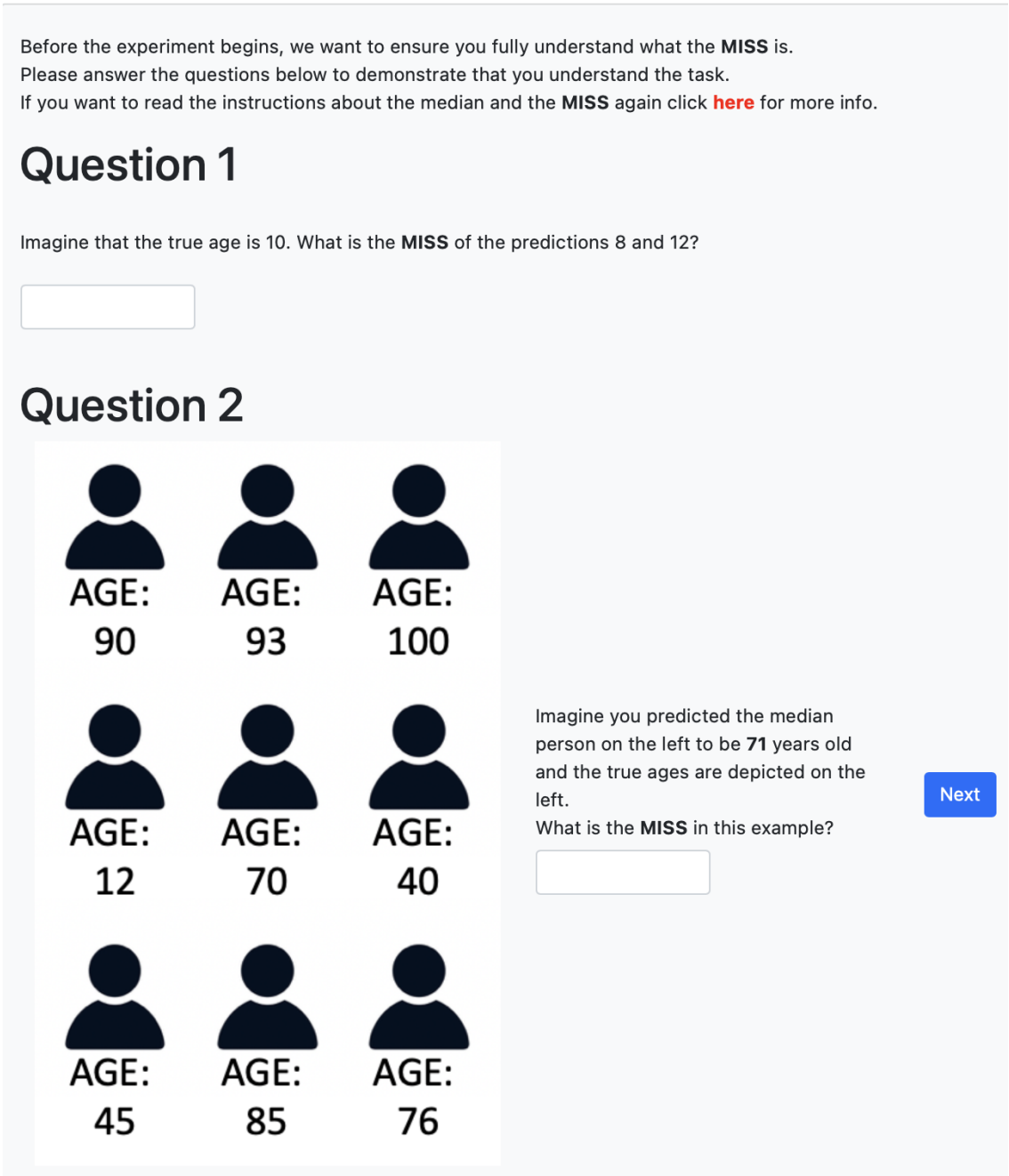


Figure A.5: Screenshot of the median and MISS comprehension question.

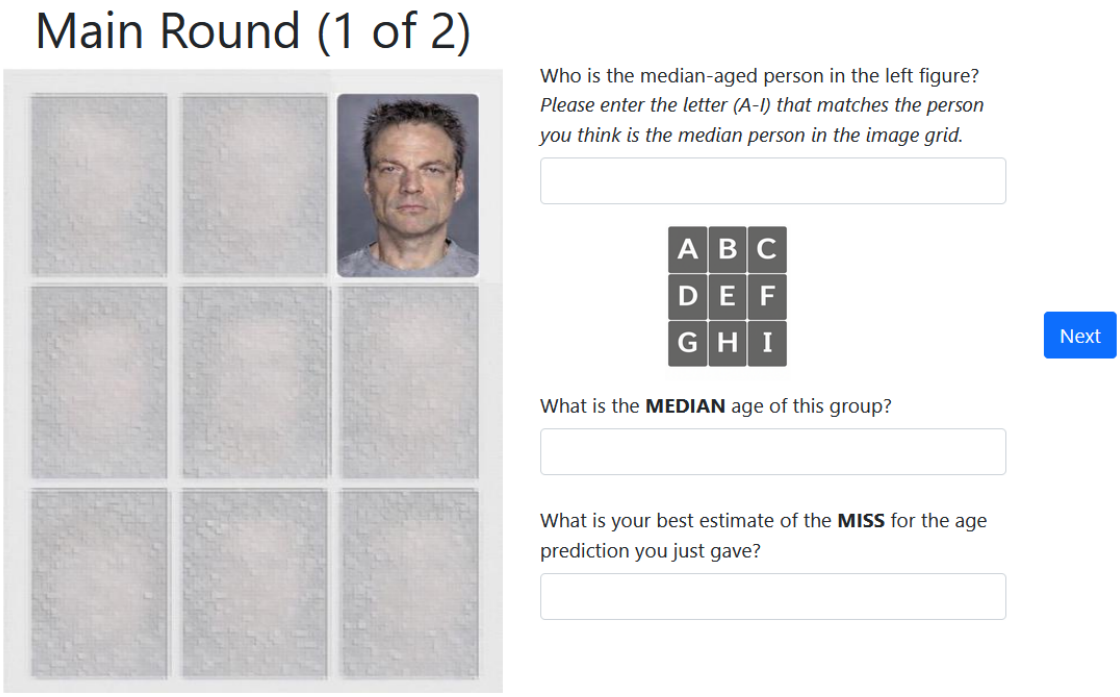


Figure A.6: Screenshot of the main task.

Notes: Due to the FACES release agreement we can only show pictures of the freely available pictures on https://faces.mpdl.mpg.de/imeji/. The other eight pictures are thus pixelated here but were visible to the participants in the experiment.

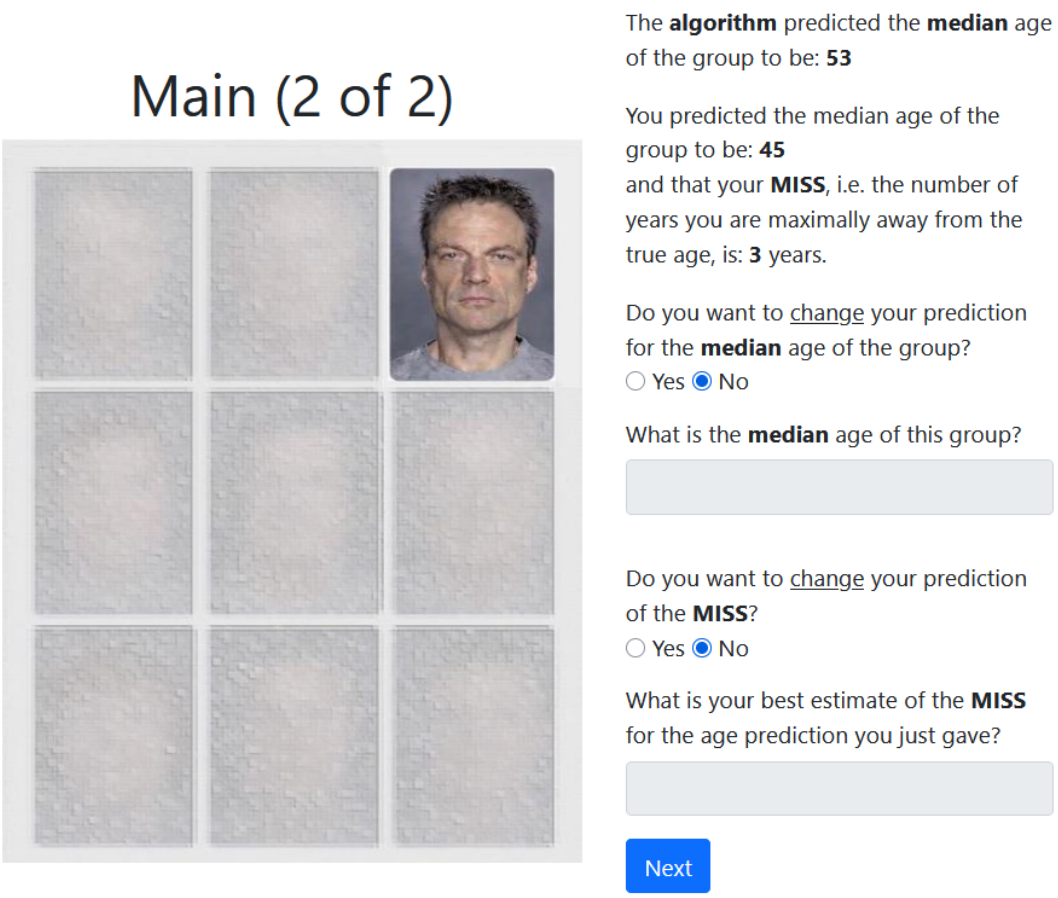


Figure A.7: Screenshot of the main task.

Notes: Due to the FACES release agreement we can only show pictures of the freely available pictures on https://faces.mpdl.mpg.de/imeji/. The other eight pictures are thus pixelated here but were visible to the participants in the experiment.