Groupe de recherche en Droit, Economie et Gestion
Gredeg

UNIVERSITÉ CÔTE D'AZUR

WHEN DOES ADVISOR CONFIDENCE IMPROVE DECISIONS? EVIDENCE FROM HUMAN AND ALGORITHMIC ADVICE

Documents de travail GREDEG
GREDEG Working Papers Series

Mathieu Chevrier
Sébastien Massoni

GREDEG WP No. 2026-09
https://ideas.repec.org/s/gre/wpaper.html

When Does Advisor Confidence Improve Decisions? Evidence from Human and Algorithmic Advice

Mathieu Chevrier*& Sébastien Massoni†

GREDEG Working Paper

Abstract

Confidence often accompanies advice, but its usefulness depends on what confidence actually reveals. This paper distinguishes between two dimensions of confidence quality: discrimination, that is, whether confidence tracks correctness at the decision level, and calibration, that is, whether average confidence matches average accuracy. In a controlled advice-taking experiment comparing human and algorithmic advisors, discrimination is the main driver of both advice adoption and post-advice accuracy, whereas calibration plays a more limited role. Source matters only in a specific case: when discrimination is high, participants are more likely to follow overconfident algorithmic advice than equally overconfident human advice. Advice taking also varies with participants' own metacognitive characteristics. Higher discrimination ability is associated with more conservative advice taking, while better-calibrated participants rely more on stated confidence, benefiting when advisor confidence has high discrimination and performing worse when it is miscalibrated.

JEL Codes: C92; D91.

This research was funded by the program FUTURE LEADER within the Program Investissements d'Avenir "Lorraine Université d'Excellence" (ANR-15-IDEX-04-LUE) operated by the French National Research Agency (ANR). We also thank participants for helpful comments on earlier versions of this paper during presentations at the University of Lille, the University of Besançon, and the University of Nice, as well as at FUR 2024 in Brisbane, the GREDEG PhD Workshop in Nice, and the ASFEE Winter School in Aussois (2023). We are particularly grateful to Brice Corgnet and Paul Pezanis-Christou for their careful reading and valuable feedback.

*Cote D'Azur University, CNRS, GREDEG, France. Email: mathieu.chevrier@univ-cotedazur.fr

†Université de Lorraine, Université de Strasbourg, CNRS, BETA, Nancy, France. Email: sebastien.massoni@gmail.com

Keywords: Algorithm, Advice, Overconfidence, Discrimination, Laboratory experiment

1 Introduction

Algorithmic advisors increasingly accompany their recommendations with a confidence score, a numerical signal intended to help decision makers judge when to follow the advice and when to rely on their own judgment. Such confidence signals are becoming common in domains ranging from medical diagnosis and risk assessment to hiring and integrity screening (Delmas et al., 2024; Dargnies et al., 2024). Yet confidence may be perceived differently depending on its source. Algorithmic confidence is presented as a statistical output derived from models and data, whereas human confidence reflects metacognitive self-assessment, that is, an evaluation of the reliability of one's own judgment (Fleming and Lau, 2014; Fleming, 2024). This raises a first question: *do decision makers respond differently to identical advice and stated confidence depending on whether it comes from a human or an algorithmic advisor?*

Confidence can also be misleading. A confidence score may be biased or uninformative, just as a human advisor's self-assessed certainty may be inflated or poorly calibrated (Offert and Bell, 2021; Cao et al., 2023). When confidence is informative, it can improve joint performance and decision quality (Bahrami et al., 2010; Massoni and Roux, 2017). We focus on two distinct dimensions of confidence quality. Calibration captures average bias: whether mean confidence matches mean accuracy. Discrimination quality captures whether confidence is higher when the advisor is correct than when the advisor is incorrect (Fleming and Lau, 2014; Fleming, 2024). These two dimensions are conceptually distinct. For example, a physician asked to predict the sex of a fetus who always gives the same answer and always reports 50% confidence would be well calibrated on average, yet confidence would have zero discrimination because it does not vary across cases. By contrast, a physician who reports higher confidence when correct than when incorrect would exhibit good discrimination even if average confidence might be too high. This leads to our second question: *which dimension of confidence quality matters more for advice adoption and decision accuracy, and how do decision makers' own metacognitive abilities interact with the advisor's confidence profile?*

To address these questions, we conducted a laboratory experiment with 307 participants who performed 240 rounds of an incentivized perceptual "dot task" (Hainguerlot et al., 2023; Massoni and Roux, 2017; Massoni et al., 2014). Perceptual tasks are standard for metacognitive measurement because they allow clean control over task difficulty

and confidence elicitation (Fleming and Lau, 2014; Fleming, 2024). In each round, two circles briefly appeared on screen and participants decided which contained more dots (the "left–right" decision). Participants reported their confidence, received advice from an advisor, and made a final decision. Both choices and confidence reports were incentivized using a matching probability mechanism (Winkler and Murphy, 1968; Holt and Smith, 2009). We implemented a Judge–Advisor System (Sniezek and Buckley, 1995) in which advice is provided after the initial judgment and before the final judgment.

We implemented a 2×3 between-subjects design varying advisor type (human vs. algorithmic) and confidence condition (no confidence vs. well-calibrated vs. overconfident). Within subjects, we independently manipulate the advisor's discrimination quality across two 120-round blocks (high vs. low), with the order randomized. Crucially, the advisor's average accuracy is held constant at 75% across all conditions, so differences in advice adoption or post-advice accuracy can be attributed to the properties of the confidence signal rather than to differences in average advisor accuracy.

Three main findings emerge. First, discrimination quality is the primary driver of both advice adoption and post-advice accuracy, whereas calibration plays a more limited role. When discrimination is low, even overconfident advice does not outperform the no-confidence baseline in terms of adoption. Second, source matters only in one case: participants are more likely to follow overconfident algorithmic advice than equally overconfident human advice when discrimination quality is high. Third, decision makers' own metacognitive characteristics are associated with distinct patterns of advice use. Better-calibrated participants rely more on stated confidence, which helps them when confidence is informative but hurts them when it is biased. Participants with higher discrimination ability adopt advice more conservatively, which protects them from misleading signals but also limits gains from high-quality advice.

These results contribute to several literatures. First, they contribute to the literature on advice and decision making (Yates et al., 1996; Price and Stone, 2004; Stanciu and Fiser, 2022) by disentangling calibration from discrimination and showing that discrimination quality is the main determinant of when decision makers follow advice and whether they benefit from it. Second, they contribute to the literature on human–algorithm interaction (Mahmud et al., 2022; Chugunova and Sele, 2022; Jussupow et al., 2020) by showing that responses to algorithmic confidence depend on the quality of the signal,

rather than on source alone. Third, they contribute to the literature on heterogeneity in responses to algorithmic advice (Caplin et al., 2025; Brynjolfsson et al., 2023; Noy and Zhang, 2023) by showing that calibration and discrimination operate through distinct behavioral patterns, and that users' own metacognitive characteristics shape whether they benefit from confidence signals.

The paper proceeds as follows. Section 2 reviews the related literature. Section 3 states our hypotheses. Section 4 describes the experimental design. Section 5 presents the results. Section 6 discusses the findings and concludes.

2 Related Literature

Our contribution lies at the intersection of three literatures: confidence quality in advice taking, human–algorithm interaction, and heterogeneity in responses to algorithmic advice.

2.1 Confidence quality in advice taking

Research shows that DMs rely more on an advisor—whether human or algorithmic—when a confidence level is provided (Gaertig and Simmons, 2023; Zhang et al., 2020; Taudien et al., 2022), and that DMs prefer advisors who express high confidence, even when this confidence is inflated (Sniezek and Van Swol, 2001; Phillips, 1999; Sniezek and Buckley, 1995; Yaniv, 1997). Stanciu and Fiser (2022) show that this preference persists even when DMs are explicitly informed of the advisor's overconfidence.

However, these studies focus almost exclusively on calibration and say little about discrimination quality. To the best of our knowledge, only Yates et al. (1996) and Price and Stone (2004) address this dimension. Yates et al. (1996) find that 78% of participants preferred a forecaster who was overconfident but highly discriminating over one who was well calibrated but weakly discriminating. Price and Stone (2004) hold discrimination quality and accuracy constant and show that increasing confidence alone raises advice adoption, which they interpret as a "confidence heuristic" whereby decision makers treat confidence as a proxy for knowledgeability.

Crucially, neither study cleanly disentangles calibration from discrimination: in both designs the two dimensions co-vary. Our experiment fills this gap by independently manipulating calibration (well-calibrated vs. overconfident) and discrimination quality (high

vs. low) while holding accuracy constant. This design allows us to test whether discrimination quality or calibration is the primary driver of advice adoption and post-advice accuracy.

2.2 Algorithm and human advisors

Perceived advisor competence is a key determinant of advice taking (Bailey et al., 2023; Chevrier et al., 2024; Bonaccio and Dalal, 2006). People tend to rely more on algorithmic advice when the task is perceived as objective or quantitative (Castelo et al., 2019). In perceptual tasks, recent work reports greater acceptance of algorithms in image-based detection tasks (Biermann et al., 2022; Logg et al., 2019), though algorithmic preferences can disappear when DMs have access to a human expert (Logg et al., 2019).

Prior work also shows that disclosing an algorithm's confidence level can increase advice taking (Zhang et al., 2020; Taudien et al., 2022), and that reliance on algorithmic advice depends on DMs' initial confidence (Chong et al., 2022; Snijders et al., 2023). However, existing studies do not directly compare algorithm and human advisors who provide identical confidence information while holding the informational content of the advice constant.

Our experiment addresses this gap by comparing human and algorithmic advisors who provide identical recommendations and identical stated confidence levels. This design allows differences in advice adoption to be traced to the perceived source rather than to differences in the informational content of the advice. Because algorithmic and human confidence are perceived as arising from different processes, decision makers may respond differently to the same confidence signal depending on its source.

2.3 Heterogeneity in responses to algorithmic advice

A growing body of evidence suggests that Artificial Intelligence (AI) tools raise average productivity while compressing the performance distribution, disproportionately benefiting initially lower-performing individuals (Brynjolfsson et al., 2023; Noy and Zhang, 2023). Caplin et al. (2025) show that AI can be equalizing, but only if users know when they are likely to be wrong and therefore when deferring to the tool is beneficial. Their analysis highlights calibration as the key metacognitive trait for benefiting from AI.

However, we still know little about how DMs' own discrimination quality, captured by

meta-d' (Maniscalco and Lau, 2012, 2014), shapes interactions with an advisor. Because calibration and discrimination rely on distinct metacognitive processes (Fleming and Lau, 2014; Fleming, 2024), they may generate different advice-taking strategies and different gains from AI. Our study addresses this gap by varying the advisor's calibration and discrimination quality, comparing algorithmic and human advisors, and examining how decision makers' own calibration and discrimination shape advice use and accuracy.

3 Hypotheses

We derive four hypotheses organized around two themes: how advisor confidence quality shapes advice adoption and post-advice accuracy (Hypotheses 1–3), and how decision makers' own metacognitive abilities shape their interaction with the advisor (Hypotheses 4a–4b).

3.1 Discrimination quality as the primary driver of advice adoption

Calibration indicates whether average confidence matches average accuracy, but it does not signal, at the trial level, when advice is likely to be correct. Discrimination quality, by contrast, provides a decision-level cue. When discrimination is high, confidence is systematically higher on correct trials than on incorrect ones, giving DMs a reliable signal about when to follow the advice.

Prior evidence is consistent with this reasoning. Yates et al. (1996) and Price and Stone (2004) suggest that DMs prefer advisors whose confidence is informative, though their designs do not fully disentangle calibration from discrimination.

> **Hypothesis 1** *(Discrimination quality & advice adoption).* Participants are more likely to follow the advisor's recommendation when the advisor's confidence is highly discriminating, regardless of whether the advisor is well calibrated or overconfident.

3.2 Algorithmic vs. human confidence

Human confidence is produced through metacognition—a subjective and potentially biased assessment of one's own judgment. Algorithmic confidence, by contrast, is often perceived as a statistical quantity derived from data and models. This difference in

source may affect how decision makers interpret confidence, particularly when the signal is salient. Moreover, overconfident advisors report higher stated confidence, which may increase the salience of the confidence cue.

> **Hypothesis 2** *(Algorithm vs. Human).* Participants are more likely to follow the advisor's recommendation when the advisor is an algorithm rather than a human, especially when the advisor is overconfident.

3.3 Discrimination quality and DM accuracy

The benefit of advice for decision accuracy depends on whether the confidence signal helps decision makers identify when to follow the advisor's recommendation. When the advisor is well calibrated and has high discrimination quality, confidence provides an informative cue. When confidence is overconfident or weakly discriminating, decision makers cannot reliably distinguish correct from incorrect advice.

> **Hypothesis 3** *(Discrimination quality & Accuracy).* Post-advice accuracy is highest when advice comes from a well-calibrated advisor with high discrimination quality, and lowest when the advisor's confidence has low discrimination quality.

3.4 DM metacognitive ability, advice adoption, and accuracy

Caplin et al. (2025) show that well-calibrated DMs benefit more from AI advice. However, no study has examined how DMs' own discrimination quality shapes their interaction with an advisor. Because calibration captures aggregate confidence bias whereas discrimination captures the trial-level informativeness of one's own confidence, we expect these two metacognitive dimensions to be associated with different advice-taking strategies.

> **Hypothesis 4a** *(DM calibration).* Well-calibrated DMs benefit more from an advisor who provides a confidence level, provided that the signal is sufficiently informative.

> **Hypothesis 4b** *(DM discrimination).* Decision makers with high discrimination quality adopt advice differently from well-calibrated decision makers, leading to distinct patterns of advice adoption and accuracy.

4 Experimental Design

Participants perform an incentivized perceptual task in which they first indicate whether the left or the right circle contains more white dots (the "left–right" decision) and then report a confidence level interpreted as their subjective probability of being correct. After these initial decisions, participants receive advice and are invited to revise both their "left–right" decision and their confidence level.

We implement a 2×3 between-subjects design combined with a within-subject manipulation. Between subjects, we vary the advisor type (Human vs. Algorithm) and the confidence information provided with advice (Without confidence vs. Calibrated vs. Overconfident). Within subjects, we vary the advisor's discrimination quality across two 120-round blocks (high vs. low discrimination), with the order counterbalanced (HL vs. LH). This within-subject manipulation increases statistical power, while the counterbalanced order helps separate discrimination effects from simple block order effects. Throughout, we hold the advisor's average accuracy constant and ensure that participants assigned to the same sequence receive identical left–right recommendations round by round. This allows us to isolate the effect of confidence properties and advisor source from differences in recommendation content.

4.1 Advisor Confidence: Calibration & Discrimination

We characterize the quality of a confidence signal along two dimensions. *Calibration* captures whether average confidence matches average accuracy. A calibrated advisor with 75% accuracy reports an average confidence of 75%, whereas an overconfident advisor with the same accuracy reports an average confidence of 85%. *Discrimination quality* (metacognitive sensitivity) captures whether confidence is higher when the advisor is correct than when incorrect.

We measure discrimination quality using the Area Under the ROC Curve (AUC), defined as the probability that confidence on a randomly chosen correct trial exceeds confidence on a randomly chosen incorrect trial (with ties receiving half weight). An AUC of 0.50 indicates no discrimination (uninformative confidence), whereas an AUC of 1 indicates perfect discrimination. In our implementation, as shown in Table 1, high-discrimination advisors have AUCs around 0.83 (approximately 0.81–0.85), whereas low-

discrimination advisors have AUCs around 0.52 (approximately 0.51–0.53).[1]

Table 1: Between- and Within-Subjects Design with Advisor Average Confidence Level (CL) and Area Under The ROC Curve (AUC).

			Without		Calibrated		Overconfident	
			1–120	121–240	1–120	121–240	1–120	121–240
Algorithm	Sequence HL	Average CL	–	–	75%	75%	85%	85%
		AUC	–	–	0.85	0.51	0.81	0.53
	Sequence LH	Average CL	–	–	75%	75%	85%	85%
		AUC	–	–	0.51	0.85	0.53	0.81
Human	Sequence HL	Average CL	–	–	75%	75%	85%	85%
		AUC	–	–	0.85	0.51	0.81	0.53
	Sequence LH	Average CL	–	–	75%	75%	85%	85%
		AUC	–	–	0.51	0.85	0.53	0.81

Note. This table summarizes the 2 × 3 between-subjects design (advisor type: Algorithm vs. Human × confidence condition: Without, Calibrated, Overconfident) and the within-subject manipulation of discrimination quality across two 120-round blocks. In the Without condition, advice consists only of the left–right recommendation (no confidence is displayed, indicated by "–"). In the Calibrated and Overconfident conditions, the advisor additionally displays a confidence level (CL) whose mean is held constant within condition: 75% in Calibrated and 85% in Overconfident. Because the advisor's average success rate is fixed at 75%, the implied calibration bias (mean confidence minus accuracy) is 0 in Calibrated and +10 percentage points in Overconfident. Discrimination quality is manipulated within subjects: HL (LH) denotes that the advisor's confidence is highly (weakly) discriminating in rounds 1–120 and weakly (highly) discriminating in rounds 121–240. Discrimination is quantified by the area under the ROC curve (AUC).

4.2 Perceptual task and decision timeline

The perceptual task follows Massoni et al. (2014) and Massoni and Roux (2017). In each round, participants observe two black circles containing white dots and indicate which circle contains more dots (see Appendix B). The task consists of 60 distinct images. In every image, one circle contains 50 dots and the other contains 47[2]. The 60-image sequence is repeated four times, so each participant completes 240 rounds. This number of repetitions ensures that each participant interacts at least 120 times with a given advisor profile, which is required for estimating participants' metacognitive measures.

Figure 1 summarizes the decisions timeline. In Step 1, a target screen displaying the two empty circles appears for 1 second. In Step 2, the dot stimulus is displayed for 750 ms. In Step 3, participants make the left–right decision. In Step 4, participants report

[1] In the results section, we measure participants' own discrimination quality using meta-d', a signal-detection-based measure of metacognitive sensitivity (Maniscalco and Lau, 2012, 2014). We rely on meta-d' to analyze experimental data because it helps disentangle metacognitive sensitivity from response bias and idiosyncratic use of the confidence scale (i.e., differences in criteria across individuals). By construction, these biases are absent for our programmed advisors.

[2] The 3-dot difference was calibrated in a preliminary study using a two-up-one-down staircase procedure (Levitt, 1971), yielding an average baseline accuracy of about 69%.

confidence by selecting one of 11 values: 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, or 100%. In Step 5, participants receive advice and are asked to revise both their left–right decision and their confidence. Importantly, the stimulus is not displayed again at this stage. Participants only see their own initial decision and confidence, together with the advisor's recommendation and confidence level (depending on the treatment). This feature ensures that revisions reflect advice use rather than re-examination of the stimulus. In Step 6, participants receive feedback on whether their final decision was correct, as well as the advisor's correctness for that round.

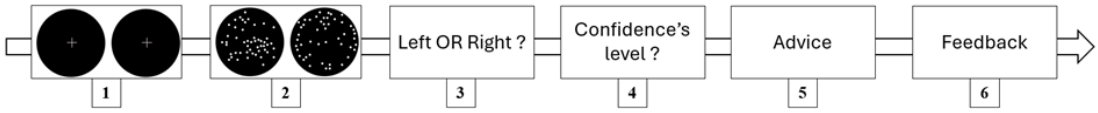


Figure 1: Decision timeline within a round.

Note. Each participant completes 240 rounds. Step 1: target screen (1s). Step 2: stimulus display (750 ms). Step 3: initial left–right decision. Step 4: initial confidence report. Step 5: advice is revealed and participants provide a final decision and confidence. Step 6: feedback on participant and advisor correctness.

4.3 Confidence Elicitation

Confidence reports are incentivized using the Matching Probability (MP) method (Winkler and Murphy, 1968; Holt and Smith, 2009). After each decision, the participant selects a confidence level $p \in \{50, 55, \ldots, 100\}$ interpreted as the subjective probability that the left–right decision is correct. We use the sequential draws as implemented in Hollard et al. (2016). Two random numbers are drawn independently and uniformly from $\{1, \ldots, 100\}$, denoted L_1 and L_2. If $L_1 \leq p$, the participant is paid if and only if the decision is correct. If $L_1 > p$, the participant is paid if and only if $L_2 \leq L_1$. Under this mechanism, expected earnings are maximized by reporting true subjective confidence (see Karni (2009) for the theoretical and Burfurd and Wilkening (2018) for the behavioral properties of the mechanism).

4.4 Advisor construction

A key requirement is to vary calibration and discrimination without deceiving participants, while holding average advisor accuracy constant. We therefore constructed advisors from the Experimental Economics Laboratory of Nice (LEEN) pre-study in which 11 participants completed 480 dot-task rounds (60 images repeated 8 times), yielding 8

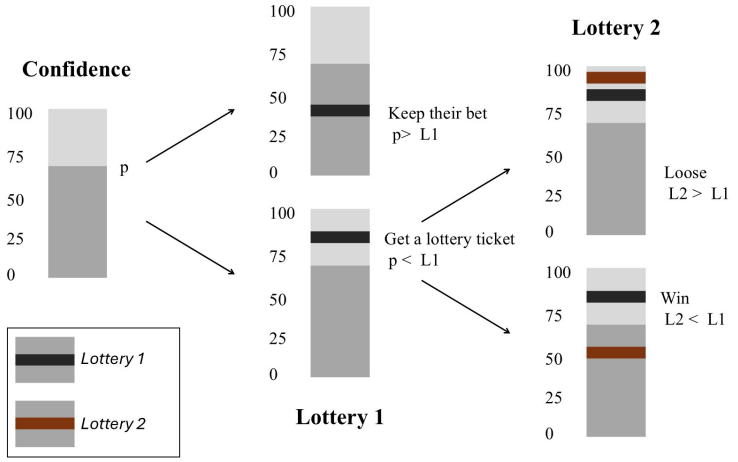


Figure 2: Matching Probability mechanism to elicit confidence.

Note. The participant selects confidence $p \in [50, 100]$. A first lottery L_1 is drawn. If $L_1 \leq p$, payment depends only on correctness. If $L_1 > p$, a second lottery L_2 is drawn and payment is obtained if $L_2 \leq L_1$.

observations per image.

From this pre-study, we selected four participants with high variability in confidence reports. For each selected participant and each image, we retained one of the eight available observations to construct an advisor profile satisfying four constraints: (i) average advisor accuracy equals 75%; (ii) average confidence equals 75% in the calibrated condition and 85% in the overconfident condition; (iii) discrimination quality matches the targeted AUC level (high vs. low); and (iv) for a given image sequence, the left–right recommendation is identical across the Without, Calibrated, and Overconfident conditions. This procedure allows us to vary calibration and discrimination while keeping average accuracy fixed and recommendation content comparable across confidence conditions. Details on the resulting advice sequences are reported in Appendix A.

4.5 Treatment conditions

4.5.1 Advisor type (between subjects)

Participants are randomly assigned to a Human or Algorithm advisor condition. In both cases, we truthfully state that advice originates from decisions made by previous participants under the same task and incentives. Specifically, for both types of advisors, participants are informed that: "*During a previous experimental session, other participants were asked to participate in the same experimental task that you are about to do,*

with the same monetary incentives. They were aware that their choices would aim to serve as advice for you." In the Human condition, we add that: "*One of these participants will be your advisor, then after 120 rounds, another participant will be assigned to you until the end of the task.*" In the Algorithm condition, we add that: "*The algorithm uses a database including the participants who played during that session. It selects certain decisions and confidence levels made by these participants. After 120 rounds, a new algorithm will be assigned to you until the end of the task.*" In both conditions, we further state: "*These two other participant [algorithmic] advisors have achieved results that are above or equal to the average of participants from the previous session.*"

This framing is designed to isolate source effects while minimizing ex ante differences in perceived competence. First, we use the generic term "algorithm" rather than "AI," since "AI" framing can increase reliance independently of content (Candrian and Scherer, 2024; Langer et al., 2022; Raux and Dreyfuss, 2025). Second, we describe the algorithm as selecting prior decisions rather than performing complex computations, so as not to confound source with perceived complexity or competence (Lehmann et al., 2022; Chevrier and Teixeira, 2024). Third, the wording "selects certain decisions" is intentionally non-specific to avoid deception, since "certain" may refer to one decision or multiple decisions. This phrasing allows us to hold the informational content of advice constant across the Human and Algorithm conditions.

4.5.2 Confidence Condition

Table 1 summarizes the full design. In the baseline — Without — condition, participants receive only the advisor's "left–right" recommendation. In the Calibrated condition, advice is accompanied by a confidence level with mean 75%, matching the advisor's 75% accuracy. In the Overconfident condition, advice is accompanied by a confidence level with mean 85%, i.e., 10 percentage points above accuracy. Within each confidence condition, participants interact with two advisors that differ only in discrimination quality, each for 120 rounds. In the HL sequence, participants first interact with a high-discrimination advisor (rounds 1–120) and then with a low-discrimination advisor (rounds 121–240). In the LH sequence, the order is reversed.

4.6 Experimental procedure and payments

We conducted two pre-studies. First, an informal unpaid pilot with 10 participants calibrated task difficulty to reach a baseline accuracy of about 69%. Second, in May 2023, we ran the advisor-generation pre-study at the LEEN (Experimental Economics Laboratory of Nice) with 11 participants over two sessions. Each participant completed the task twice (480 rounds total). Average earnings were €26.80 for approximately 1.5 hours, including a €5 show-up fee.

We initially preregistered our hypotheses on AsPredicted. Because the original link later became inaccessible, we reposted the same preregistration text on OSF for transparency. [3] To estimate the required sample size, we ran 1,000 Monte Carlo simulations assuming an effect size of 0.10, 120 repetitions per participant, and a significance level of 0.05. The simulations yielded power of 0.92, indicating that 50 participants per treatment condition are sufficient.

The main experiment was conducted in October 2023 at the LEEN. We recruited participants using the ORSEE database (Greiner, 2015) and programmed the experiment in oTree (Chen et al., 2016), deployed on a Heroku server. Data cleaning and descriptive analyses were performed in R, and econometric analyses in Stata. The main study consists of 30 sessions with an average of 11 participants per session, for a total of 307 participants.[4] Each between-subject treatment includes at least 50 participants, with at least 25 per within-subject sequence.[5]

Before the main task, participants completed five practice rounds with step-by-step explanations of the MP payoff rule. After the 240 rounds, participants completed a questionnaire including demographics, CRT-6 (Primi et al., 2016), CRT-4 (Thomson and Oppenheimer, 2016), and the Social Interaction subscale of the NARS (Nomura et al., 2006; Dinet and Vivian, 2014). All items are reported in Appendix C.

Payments for the dot task were based on four randomly selected rounds, two drawn from rounds 1–120 and two from rounds 121–240. Only final decisions and confidence

[3] Link: https://osf.io/t8kvw/overview

[4] Precise number of participants across treatments with within-subject sequence 1 (2): Algo & Without: 25 (25), Human & Without: 25 (27), Algo & Calibrated: 26 (25), Human & Calibrated: 26 (25), Algo & Overconfident: 27 (26), Human & Overconfident: 25 (25). Participants who took part in the pre-study were excluded.

[5] Participants' ages range from 18 to 25 years. On average, 35.5% are male across conditions. A rank-sum test indicates no systematic differences across treatments ($p > 0.10$), except in the human - overconfident condition, which has a slightly higher share of participants aged 22–25.

reports were eligible for payment[6]. Participants earned up to €3.50 per selected round (maximum €14), plus a €5 show-up fee. Average total earnings were €15.53 for approximately one hour of participation.

5 Results

5.1 Dependent variables and measures

We rely on three main outcome measures. First, advice adoption is captured by the switch ratio, defined as the proportion of disagreement trials in which a participant revises their initial decision to match the advisor's recommendation. Let d_{it}^{before} denote participant i's initial decision at round t, d_{it}^{after} the final decision, and d_{it}^{a} the advisor's recommendation. On disagreement trials ($d_{it}^{\text{before}} \neq d_{it}^{a}$), a switch occurs when $d_{it}^{\text{after}} = d_{it}^{a}$. The participant-level switch ratio is

$$SR_i = \frac{\sum_t Switch_{it}}{\text{number of disagreement trials for } i}.$$

In the econometric analysis, we estimate random-effects panel logit models at the decision level with $Switch_{it} \in \{0, 1\}$ as the dependent variable. As a robustness check, we also report results using "Follow", defined for all rounds as a dummy equal to 1 whenever $d_{it}^{\text{after}} = d_{it}^{a}$, regardless of whether there was initial disagreement[7].

Second, accuracy is measured by the success rate, i.e., the share of correct decisions. We distinguish the pre-advice success rate (i.e., the success rate before advice) from the post-advice success rate (i.e., the success rate after advice). In regressions, we use a decision-level indicator $Success_{it} \in \{0, 1\}$.

Third, we compute two participant-level metacognitive measures. Calibration captures average bias:

$$\text{Calibration}_i = -\left| \frac{1}{n_i} \sum_{t=1}^{n_i} (k_{it} - x_{it}) \right|,$$

[6]To measure advice adoption, we implement a standard Judge–Advisor System (Sniezek and Buckley, 1995): an initial judgment is elicited before advice and a final judgment after advice. Only the decision after the advice is used for payoffs (see Section 4.6), which isolates the causal effect of advice on final performance. Participants are nonetheless naturally incentivized to report their true initial beliefs because the stimulus disappears before advice is shown.

[7]Throughout the results section, predicted probabilities from panel logit models are evaluated conditional on the random effect set to zero. Confidence intervals for differences in predicted probabilities are computed from contrasts of predictive margins.

where k_{it} is confidence in being correct and $x_{it} \in \{0, 1\}$ indicates correctness; values closer to zero indicate better calibration. Meta-d' is a signal-detection-based measure of metacognitive sensitivity (Maniscalco and Lau, 2012, 2014) that captures how well a participant's confidence distinguishes correct from incorrect decisions; higher values indicate better discrimination quality.[8] Both measures are standardized for the econometric analysis.

5.2 Descriptive Statistics

Table 2 reports descriptive statistics. Before advice, participants' success rate averages approximately 69% across all conditions (all $p > 0.10$), confirming that our advisor, whose accuracy is fixed at 75%, outperforms participants on average. Participants switch to follow the advisor's recommendation after disagreement in about 37% of cases. On average, post-advice success exceeds pre-advice success by roughly 5 percentage points.

Participants' calibration before advice is approximately -0.09, very close to calibration measured after advice (-0.08), and it does not differ significantly across treatment conditions (rank-sum test, $p > 0.10$). Meta-d' differs between the pre- and post-advice measurements, but we detect no pre-advice differences across treatments. Thus, although post-advice calibration and meta-d' may be endogenous to advice exposure, baseline metacognitive measures appear comparable across treatment conditions. This is reassuring for the mediation analysis in Section 5.5, as it indicates that pre-advice metacognitive differences are unlikely to confound treatment comparisons.

5.3 Advice Adoption

5.3.1 Disentangling calibration and discrimination effects

Overall, providing confidence information increases advice adoption, but this effect depends critically on the quality of the confidence signal. When the advisor has low discrimination quality, switch ratios do not differ from the no-confidence baseline, whether the advisor is well calibrated (33.5% vs. 33.2%, rank-sum test, $p = 0.90$) or overconfident (33.5% vs. 36.3%, rank-sum test, $p = 0.50$). By contrast, when discrimination quality is high, participants switch substantially more often than in the no-confidence condition, both for a well-calibrated advisor (33.5% vs. 41.4%, signed-rank test, $p < 0.003$) and for

[8] We compute meta-d' using the R package *craddm/metaSDT*.

Table 2: Main dependent variable summary

		Without	Calibrated		Overconfident	
		.	Low Disc.	High Disc.	Low Disc.	High Disc.
Switch Ratio		33.5%(.20)	33.2%(.21)	41.4%(.20)	36.3%(.22)	43.6%(.24)
Success Rate	Before	69.5%(.05)	69.6%(.06)	69.2%(.06)	69.4%(.06)	69.1%(.07)
Success Rate	After	73.4%(.05)	71.6%(.05)	78.2%(.06)	72.9%(.05)	77.1%(.06)
Calibration	Before	−0.08(.06)	−0.08(.06)	−0.09(.07)	−0.11(.08)	−0.11(.08)
Calibration	After	−0.09(.07)	−0.09(.06)	−0.08(.06)	−0.11(.08)	−0.09(.06)
Meta-d'	Before	0.6536(.45)	0.6476(.49)	0.6750(.46)	0.7591(.53)	0.8314(.57)
Meta-d'	After	1.1309(.50)	0.9656(.72)	1.4866(.64)	1.0266(.58)	1.4690(.57)

Notes. This table pools algorithmic and human advisors. Mean levels with standard deviations in parentheses are reported for switch ratio, success rate, and metacognitive abilities (calibration and meta-d') across treatment conditions. Except for the switch ratio, all variables are recorded before and after advice is given. The success rate is the proportion of correct decisions. Calibration is computed as the negative absolute difference between average confidence and average success rate, so that values closer to zero indicate better calibration. The meta-d' score follows Maniscalco and Lau (2012) and Maniscalco and Lau (2014). The switch ratio captures the proportion of trials in which participants change their initial decision to match the advisor's recommendation.

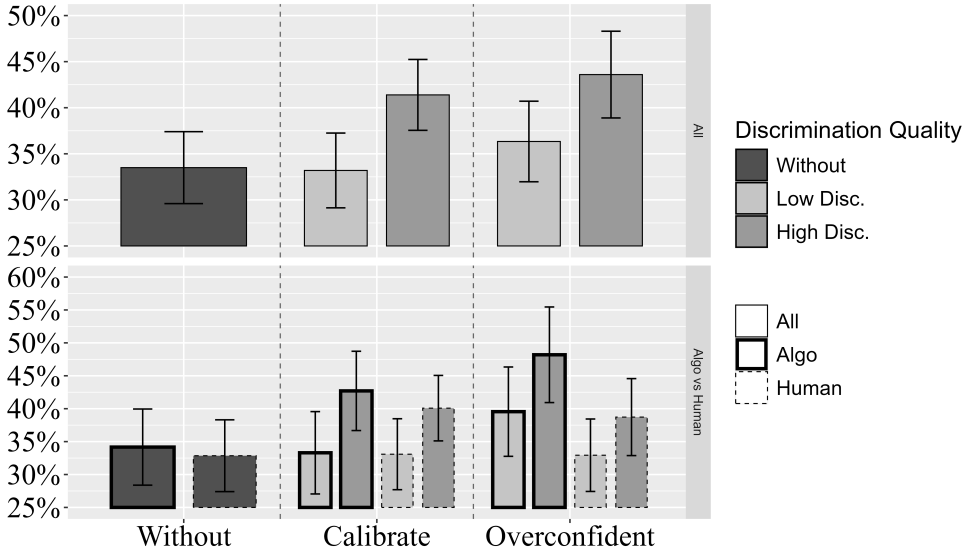


Figure 3: Switch ratio by advisor calibration and discrimination quality.

an overconfident advisor (33.5% vs. 43.6%, signed-rank test, $p < 0.002$). Overall, switch rates are 18.1% higher under high than under low discrimination. Moreover, participants are more likely to follow a well-calibrated advisor with high discrimination than an overconfident advisor with low discrimination (41.4% vs. 36.3%, rank-sum test, $p = 0.04$; see Figure 3).

These patterns are confirmed by the random-effects panel logit models reported in Table 3. The predicted probability of switching is 30.9% (95% CI [27.3, 34.6]) for weakly discriminating calibrated advice and 40.3% (95% CI [36.5, 44.0]) for strongly discrimi-

nating calibrated advice, an increase of about 9.4 percentage points. For overconfident advice, the corresponding probabilities are 38.5% (95% CI [34.2, 42.9]) and 46.7% (95% CI [42.3, 51.1]), an increase of about 8.2 points. Overconfidence also increases switching, but more modestly, by about 2.6 points when discrimination is weak and 2.7 points when discrimination is strong. We do not detect a significant interaction between the two dimensions (Over × High Discrimination (p = 0.590, 95% CI [-5.6, 3.2]), suggesting that discrimination and overconfidence contribute separately to advice adoption.

Result 1. Consistent with Hypothesis 1, discrimination quality is the main driver of advice adoption. Overconfidence plays a secondary role, and we find no evidence of an interaction between the two dimensions.

Table 3: Random-effects panel logit regressions of "Switch", "Follow" and "Success".

Dependent Variable	Switch		Follow		Success	
	(1)	(2)	(3)	(4)	(5)	(6)
Over	0.424**	0.405**	0.387**	0.366**		
	(0.178)	(0.180)	(0.154)	(0.156)		
Calibrated	-0.042	-0.073	-0.056	-0.082	-0.132**	-0.142**
	(0.171)	(0.174)	(0.144)	(0.147)	(0.072)	(0.069)
High Disc.	0.560***	0.560***	0.509***	0.509***	0.341***	0.331***
	(0.083)	(0.083)	(0.075)	(0.075)	(0.043)	(0.043)
Over × High	-0.099	-0.099	-0.105	-0.105		
Disc.	(0.129)	(0.129)	(0.116)	(0.116)		
Calibrated × High					0.255***	0.265***
Disc.					(0.070)	(0.069)
Algorithm	0.110	0.087	0.083	0.069		
	(0.133)	(0.142)	(0.114)	(0.122)		
Success Rate					3.333***	3.333***
Before Advice					(0.004)	(0.004)
Constant	4.203***	4.503***	3.538***	3.745***	-0.504***	-2.659***
	(0.297)	(0.392)	(0.256)	(0.333)	(0.306)	(3.071)
Control	No	Yes	No	Yes	No	Yes
Observations	29,157	29,157	73,680	73,680	36,720	36,720
Participants	307	307	307	307	153	153
Wald χ^2	504.9	510.2	1539	1547	1193	1261
p-value	0.000	0.000	0.000	0.000	0.000	0.000

Notes: Robust standard errors at the individual level in parentheses. "Switch" includes observations where participants initially disagreed with the advisor; "Follow" includes all decisions. "Success" represents post-advice success rate; it is a dummy variable equal to 1 if the answer is correct, and 0 otherwise. Binary variables: Over (overconfident advisor), Calibrated (well-calibrated advisor), High Discrimination, Algorithm. In columns (5) and (6), we add pre-advice success. In columns (2), (4), and (6), we add "NARS", "CRT score", "Male", "Years of Study", "Age", "Study Domain", "round", and "Sequence HL" (if the advisor first has high discrimination quality and then low, it equals 1; 0 otherwise). Except for "Advisor more confident", no other variables are significant. We discuss further the role of advisor confidence in Section 5.5. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

5.3.2 Algorithmic vs. Human Advisors

Overall, we find no systematic source differences in advice adoption. Switch ratios are similar for algorithmic and human advisors in the Without condition, in both well-calibrated conditions, and in the overconfident–low-discrimination condition.

An exception arises when the advisor is overconfident and highly discriminating. In that condition, participants are 24.5% more likely to switch to algorithmic than to human advice (48.2% vs. 38.7%; t-test, $p = 0.021$, rank-sum test, $p = 0.045$). Panel logit models reported in Table 4 confirm this pattern: in the overconfident–high-discrimination condition, algorithmic advice is more likely to be followed than human advice. The predicted probability of switching the advice is 49.1% (95% CI [42.1, 56.2]) for algorithmic advisors, compared with 38.7% (95% CI [33.8, 43.6]) for human advisors, a difference of 10.4 percentage points ($p < 0.01$, 95% CI [1.8, 19.1]). In all other conditions, we do not detect significant differences between algorithmic and human advisors.[9]

Result 2. We find partial support for Hypothesis 2. In most conditions, participants respond similarly to algorithmic and human advisors. A source effect emerges only when the advisor is both overconfident and highly discriminating.

5.4 Effect of Advisor Confidence Quality on DM Accuracy

5.4.1 Discrimination quality, overconfidence, and accuracy

Pre-advice success rate is stable at approximately 69% across all conditions. Figure 4 shows post-advice success rates by treatment condition. Receiving advice without an explicit confidence level increases success by 5.5 percentage points relative to pre-advice performance (signed-rank test, $p < 0.001$). Providing confidence information yields an additional gain of 2.2 percentage points on average relative to advice without confidence (rank-sum test, $p < 0.01$).

However, this additional benefit depends on discrimination quality. When discrimination is low, post-advice success under a well-calibrated advisor (71.6%) or an overconfident advisor (72.2%) does not differ significantly from the no-confidence benchmark (73.4%; rank-sum tests, $p = 0.23$ and $p = 0.90$, respectively). By contrast, when dis-

[9] Additional estimates suggest that the marginal effect of advisor confidence differs slightly by source only in the overconfident–high-discrimination condition; see Table 4 and Appendix D. Robustness checks using "Follow" as the dependent variable are reported in Appendix D, Table 9, column (5).

Table 4: Random-effects panel logit regressions of "Switch" by advisory profile.

Dependent variable	Switch				
	Without	Well-calibrated		Overconfident	
	.	Weak Disc.	Strong Disc.	Weak Disc.	Strong Disc.
	(1)	(2)	(3)	(4)	(5)
Algorithm	0.090	-0.043	0.780	1.063	2.034**
	(0.178)	(0.850)	(0.694)	(0.933)	(0.740)
Advisor more confident		0.057***	0.071***	0.053***	0.071***
		(0.008)	(0.007)	(0.007)	(0.007)
Algorithm × Advisor		0.001	-0.009	-0.009	-0.019*
more confident		(0.011)	(0.010)	(0.010)	(0.008)
Constant	-0.726***	-4.711***	-5.521***	-5.116***	-6.215***
	(0.159)	(0.556)	(0.500)	(0.500)	(0.500)
Observations	9,679	4,678	4,986	4,644	5,170
Participants	102	102	102	103	103
Log Likelihood	−6,190	−2,841	−2,894	−2,977	−3,197
AIC	12,387	5,693	5,799	5,964	6,403

Notes: Robust standard errors clustered at the participant level are reported in parentheses. The sample includes only trials in which the participant's initial "left–right" decision differs from the advisor's recommendation (disagreement trials). We control for each treatment variation: "Without", Well-calibrated & Low Discrimination, Well-calibrated & High Discrimination, Overconfidence & Low Discrimination, and Overconfidence & High Discrimination. "Advisor more confident" is a dummy equal to 1 if the advisor's confidence level is higher than the participant's confidence level, and 0 otherwise. We also control for "round number" and "Sequence HL" (if the advisor first has high discrimination quality and then low, it equals 1; 0 otherwise). No control variable is significant. Appendix D, Figure 9 shows similar results using the same model with "Follow" as the dependent variable. We discuss further the role of advisor confidence in Section 5.5. See Table 9 in Appendix D for a robustness check with "Follow" as the dependent variable. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

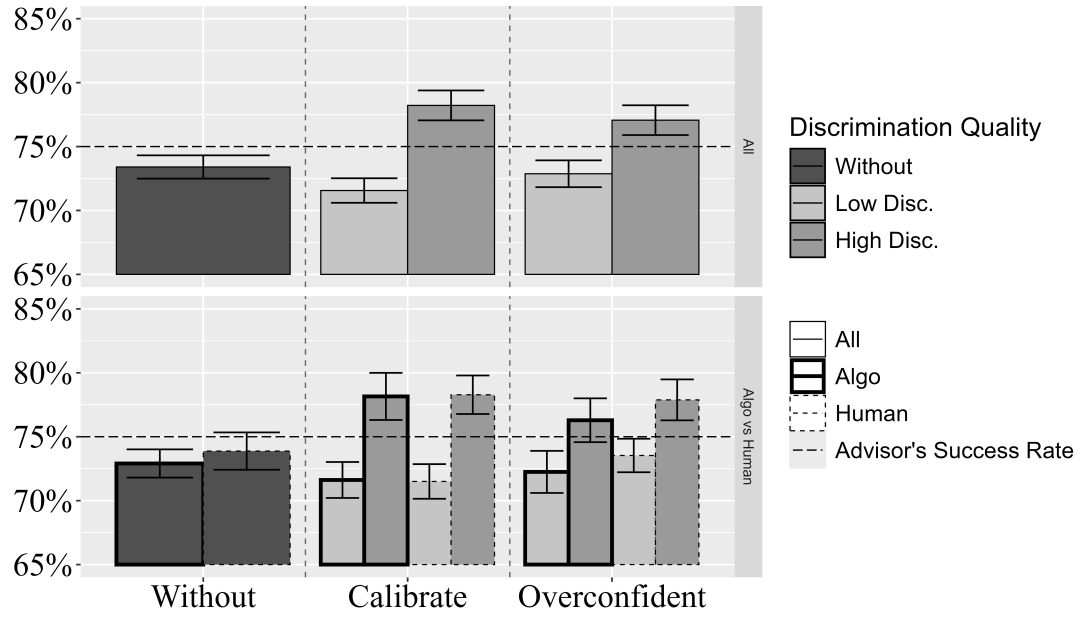


Figure 4: Average success rate after receiving advice.

crimination is high, post-advice success is substantially higher both for a well-calibrated advisor (78.2%) and for an overconfident advisor (77.1%) than in the no-confidence condition (73.4%; rank-sum tests, both $p < 0.001$). The difference between well-calibrated and overconfident advice is not statistically significant, whether discrimination is low (71.6% vs. 72.9%, rank-sum test, $p = 0.25$) or high (78.2% vs. 77.1%, rank-sum test, $p = 0.19$). Regression estimates in Table 3 confirm these patterns. The interaction Calibrated $\times$ High Discrimination increases by 2.3% ($p < 0.01$, 95% CI [0.009, 0.03]) the accuracy, whereas Over $\times$ High Discrimination is not significant ($p > 0.1$).

Result 3. Consistent with Hypothesis 3, discrimination quality is the main determinant of post-advice accuracy. The highest post-advice success rate is achieved when the advisor is well calibrated and highly discriminating.

5.5 DM Metacognition, Advice Adoption, and Accuracy

The preceding analyses show that discrimination quality shapes both advice adoption and post-advice accuracy, and that a source effect emerges only in a specific condition. These aggregate patterns may nevertheless mask substantial heterogeneity across decision makers. If participants differ in how well they evaluate their own judgments, they may respond differently to the same confidence signal and, as a result, benefit differently from advice. We therefore examine how participants' own metacognitive characteristics

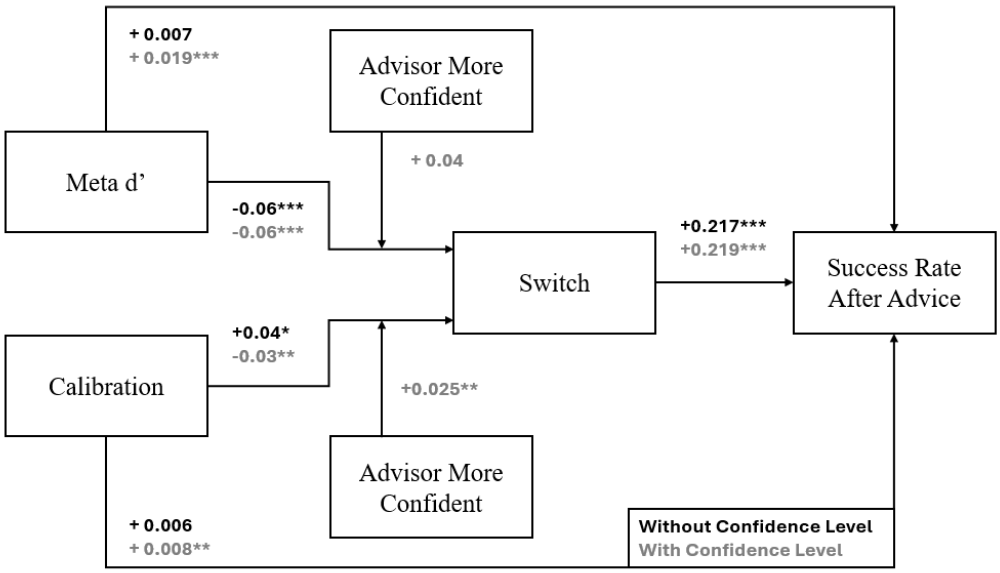


Figure 5: **Path analysis linking metacognitive ability, switching, and post-advice success.** The figure summarizes the estimated relationships on disagreement trials, that is, trials in which the participant's initial "left–right" decision differs from the advisor's recommendation. Estimates are obtained from random-effects panel logit models with standard errors clustered at the participant level, controlling for advisor type, series order, and round. Path coefficients are reported as average marginal effects (AME), except for the effect of switching on post-advice success, which is reported as a risk difference (RD) based on predicted probabilities. The figure is estimated on pooled data including both human and algorithmic advisors. See Appendix D, Figure 17, for the corresponding pooled logit specification. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

are associated with switching behavior and post-advice success.

To do so, we estimate a path analysis at the decision level, restricted to disagreement trials. We use random-effects panel logit models with participant-level random intercepts, controlling for advisor type and series fixed effects. The first set of models relates metacognitive ability to switching; in the conditions with confidence, we additionally include an indicator for whether the advisor reports higher confidence than the participant. The second set of models relates switching to post-advice success. We summarize these relationships graphically in Figure 5. Total, indirect, and direct effects (TE/NIE/NDE) for a +1 SD change in meta-d' or calibration are obtained by G-computation with a participant-cluster bootstrap (1000 repetitions). In the Without confidence condition, the estimated total effects are $TE_{\text{meta-}d'} = -0.006$ (n.s.) and $TE_{\text{calibration}} = 0.010$ ($p < 0.05$). In the With confidence conditions, they are $TE_{\text{meta-}d'} = 0.004$ (n.s.) and $TE_{\text{calibration}} = -0.010$ ($p < 0.05$). Appendix D reports a pooled logit specification as a robustness check and yields the same qualitative conclusions.

Calibration and meta-d' capture distinct aspects of metacognitive ability. Figure 5 shows that switching is strongly associated with post-advice success: switching increases success by about 22 percentage points both without confidence (Risk Difference = 0.217, $p < 0.001$) and with confidence (Risk Difference = 0.219, $p < 0.001$). The figure is estimated on pooled data including both human and algorithmic advisors, whereas the panel analyses below focus on algorithmic-advisor trials only (Tables 5 and 6). Corresponding pooled tables are reported in Appendix D and lead to the same qualitative conclusions.

Table 5 shows that participants with higher meta-d' switch more conservatively. In the pooled specification with confidence (col. 2), a one-standard-deviation increase in meta-d' reduces the probability of switching by 7.7 percentage points when the advisor is not more confident than the participant ($p < 0.001$, 95% CI[-11.6,-3.8]), and by 4.2 percentage points when the advisor is more confident ($p < 0.05$, 95% CI[-8.2, - 0.1]). Consistent with this pattern, the predicted probability of switching falls from 27.0% to 12.3% as meta-d' increases from -1 SD to +1 SD when the advisor is not more confident, and from 52.8% to 44.4% when the advisor is more confident. Thus, although the advisor's confidence attenuates the negative association between meta-d' and switching, participants with higher meta-d' remain less likely to follow the advisor even when the advisor reports higher confidence. Because switching is beneficial on average, this conservative pattern limits gains from high-quality advice but also reduces exposure to misleading signals.

Calibration displays a different pattern. In the pooled specification, calibration is not strongly related to switching on the probability scale: a one-standard-deviation increase in calibration changes switching by -2.0 percentage points when the advisor is not more confident ($p = 0.197$, 95% CI [-5.0, 1.0]) and by only +0.3 percentage points when the advisor is more confident ($p = 0.901$, 95% CI [-3.7, 4.3]). Table 6 shows that calibration is positively associated with post-advice success overall, but that the returns to switching depend on the advisor's confidence quality. When the advisor is well calibrated and highly discriminating, better-calibrated participants benefit from switching . By contrast, when the advisor is overconfident, greater reliance on stated confidence becomes detrimental, especially when overconfidence is paired with high discrimination.

The contrast between these two metacognitive dimensions is particularly clear in the overconfident–high-discrimination condition. In that cell, participants with higher meta-d' do not obtain an additional switching-related advantage, but they do achieve

higher success overall. This is consistent with the idea that a more conservative switching style protects them from misleading confidence cues. Better-calibrated participants, by contrast, benefit when advisor confidence is informative but are more exposed when confidence is biased.

Result 4. Consistent with Hypotheses 4a and 4b, calibration and discrimination are associated with distinct patterns of advice use. Better-calibrated decision makers respond more strongly to advisor confidence, which helps when confidence is informative but hurts when it is biased. Decision makers with higher meta-d' adopt advice more conservatively, which protects them from misleading signals but limits gains from high-quality advice.

Table 5: Random-effects panel logit regressions of "Switch" to algorithmic advice as a function of participants' metacognitive skills.

Dependent variable	Switch					
	All		Well-calibrated		Overconfident	
	ALL		Weak Disc.	Strong Disc.	Weak Disc.	Strong Disc.
	(1)	(2)	(3)	(4)	(5)	(6)
Advisor more confident	1.814***	1.870***	1.512***	2.471***	1.222***	2.219***
	(0.114)	(0.111)	(0.158)	(0.213)	(0.230)	(0.194)
Meta-d'	-0.313***	-0.611***	-0.370	-0.472*	-0.736***	-0.835***
	(0.105)	(0.147)	(0.320)	(0.247)	(0.217)	(0.242)
Advisor more confident		0.384***	0.434*	0.515***	0.470***	0.282
× Meta-d'		(0.111)	(0.243)	(0.189)	(0.176)	(0.191)
Calibration	-0.031	-0.163	-0.070	-0.051	-0.219	-0.120
	(0.108)	(0.126)	(0.314)	(0.349)	(0.179)	(0.189)
Advisor more confident		0.177*	0.205	0.431	0.001	0.050
× Calibration		(0.098)	(0.189)	(0.291)	(0.141)	(0.143)
Round	0.002***	0.002***	0.007***	0.003	0.003*	-0.001
	(0.001)	(0.001)	(0.003)	(0.002)	(0.002)	(0.002)
Series HL	-0.460**	-0.485**	-0.422	-1.050**	0.055	0.085
	(0.198)	(0.200)	(0.576)	(0.486)	(0.442)	(0.478)
Constant	-2.733***	-2.749***	-2.430***	-1.435***	-1.808***	-1.255***
	(0.215)	(0.219)	(0.557)	(0.340)	(0.433)	(0.432)
Observations	14,755	14,755	2,353	2,496	2,404	2,717
Participants	154	154	51	51	53	53
Wald χ^2	306.9	332.4	105.5	267.1	62.98	235.9
p-value	0.000	0.000	0.000	0.000	0.000	0.000

Notes. Robust standard errors clustered at the participant level are reported in parentheses. The sample includes only trials with an algorithmic advisor in which the participant's initial "left–right" decision differs from the advisor's recommendation (disagreement trials). In columns (1) and (2), we include treatment-condition fixed effects. All specifications additionally control for series and round. "Switch" is a dummy equal to 1 if the participant follows the advice, and 0 otherwise. "Advisor more confident" is a dummy equal to 1 if the advisor's confidence level is higher than the participant's confidence level, and 0 otherwise. Meta-d' and Calibration are standardized to make their coefficients comparable across specifications. Higher values indicate higher metacognitive ability. See Appendix D, Table 12, for the same model estimated on the pooled sample including both human and algorithmic advisors, and Table 14 for the same model including individual controls (NARS, CRT score, age, male, years of study, and round). Both appendix tables support the results reported in this table. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table 6: Random-effects panel logit regressions of "Post-Advice Success Rate" from an algorithmic advisor on switching and metacognitive skills.

Dependent variable	Post-Advice Success					
	All		Well-calibrated		Overconfident	
	ALL		Weak Disc.	Strong Disc.	Weak Disc.	Strong Disc.
	(1)	(2)	(3)	(4)	(5)	(6)
Pre-Advice Success	1.392***	1.533***	1.465***	1.407***	0.962***	0.502
	(0.117)	(0.280)	(0.307)	(0.400)	(0.308)	(0.356)
Switch	1.082***	1.040***	0.778***	1.522***	0.767***	0.830***
	(0.055)	(0.120)	(0.129)	(0.264)	(0.103)	(0.159)
Meta-d'	0.074***	0.047	0.081**	0.108*	0.191***	0.156**
	(0.017)	(0.038)	(0.041)	(0.057)	(0.052)	(0.075)
Switch		-0.043	-0.136	0.227	-0.161*	0.005
× Meta-d'		(0.065)	(0.103)	(0.148)	(0.087)	(0.085)
Calibration	0.034**	0.075**	0.078**	0.167**	0.104***	0.132**
	(0.015)	(0.032)	(0.038)	(0.073)	(0.038)	(0.059)
Switch		-0.154*	-0.195	-0.045	-0.138**	-0.326***
× Calibration		(0.092)	(0.129)	(0.142)	(0.068)	(0.060)
Series	0.001	0.045	-0.042	0.060	-0.039	0.030
	(0.031)	(0.076)	(0.088)	(0.118)	(0.091)	(0.101)
Constant	-0.703***	-0.872***	-0.851***	-0.526**	-0.594***	-0.083
	(0.071)	(0.151)	(0.157)	(0.257)	(0.174)	(0.245)
Observations	29,157	4,785	2,353	2,496	2,404	2,717
Participants	307	50	51	51	53	53
Wald χ^2	496.5	111.2	62.57	84.94	94.66	196.4
p-value	0.000	0.000	0.000	0.000	0.000	0.000

Notes. Robust standard errors clustered at the participant level are reported in parentheses. The sample includes only trials with an algorithmic advisor in which the participant's initial "left–right" decision differs from the advisor's recommendation (disagreement trials). In columns (1) and (2), we include treatment-condition fixed effects. All specifications additionally control for series and round. "Post-Advice Success" is a dummy equal to 1 if the participant's second "left–right" decision is correct, and 0 otherwise. "Pre-Advice Success" is a dummy equal to 1 if the participant's first "left–right" decision is correct, and 0 otherwise. "Switch" is a dummy equal to 1 if the participant follows the advice, and 0 otherwise. "Advisor more confident" is a dummy equal to 1 if the advisor's confidence level is higher than the participant's confidence level, and 0 otherwise. Meta-d' and Calibration are standardized to make their coefficients comparable across specifications. Higher values indicate higher metacognitive ability. See Appendix D, Table 13, for the same model estimated on the pooled sample including both human and algorithmic advisors, and Table 15 for the same model including individual controls (NARS, CRT score, age, male, years of study, and round). Both appendix tables support the results reported in this table. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

6 Discussion and Conclusion

This paper asks which dimension of confidence quality matters most for advice adoption and post-advice accuracy, and how decision makers' own metacognitive abilities interact with the advisor's confidence profile. We study these questions by independently manipulating calibration and discrimination quality in a controlled perceptual environment and by comparing human and algorithmic advisors.

Our first main finding is that discrimination quality—whether the advisor's confidence tracks correctness at the decision level—is the primary driver of both advice adoption and accuracy gains. This helps reconcile two strands of the literature. The confidence heuristic suggests that decision makers favor confident advisors, yet other work finds that confidence disclosure does not always increase advice use. Our results suggest that the key distinction is whether confidence is informative: when confidence tracks correctness, it functions as a useful cue; when it does not, even high stated confidence fails to increase adoption. Overconfidence, once discrimination is accounted for, plays a secondary role and does not appear to amplify the effect of discrimination. This qualifies prior findings that do not fully disentangle calibration from discrimination. In our repeated design, participants can observe the relationship between confidence and accuracy over time, which may help them discount uninformative overconfidence more effectively than in single-shot settings.

Our second main finding concerns source. Across most conditions, participants respond similarly to algorithmic and human confidence. However, a source effect emerges when the advisor is overconfident and highly discriminating: in that condition, participants follow algorithmic advice more often than equivalent human advice. One interpretation is that decision makers place greater weight on algorithmic confidence when it appears informative, perhaps because they perceive it as more objective or data-driven than human confidence. More broadly, this result suggests that source effects are conditional on the quality and salience of the confidence signal, rather than uniform across settings.

Our third main finding is that calibration and discrimination are associated with distinct patterns of advice use. Participants with higher meta-d' adopt advice more conservatively: they switch less often, which limits gains from high-quality advice but also protects them from misleading signals. Better-calibrated participants, by contrast, re-

spond more strongly to the advisor's stated confidence, which helps them when the signal is informative but hurts them when it is biased. This extends the results of Caplin et al. (2025), who emphasize calibration as a key metacognitive trait for benefiting from algorithmic advice, by showing that discrimination quality is a complementary and independently important dimension. More broadly, our results suggest that the value of advisor confidence depends on the match between the advisor's confidence profile and the user's metacognitive strengths.

These findings also carry practical implications for the design of decision-support systems. Our results suggest that improving the discrimination quality of confidence signals may matter more than calibration alone for effective advice use. They also suggest that confidence disclosure by itself may be insufficient: what matters is whether the disclosed confidence is genuinely informative. More broadly, heterogeneity in users' metacognitive profiles implies that a uniform confidence display may not serve all users equally, raising the possibility that adaptive interfaces tailored to users' metacognitive strengths could improve decision quality.

Our study has limitations. The binary and repeated perceptual task provides extensive feedback that is rarely available in field settings. Participants' metacognitive measures may be endogenous, although our within-cell analyses mitigate this concern. Finally, the advisor-source manipulation varies labels while holding informational content constant, and therefore does not capture settings in which algorithmic and human advisors differ in actual accuracy or error structure. Testing these mechanisms in richer decision environments, with adaptive confidence displays, and in generative-AI contexts would be a natural next step.

Declarations

Conflicts of interest: The authors declare that they have no conflict of interest.

Use of Artificial Intelligence: The authors declare that they used an AI-based large language model (LLM) solely to identify and correct errors in the manuscript. However, the authors certify that AI-LLMs were not used for any other tasks in this project.

Bibliography

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., and Frith, C. D. (2010). Optimally interacting minds. Science, 329(5995):1081–1085.

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.

Biermann, J., Horton, J. J., and Walter, J. (2022). Algorithmic advice as a credence good. ZEW-Centre for European Economic Research Discussion Paper, (22-071).

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.

Brynjolfsson, E., Li, D., and Raymond, L. (2023). Generative ai at work «, national bureau of economic research working paper 31161.

Burfurd, I. and Wilkening, T. (2018). Experimental guidance for eliciting beliefs with the stochastic becker–degroot–marschak mechanism. Journal of the Economic Science Association, 4(1):15–28.

Candrian, C. and Scherer, A. (2024). How terminology affects users' responses to system failures. Human factors, 66(8):2082–2103.

Cao, S., Jiang, W., Yang, B., and Zhang, A. L. (2023). How to talk when a machine is listening: Corporate disclosure in the age of ai. The Review of Financial Studies, 36(9):3603–3642.

Caplin, A., Deming, D., Li, S., Martin, D., Marx, P., Weidmann, B., and Ye, K. J. (2025). The abcs of who benefits from working with ai: Ability, beliefs, and calibration. Management Science.

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

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. Journal of Behavioral and Experimental Finance, 9:88–97.

Chevrier, M., Corgnet, B., Guerci, E., and Rosaz, J. (2024). Algorithm credulity: Human and algorithmic advice in prediction experiments. Available at SSRN 4828701.

Chevrier, M. and Teixeira, V. (2024). Algorithm delegation and responsibility: Shifting blame to the programmer? Technical report, Groupe de REcherche en Droit, Economie, Gestion (GREDEG CNRS), Université ....

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.

Chugunova, M. and Sele, D. (2022). We and it: An interdisciplinary review of the experimental evidence on how humans interact with machines. Journal of Behavioral and Experimental Economics, 99:101897.

Dargnies, M.-P., Hakimov, R., and Kübler, D. (2024). Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. Management Science.

Delmas, H., Denault, V., Burgoon, J. K., and Dunbar, N. E. (2024). A review of automatic lie detection from facial features. Journal of Nonverbal Behavior, 48(1):93–136.

Dinet, J. and Vivian, R. (2014). Exploratory investigation of attitudes towards assistive robots for future users. Le travail humain, 77(2):105–125.

Fleming, S. M. (2024). Metacognition and confidence: A review and synthesis. Annual Review of Psychology, 75(1):241–268.

Fleming, S. M. and Lau, H. C. (2014). How to measure metacognition. Frontiers in human neuroscience, 8:443.

Gaertig, C. and Simmons, J. P. (2023). Are people more or less likely to follow advice that is accompanied by a confidence interval? Journal of Experimental Psychology: General, 152(7):2008.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with orsee. Journal of the Economic Science Association, 1(1):114–125.

Hainguerlot, M., Gajdos, T., Vergnaud, J.-C., and de Gardelle, V. (2023). How overconfidence bias influences suboptimality in perceptual decision making. Journal of Experimental Psychology: Human Perception and Performance, 49(4):537.

Hollard, G., Massoni, S., and Vergnaud, J.-C. (2016). In search of good probability assessors: an experimental comparison of elicitation rules for confidence judgments. Theory and Decision, 80(3):363–387.

Holt, C. A. and Smith, A. M. (2009). An update on bayesian updating. Journal of Economic Behavior & Organization, 69(2):125–134.

Jussupow, E., Benbasat, I., and Heinzl, A. (2020). Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion.

Karni, E. (2009). A mechanism for eliciting probabilities. Econometrica, 77(2):603–606.

Langer, M., Hunsicker, T., Feldkamp, T., König, C. J., and Grgić-Hlača, N. (2022). "look! it'sa computer program! it's an algorithm! it's ai!": Does terminology affect human perceptions and evaluations of algorithmic decision-making systems? In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, pages 1–28.

Lehmann, C. A., Haubitz, C. B., Fügener, A., and Thonemann, U. W. (2022). The risk of algorithm transparency: How algorithm complexity drives the effects on the use of advice. Production and Operations Management, 31(9):3419–3434.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. The Journal of the Acoustical society of America, 49(2B):467–477.

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.

Maniscalco, B. and Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Consciousness and cognition, 21(1):422–430.

Maniscalco, B. and Lau, H. (2014). Signal detection theory analysis of type 1 and type 2 data: meta-d, response-specific meta-d, and the unequal variance sdt model. In The cognitive neuroscience of metacognition, pages 25–66. Springer.

Massoni, S., Gajdos, T., and Vergnaud, J.-C. (2014). Confidence measurement in the light of signal detection theory. Frontiers in psychology, 5:1455.

Massoni, S. and Roux, N. (2017). Optimal group decision: A matter of confidence calibration. Journal of Mathematical Psychology, 79:121–130.

Nomura, T., Kanda, T., and Suzuki, T. (2006). Experimental investigation into influence of negative attitudes toward robots on human–robot interaction. Ai & Society, 20:138–150.

Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Science, 381(6654):187–192.

Offert, F. and Bell, P. (2021). Perceptual bias and technical metapictures: critical machine vision as a humanities challenge. Ai & Society, 36(4):1133–1144.

Phillips, J. M. (1999). Antecedents of leader utilization of staff input in decision-making teams. Organizational Behavior and Human Decision Processes, 77(3):215–242.

Price, P. C. and Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. Journal of Behavioral Decision Making, 17(1):39–57.

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., and Hamilton, J. (2016). The development and testing of a new version of the cognitive reflection test applying item response theory (irt). Journal of Behavioral Decision Making, 29(5):453–469.

Raux, R. and Dreyfuss, B. (2025). Human learning about ai. In Proceedings of the 26th ACM Conference on Economics and Computation, pages 1106–1106.

Sniezek, J. A. and Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision making. Organizational behavior and human decision processes, 62(2):159–174.

Sniezek, J. A. and Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge-advisor system. Organizational behavior and human decision processes, 84(2):288–307.

Snijders, C., Conijn, R., de Fouw, E., and van Berlo, K. (2023). Humans and algorithms detecting fake news: Effects of individual and contextual confidence on trust in algorithmic advice. International Journal of Human–Computer Interaction, 39(7):1483–1494.

Stanciu, O. and Fiser, J. (2022). Do humans recalibrate the confidence of advisers or take their confidence at face value? In Proceedings of the Annual Meeting of the Cognitive Science Society, volume 44.

Taudien, A., Fügener, A., Gupta, A., and Ketter, W. (2022). Calibrating users' mental models for delegation to ai.

Thomson, K. S. and Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. Judgment and Decision making, 11(1):99–113.

Toplak, M. E., West, R. F., and Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. Thinking & reasoning, 20(2):147–168.

Winkler, R. L. and Murphy, A. H. (1968). " good" probability assessors. Journal of Applied Meteorology (1962-1982), pages 751–758.

Yaniv, I. (1997). Weighting and trimming: Heuristics for aggregating judgments under uncertainty. Organizational behavior and human decision processes, 69(3):237–249.

Yates, J. F., Price, P. C., Lee, J.-W., and Ramirez, J. (1996). Good probabilistic forecasters: The 'consumer's' perspective. International Journal of Forecasting, 12(1):41–56.

Zhang, Y., Liao, Q. V., and Bellamy, R. K. (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, pages 295–305.

A Advisor advice depending on the treatment

Table 7: Advisor advice depending on the treatment

	High Discrimination				Low Discrimination			
Round	**Choice**	**Calib**	**Over**	**Accuracy**	**Choice**	**Calib**	**Over**	**Accuracy**
1	Left	95	85	Correct	Left	65	95	Correct
2	Left	60	90	Correct	Right	65	80	False
3	Left	70	75	Correct	Right	60	85	False
4	Right	100	95	Correct	Right	60	80	Correct
5	Right	50	50	Incorrect	Right	65	75	Correct
6	Right	60	60	Incorrect	Left	70	85	Correct
7	Right	90	95	Correct	Right	90	95	Correct
8	Right	85	90	Correct	Right	70	80	Correct
9	Left	75	90	Correct	Right	75	85	False
10	Left	70	70	Correct	Left	75	70	Correct
11	Right	50	65	Incorrect	Right	70	100	False
12	Left	90	70	Correct	Right	75	80	False
13	Right	90	100	Correct	Left	65	75	False
14	Right	70	90	Correct	Right	65	85	Correct
15	Right	100	90	Incorrect	Right	75	75	False
16	Right	70	85	Correct	Right	80	95	Correct
17	Left	95	80	Correct	Left	75	85	Correct
18	Left	70	95	Correct	Left	85	80	Correct
19	Left	90	100	Correct	Right	95	85	False
20	Right	100	100	Correct	Right	70	80	Correct
21	Right	70	85	Correct	Right	80	85	Correct
22	Left	65	55	Incorrect	Right	70	80	Correct
23	Left	65	60	Correct	Left	70	85	Correct
24	Left	100	60	Correct	Left	75	85	Correct
25	Right	75	80	Correct	Right	75	95	Correct
26	Right	60	65	Correct	Right	70	80	Correct
27	Left	85	95	Incorrect	Right	80	95	Correct
28	Left	75	75	Correct	Left	65	85	Correct
29	Left	80	95	Incorrect	Left	85	75	False
30	Right	75	85	Correct	Left	85	80	False
31	Right	100	75	Correct	Right	75	90	Correct
32	Right	55	90	Correct	Left	70	95	False
33	Left	65	60	Correct	Left	65	85	Correct
34	Right	75	85	Correct	Right	75	90	Correct
35	Right	75	80	Incorrect	Left	75	85	Correct
36	Right	55	85	Incorrect	Left	65	85	Correct
37	Right	60	85	Correct	Left	75	100	False
38	Left	85	100	Incorrect	Right	80	80	Correct
39	Right	100	100	Incorrect	Right	80	80	False
40	Left	60	95	Incorrect	Right	80	80	Correct
41	Right	60	100	Correct	Right	80	90	Correct
42	Left	55	85	Correct	Left	80	85	Correct
43	Right	65	85	Correct	Right	75	90	Correct
44	Left	75	95	Correct	Left	75	95	Correct
45	Left	95	90	Correct	Left	70	70	Correct
46	Left	80	90	Incorrect	Right	80	90	Correct
47	Right	65	80	Correct	Right	75	90	Correct
48	Right	80	100	Correct	Right	70	95	Correct
49	Left	90	100	Correct	Left	85	80	Correct
50	Left	90	85	Incorrect	Left	70	90	Correct
51	Right	100	100	Correct	Right	70	80	Correct
52	Left	80	95	Correct	Right	80	95	Correct
53	Left	50	100	Correct	Left	75	75	Correct
54	Right	75	85	Correct	Right	60	90	False
55	Left	55	95	Correct	Left	85	80	Correct
56	Left	85	90	Correct	Left	75	95	Correct
57	Left	80	100	Incorrect	Right	75	85	Correct
58	Right	70	70	Correct	Right	95	85	Correct
59	Left	55	75	Correct	Left	80	90	Correct
60	Left	70	85	Correct	Right	80	95	False
Average	48%	75.58%	84.83%	75%	58%	74.66%	85.50%	75%
AUC	.	0.85	0.81	.	.	0.51	0.53	.

B Instructions

Instructions translated from French.

Welcome instructions (for every experiment conducted at LEEN):

Welcome to the Experimental Economics Laboratory of Nice (the LEEN). By agreeing to participate in this experiment, you express your full agreement with the Laboratory's regulations, available on the website or upon request. You will participate in an experiment where your decisions will be anonymous and will partly determine your final payment. Therefore, we invite you to read the following instructions carefully. In addition to earnings collected in the experiment and regardless of your decisions, a fixed amount of 5 euros will be given to you to cover your travel expenses. A variable amount will be added based on your decisions during the experiment. Your earnings will be paid individually and confidentially at the end of the experiment.In order not to alter the results of the experiment, we ask that you do not communicate with other participants. We also ask that you kindly turn off your mobile phones and refrain from using them throughout the duration of the experiment. If these rules are breached, the experiment will be interrupted, and earnings will be forfeited. If you encounter a technical problem, please simply raise your hand, and wait for the experimenter to come to you. Everybody in this room has access to the same instructions and will participate in the same experiment.

Experiment starting :
General presentation :

This experiment lasts on average 1 hour and is divided into 4 stages. In addition to the 5 euros for travel expenses, and 2 euros for your participation, you can earn up to 14 euros during this experimental task, depending on your choices. Therefore, you can earn a maximum of 21 euros.

Course of the Experiment:

1. You will perform the classic dot task for 5 rounds to practice. This part of the experiment is not paid.

2. You will perform the dot task with a counselor for 240 rounds. This part of the experiment is paid. Between rounds 120 and 121, you will be asked to fill out a questionnaire that is not paid.

3. You will fill out 1 questionnaire. This part of the experiment is not paid.

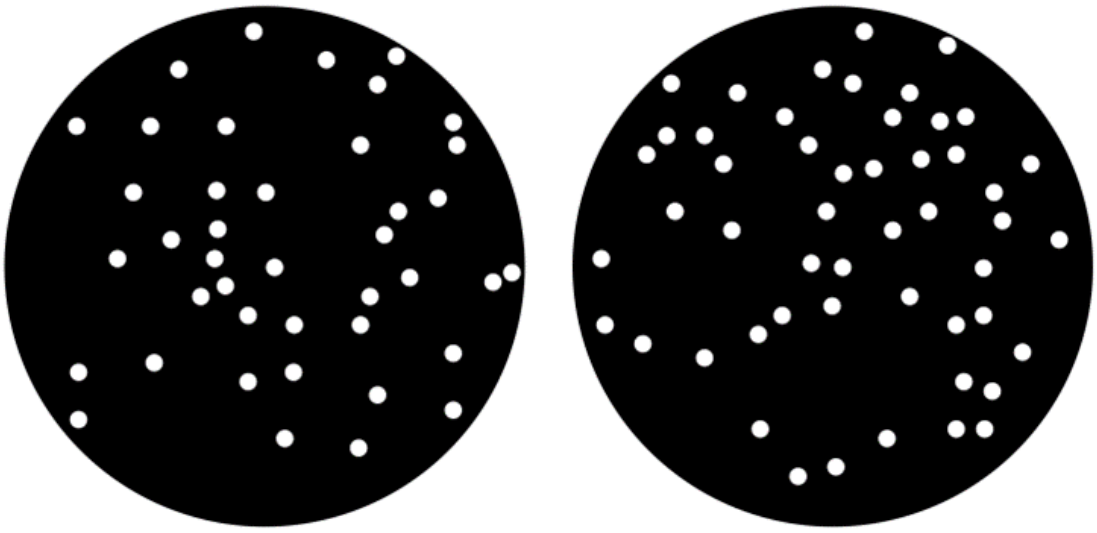
4. Payments.

Instruction : Perception Task

1) Procedure of the Task:

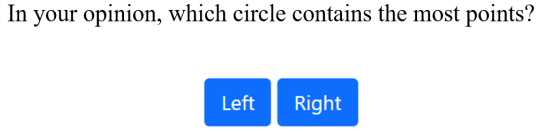
Two black circles containing white dots will appear on the screen for 750ms (Figure 6)

Figure 6: Blacks circles with white dots



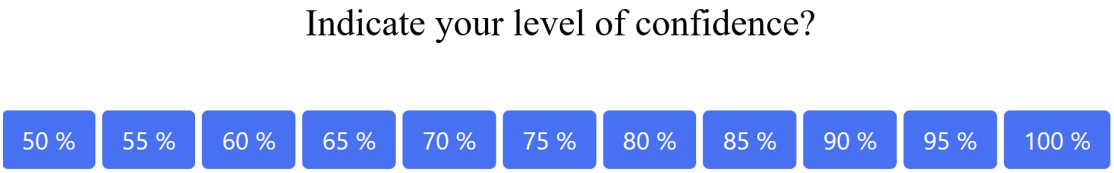
1. Determine which circle contains the most white dots (Figure 15).

Figure 7: Left-Right Decision



2. Determine your level of confidence (Figure 16).

Figure 8: Confidence Level



3. How sure are you that the circle you selected contains more white dots? For example, if you select Left and 60%, this means you are 60% sure that the left circle contains more white dots than the right circle.

How sure are you that the circle you selected contains more white dots? For example, if you select Left and 60%, this means you are 60% sure that the left circle contains more white dots than the right circle.

2) Payment:

For this training, we will randomly select one of the 5 rounds you will play to determine a fictitious payment that will serve as an example.

The higher your confidence level, the more your payment depends on your choice.

The lower your confidence level, the more your payment depends on luck.

A number between 0 and 100 (**called lottery 1**) is randomly drawn by the computer. Here are the possible situations:

Situation 1. If lottery 1 is lower than (or equal to) the confidence level of the selected round, you win 3.50 euros if you indicated the correct circle, and 0 euro otherwise.

Situation 2. If lottery 1 is higher than the confidence level of the selected round, the computer draws a second number between 0 and 100 (called lottery 2).

Situation 2.a If lottery 2 is lower than (or equal to) lottery 1, you win 3.50 euros.

Situation 2.b If lottery 2 is higher than lottery 1, you win 0 euro.

Practical example:

Situation 1. Imagine you chose Left with a confidence of 95%. Lottery 1 is equal to 72 and lottery 2 is equal to 69. The correct answer was Right, so you win 0 euro.

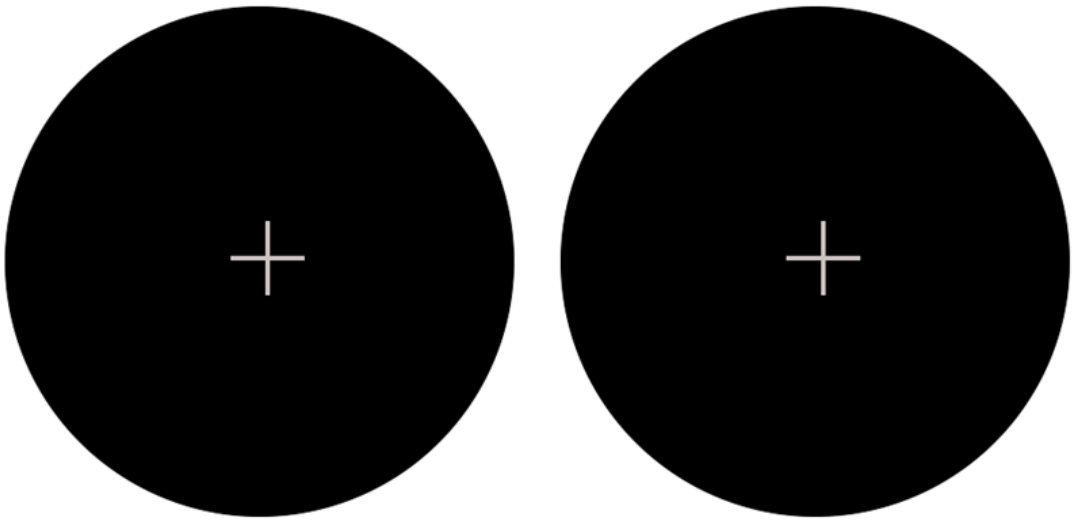
Situation 2.a But if you had chosen a confidence level of 70%, then you would have won 3.50 euros because lottery 2 is lower than lottery 1.

Situation 2.b Imagine you chose Left with a confidence of 60%. Lottery 1 is equal to 72 and lottery 2 is equal to 75. The correct answer is Left, but you win 0 euro because lottery 1 is higher than your confidence level and lottery 2 is higher than lottery 1.

Trial (Participant do the task 5 times):

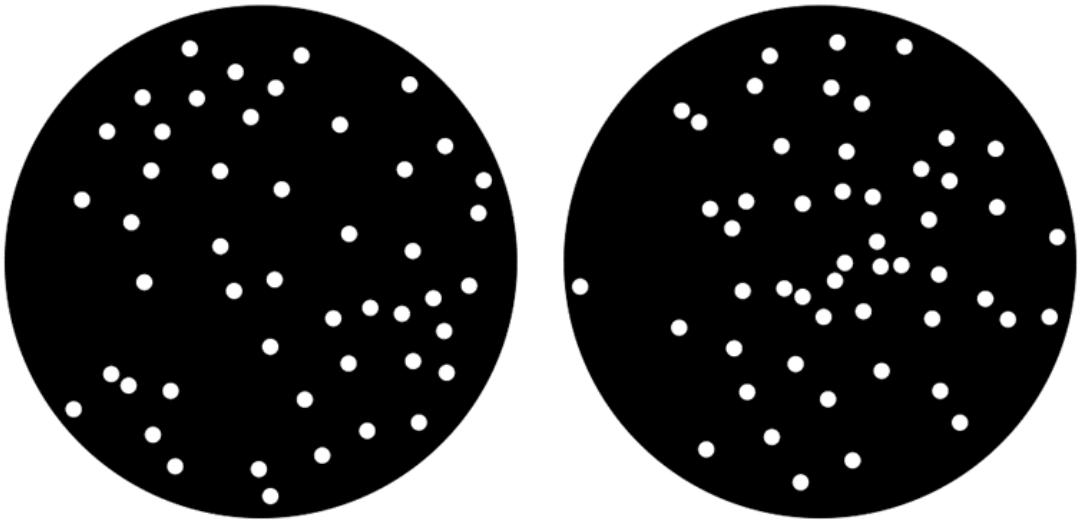
First step: Calibration. (Figure 13)

Figure 9: Calibration



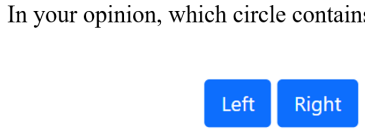
Second step: Stimuli. (Figure 14)

Figure 10: Stimuli



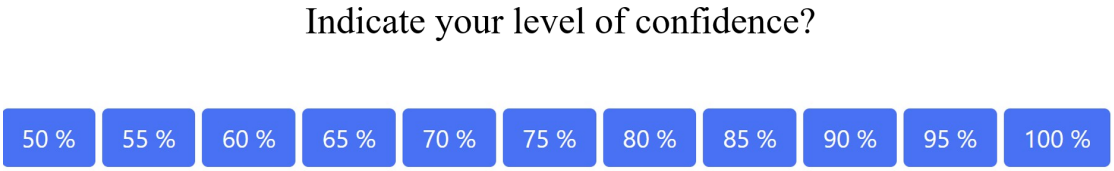
Third step: Decision. (Figure 15)

Figure 11: Decision



Fourth step: Confidence Level. (Figure 16)

Figure 12: Stimuli



Fifth step: Feedback.

You selected the **right circle**, and your confidence level was **75%** The 2 lotteries that were drawn: Lottery 1: **57** Lottery 2: **53**

Your prediction is:

In this example, you won **3 euros 50** because your prediction was correct, and your

confidence level (75) is higher than lottery 1 (57).

If your confidence level had been lower than lottery 1 (57), you would have also won **3 euros 50** because lottery 2 (53) is lower than lottery 1 (57).

However, if lottery 2 had been higher than lottery 1 and your confidence level lower than lottery 1, you would have won **0 euro**.

Reminder:

The higher your confidence level, the more your payment depends on your choice.

The lower your confidence level, the more your payment depends on luck.

To earn as much money as possible, you should provide a confidence level that is as accurate as possible, neither too high nor too low.

After five rounds of trials:

Round 1 was randomly selected:

Your hypothetical payment amounts to: **€3.50.** Payment for round 1: **€3.50.** Your prediction was: **Correct.** Your level of confidence: **75.** The 2 lotteries that were drawn:

Lottery 1: 24

Lottery 2: 70

Reminder:

The higher your confidence level, the more your payment depends on your choice.

The lower your confidence level, the more your payment depends on luck.

A number between 0 and 100 (called lottery 1) is randomly drawn by the computer. Here are the possible situations:

Situation 1.

If lottery 1 is lower than (or equal to) the confidence level of the selected round, you win €3.50 if you have indicated the correct circle, and €0 otherwise.

Situation 2.

If lottery 1 is higher than the confidence level of the selected round, the computer draws a second number between 0 and 100 (called lottery 2).

Situation 2.a

If lottery 2 is lower than (or equal to) lottery 1, you win €3.50.

Situation 2.b

If lottery 2 is higher than lottery 1, you win €0.

For the following part of the instructions, when the advisor is human, we use blue. Otherwise, we use red for an algorithm. When the advisor provide a confidence level, we put the text in green.

INSTRUCTION: Perception Task with Advisor

Procedure of the Task:

This is exactly the same task as before, except that you will be advised by another participant [OR] an algorithm. After observing the two circles, indicate your choice (right or left) and your level of confidence. Then, you will receive advice from the another participant [OR] an algorithm

Presentation of the other participant [OR] an algorithmic advisor:

During a previous experimental session, other players were asked to participate in the same experimental task that you are about to play, with the same monetary incentives. They were aware that their choices would aim to serve as advice for you.

[**Human Condition**] One of these participants will be your advisor, then **after 120 rounds, another participant will be assigned to you until the end ofthe task.**

[**Algorithm Condition**] The algorithm uses a database including the participants who played during that session. It selects certain decisions and confidence levels made by these participants.
After 120 rounds, a new algorithm will be assigned to you until the end of the task.

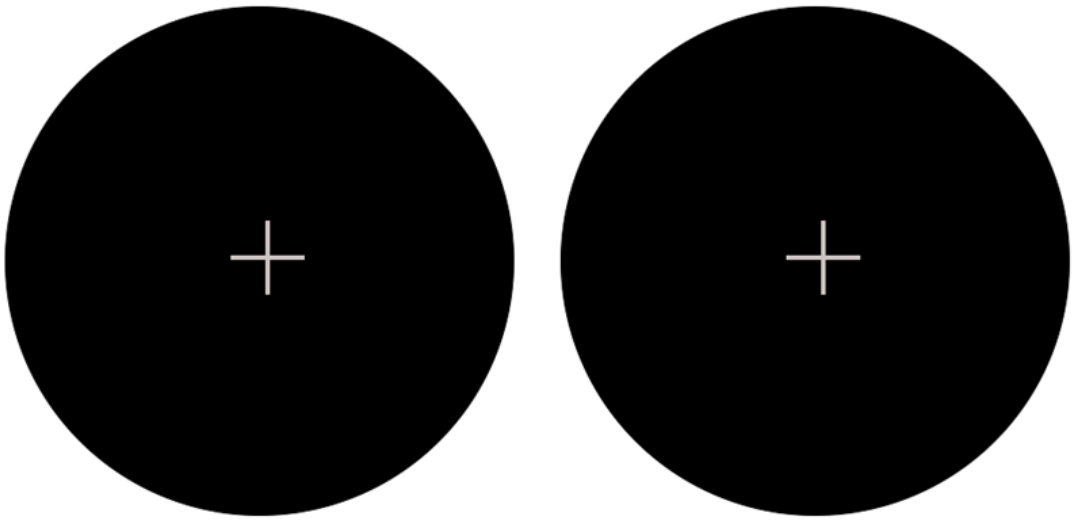
These two other participant [OR] algorithmic advisors have achieved results that are above or equal to the average of participants from the previous session.

Payment:

We use the same payment system as before, except that the decision considered will be the one made after the advice. During the 240 rounds, 4 rounds will be randomly selected (2 rounds between 1 and 120, 2 rounds between 121 and 240). For each selected round, you have a chance to win **€3.50 or €0**. Therefore, you can win a maximum of **€14** for this task.

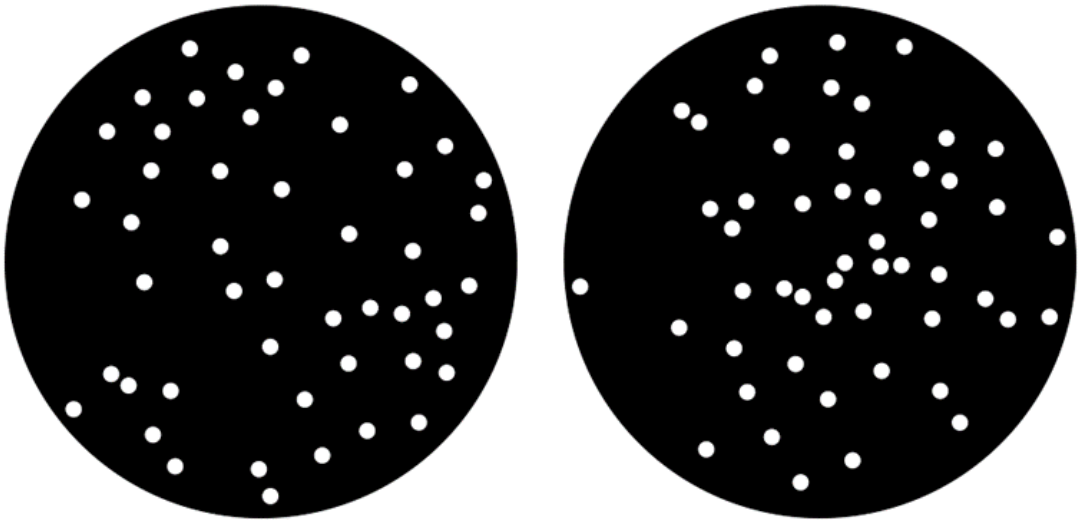
First step: Calibration. (Figure 13)

Figure 13: Calibration



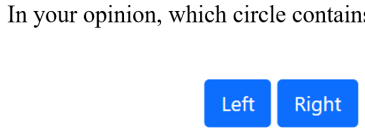
Second step: Stimuli. (Figure 14)

Figure 14: Stimuli



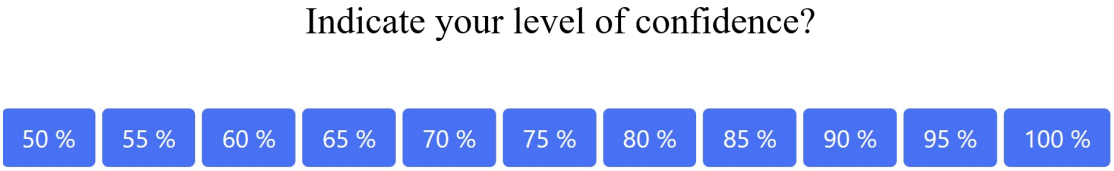
Third step: Decision. (Figure 15)

Figure 15: Decision



Fourth step: Confidence Level. (Figure 16)

Figure 16: Stimuli



Fifth step: Advice. (Figure 16)

The other participant **[OR]** The algorithm advises: **Left** [For confidence level condition] with a confidence level of: 75%.

You have selected:**Right** with a confidence level of: **80%.**

In your opinion, which circle contains the most points?

Left ◯◯ **Right**

Indicate your level of confidence?

Sixth step: Feedback. (Figure 16)

The other participant **[OR]** The algorithm selected the **Left** circle with a confidence level of: **75%**.

You selected the **Right** circle, and your confidence level was **90%**.

The other participant's **[OR]** The algorithm's prediction is: Correct

Your prediction is: Wrong

After 120 rounds:

Participants complete the CRT Test, provide personal data, and the NARS (see).

Change of the participant [OR] algorithm advising you.

This advisor has also achieved results equal to or higher than the average of participants from the previous session.

After 120 rounds:

Results:

Rounds 23, 99, 156, and 233 have been randomly selected:

Your total payment amounts to **7€.**

Payment for round 23: **3€ 50.**

Your prediction was: **Correct.**

confidence level: **90%.**

The 2 lotteries that were drawn: Lottery 1: **78** Lottery 2: **23.**

Payment for round 99: **0.**

Your prediction was: **Correct.**

confidence level: **50%.**

The 2 lotteries that were drawn: Lottery 1: **58** Lottery 2: **89.**

Payment for round 156: **3€ 50.**

Your prediction was: **False.**

confidence level: **50%.**

The 2 lotteries that were drawn: Lottery 1: 67 Lottery 2: **24.**

Payment for round 216: **0€.**

Your prediction was: **False.**

confidence level: **95%.**

The 2 lotteries that were drawn: Lottery 1: 7 Lottery 2: **29.**

C Questionnaire

C.1 NARS (Negative Attitude toward Algorithm)

Original questionnaire (Nomura et al., 2006), translated from English to French (Dinet and Vivian, 2014). The following questions assess the subjects' level of sociability with robots. Subjects could select from 5 response levels, ranging from "Strongly disagree" to "Strongly agree".

1. I would feel uneasy if I was given a job where I had to use robots.

2. The word "robot" means nothing to me.

3. I would feel nervous operating a robot in front of other people.

4. I would hate the idea that robots or artificial intelligences were making judgments about things.

5. I would feel very nervous just standing in front of a robot.

6. I would feel paranoid talking with a robot.

C.2 CRT (Cognitive Reflection Test)

We combined CRT 6 (modify from (Toplak et al., 2014)), and CRT 4 (from (Thomson and Oppenheimer, 2016)). Participants have 8 minutes to answer all the CRT 6 and CRT 4 questions.

CRT 6:

1. If 2 nurses take 2 minutes to measure the blood pressure of 2 patients, how long would it take 200 nurses to measure the blood pressure of 200 patients? [Correct answer: 2 minutes; intuitive answer: 200 minutes]

2. Soup and salad cost a total of 5.50 euros. The soup costs 5 euros more than the salad. How much does the salad cost? [Correct answer: 0.25 euro; intuitive answer: 0.5 euro]

3. Sally is making sun tea. Every hour, the concentration of the tea doubles. If it takes 6 hours for the tea to be ready, how long would it take for the tea to reach half of its final concentration? [Correct answer: 5 hours; intuitive answer: 3 hours]

4. If John can drink one barrel of water in 6 days, and Mary can drink one barrel of water in 12 days, how long would it take them to drink one barrel of water

together? [Correct answer: 4 days; intuitive answer: 9 days]

5. Jerry received both the 15th highest and the 15th lowest marks in the class. How many students are in the class? [Correct answer: 29 students; intuitive answer: 30 students]

6. A man buys a pig for 60 euros, sells it for 70 euros, buys it back for 80 euros, and then sells it again for 90 euros. How much has he made? [Correct answer: 20 euros; intuitive answer: 10 euros]

CRT 4:

1. If you're running a race and you pass the person in second place, what place are you in? [Correct answer: second; intuitive answer: first]
2. A farmer had 15 sheep and all but 8 died. How many are left? [Correct answer: 8; intuitive answer: 7]
3. Emily's father has three daughters. The first two are named April and May. What is the third daughter's name? [Correct answer: Emily; intuitive answer: June]
4. How many cubic feet of dirt are there in a hole that is 3' deep x 3' wide x 3' long? [Correct answer: none intuitive answer: 27]

C.3 Demographic Questionnaire:

Your age:

1. 18 - 21 years old.
2. 22 - 25 years old.
3. 26 - 29 years old.

Your gender (For anonymity reasons, you are not required to specify your gender if other):

1. Female
2. Male
3. Other

Education level (If you are in your first or second year, tick "High School Diploma".If you are in your third year, tick "2 years post-high school diploma"...):

1. Vocational certificate
2. High School Diploma
3. 2 years post-high school diploma
4. 3 years post-high school diploma
5. 5 years post-high school diploma
6. 8 years post-high school diploma
7. Other

Main field of study of your past or current education:

1. Biology, Health, Sports
2. Law, Political Science, Economics, and Management
3. Education, Teaching, Training
4. Letters, Languages, Arts, and Communication
5. Science, Engineering, Technology, Environment
6. Humanities and Social Science

D Econometrics model

Table 8: Switch ratio comparison between Algorithm and Human.

Condition	Algorithm	Human	Wilcoxon p	t-test p	n_{algo}	n_{human}
Without	34.2% (0.2)	32.9% (0.2)	0.849	0.741	50	52
Well-calibrated – Low Disc.	33.3% (0.22)	33.1% (0.19)	0.981	0.958	51	51
Well-calibrated – High Disc.	42.7% (0.21)	40.1% (0.17)	0.632	0.502	51	51
Overconfident – Low Disc.	39.5% (0.24)	32.9% (0.19)	0.214	0.132	53	50
Overconfident – High Disc.	48.2% (0.26)	38.7% (0.20)	0.045	0.021	53	50
All	39.0% (0.21)	35.2% (0.18)	0.200	0.100	154	153

Notes: Cells report mean switch ratio with standard deviation in parentheses. p-values from Wilcoxon rank-sum tests and two-sided t-tests.

Table 9: Panel logit models with random effects for "Follow".

Dependent variable	Follow				
	Without	Well-calibrated		Overconfident	
	.	Weak Disc.	Strong Disc.	Weak Disc.	Strong Disc.
	(1)	(2)	(3)	(4)	(5)
Algorithm	0.088	-0.160	0.617	1.329	1.240**
	(0.140)	(0.723)	(0.375)	(0.827)	(0.485)
Advisor more confident		0.035***	0.028***	0.047***	0.025***
		(0.006)	(0.003)	(0.006)	(0.004)
Algorithm × Advisor More confident		0.002	-0.007	-0.013	-0.011*
		(0.009)	(0.004)	(0.009)	(0.005)
Series	0.012	-0.406**	-0.147	-0.092	-0.038
	(0.140)	(0.149)	(0.124)	(0.148)	(0.142)
Constant	-1.917***	-4.343***	-3.730***	-5.809***	-3.730***
	(0.122)	(0.462)	(0.250)	(0.667)	(0.375)
Observations	24,480	12,240	12,240	12,360	12,360
Participants	102	102	102	103	103
Log Likelihood	−9,626.690	−4,604.217	−5,461.876	−4,890.777	−5,864.216
AIC	19,259.380	9,218.434	10,933.750	9,791.553	11,738.430

Notes: Robust standard errors clustered at the participant level are reported in parentheses. "Follow" equals 1 if the participant's final decision matches the advisor's decision, and 0 otherwise. We include observations where participants disagreed with the advisor during the initial choice. By construction, "Advisor more confident" (i.e., a dummy equal to 1 if the advisor's confidence level is higher than the participant's confidence level, and 0 otherwise) is not defined in the Without condition. "Algorithm" is a binary variable equal to 1 if the advisor is an algorithm, and 0 otherwise. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table 10: Random-effects logit panel regressions for "Post-Advice Success" in human-advisor conditions.

Dependent Variable	Success			
	(1)	(2)	(3)	(4)
Calibrated	-0.132**		-0.142**	
	(0.072)		(0.069)	
Overconfident		-0.062		-0.090
		(0.068)		(0.070)
High Disc.	0.341***	0.549***	0.331***	0.543***
	(0.043)	(0.038)	(0.043)	(0.038)
Calib. × High Disc.	0.255***		0.265***	
	(0.070)		(0.069)	
Over × High Disc.		-0.172		-0.165
		(0.108)		(0.107)
Pre-advice success	3.333***	3.331***	3.333***	3.331***
	(0.004)	(0.004)	(0.004)	(0.004)
Advisor more confident	0.025***	0.025***	0.025***	0.025***
	(0.002)	(0.002)	(0.002)	(0.002)
Sequence HL	-0.052	-0.053	-0.043	-0.046
	(0.064)	(0.064)	(0.066)	(0.067)
NARS			-0.007	-0.005
			(0.008)	(0.008)
CRT			0.005	0.005
			(0.018)	(0.018)
Age			0.111**	0.113**
			(0.050)	(0.051)
Male dummy			-0.066	-0.070
			(0.075)	(0.075)
Years of Study			-0.036	-0.026
			(0.036)	(0.036)
Constant	-2.749***	-2.781***	-2.659***	-2.749***
	(2.891)	(2.952)	(3.071)	(3.328)
Observations	36,720	36,720	36,720	36,720
Participants	153	153	153	153
Wald Chi2	1193	1182	1261	1236
p-value	0	0	0	0

Notes: Robust standard errors clustered at the participant level are reported in parentheses. The sample includes only trials with a human advisor. All specifications additionally control for series and round. "Post-Advice Success" is a dummy equal to 1 if the participant's second "left–right" decision is correct, and 0 otherwise. "Pre-Advice Success" is a dummy equal to 1 if the participant's first "left–right" decision is correct, and 0 otherwise. "Advisor more confident" is a dummy equal to 1 if the advisor's confidence level is higher than the participant's confidence level, and 0 otherwise. See Appendix D, Table 13, for the same model estimated on the pooled sample including both human and algorithmic advisors. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table 11: Random-effects panel logit regressions of "Post-Advice Success".

Dependent Variable	Success			
	(1)	(2)	(3)	(4)
Overconfident			-0.130***	-0.146***
			(0.051)	(0.052)
Calibrated	-0.065	-0.064		
	(0.048)	(0.048)		
High Disc.	0.331***	0.326***	0.502***	0.499***
	(0.029)	(0.030)	(0.028)	(0.028)
Over. × High Disc.			-0.124**	-0.122
			(0.070)	(0.070)
Calib. × High Disc.	0.208***	0.212***		
	(0.050)	(0.050)		
Algo Dummy	0.026	0.037	0.028	0.043
	(0.041)	(0.041)	(0.041)	(0.040)
Pre-advice success	3.185***	3.185***	3.183***	3.182***
	(0.004)	(0.004)	(0.004)	(0.004)
Sequence HL	-0.025	-0.026	-0.025	-0.026
	(0.043)	(0.044)	(0.043)	(0.044)
NARS (Social Influence)		-0.002		-0.002
		(0.005)		(0.005)
CRT		-0.007		-0.005
		(0.013)		(0.013)
Male dummy		-0.071		-0.081
		(0.049)		(0.050)
Years of Study		0.013		0.014
		(0.024)		(0.024)
Constant	-2.882***	-2.882***	-2.882***	-2.900***
	(2.482)	(2.857)	(2.393)	(2.764)
Observations	73,680	73,680	73,680	73,680
Participants	307	307	307	307
Wald Chi2	2034	2128	2035	2137
p-value	0	0	0	0

Notes: Robust standard errors clustered at the participant level are reported in parentheses. The sample includes trials with both human and algorithmic advisors. All specifications additionally control for series and round. "Post-Advice Success" is a dummy equal to 1 if the participant's second "left–right" decision is correct, and 0 otherwise. "Pre-Advice Success" is a dummy equal to 1 if the participant's first "left–right" decision is correct, and 0 otherwise. "Advisor more confident" is a dummy equal to 1 if the advisor's confidence level is higher than the participant's confidence level, and 0 otherwise. See Appendix D, Table 13, for the same model estimated on the pooled sample including both human and algorithmic advisors. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table 12: Random-effects panel logit regressions of "Switch" to advice as a function of participants' metacognitive skills.

Dependent variable	Switch					
	All		Well-calibrated		Overconfident	
	ALL		Weak Disc.	Strong Disc.	Weak Disc.	Strong Disc.
	(1)	(2)	(3)	(4)	(5)	(6)
Algorithm	0.109	0.110	-0.213	-0.044	0.163	0.424*
	(0.154)	(0.154)	(0.248)	(0.229)	(0.224)	(0.266)
Meta-d'	-0.233**	-0.472***	-0.378**	-0.327**	-0.724***	-0.566***
	(0.097)	(0.115)	(0.190)	(0.136)	(0.177)	(0.180)
Calibration	-0.115	-0.240***	-0.060	-0.093	-0.472***	-0.309**
	(0.081)	(0.088)	(0.155)	(0.157)	(0.130)	(0.142)
Advisor more confident	1.983***	1.983***	1.655***	2.332***	1.749***	2.439***
	(0.081)	(0.078)	(0.118)	(0.145)	(0.159)	(0.152)
Advisor more confident		0.364***	0.421***	0.342**	0.581***	0.202
× Meta-d'		(0.084)	(0.151)	(0.160)	(0.160)	(0.145)
Advisor more confident		0.203***	0.103	0.262	0.290***	0.228**
× Calibration		(0.068)	(0.124)	(0.168)	(0.109)	(0.113)
Constant	-1.973***	-1.981***	-2.226***	-1.390***	-2.056***	-1.871***
	(0.180)	(0.188)	(0.361)	(0.201)	(0.305)	(0.279)
Observations	19,478	19,478	4,678	4,986	4,644	5,170
Participants	205	205	102	102	103	103
Wald χ^2	658.6	714.5	228.9	330.6	178.1	358.7
p-value	0.000	0.000	0.000	0.000	0.000	0.000

Notes. Robust standard errors clustered at the participant level are reported in parentheses. The sample includes only trials with an algorithmic advisor in which the participant's initial "left–right" decision differs from the advisor's recommendation (disagreement trials). In columns (1) and (2), we include treatment-condition fixed effects. All specifications additionally control for series and round. "Switch" is a dummy equal to 1 if the participant follows the advice, and 0 otherwise. "Algorithm" is a dummy equal to 1 if the advisor is an algorithm, and 0 otherwise. "Advisor more confident" is a dummy equal to 1 if the advisor's confidence level is higher than the participant's confidence level, and 0 otherwise. Meta-d' and Calibration are standardized to make their coefficients comparable across specifications. Higher values indicate higher metacognitive ability. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table 13: Random-effects panel logit regressions of "Post-Advice Success" from an advisor on switching and metacognitive skills.

Dependent variable	Post-Advice Success					
	All		Well-calibrated		Overconfident	
	ALL		Weak Disc.	Strong Disc.	Weak Disc.	Strong Disc.
	(1)	(2)	(3)	(4)	(5)	(6)
Algorithm	0.018	0.001	0.049	-0.025	-0.015	-0.064
	(0.041)	(0.052)	(0.064)	(0.081)	(0.060)	(0.070)
Pre-Advice Success	1.170***	1.622***	1.468***	1.580***	1.242***	0.954***
	(0.104)	(0.195)	(0.204)	(0.257)	(0.206)	(0.246)
Switch	1.081***	1.027***	0.779***	1.577***	0.835***	0.993***
	(0.054)	(0.090)	(0.087)	(0.179)	(0.070)	(0.126)
Meta-d'	0.073***	0.080***	0.105***	0.071*	0.155***	0.066
	(0.001)	(0.030)	(0.030)	(0.042)	(0.029)	(0.041)
Switch		-0.090**	-0.129*	0.125	-0.090	0.028
× Meta-d'		(0.043)	(0.074)	(0.094)	(0.058)	(0.059)
Calibration	0.030**	0.047*	0.069***	0.157***	0.112***	0.117***
	(0.016)	(0.026)	(0.024)	(0.039)	(0.023)	(0.036)
Switch		-0.065	-0.242***	-0.097	-0.163***	-0.265***
× Calibration		(0.066)	(0.082)	(0.077)	(0.054)	(0.047)
Constant	-0.206***	-0.853***	-0.962***	-0.576***	-0.713***	-0.246
	(0.063)	(0.113)	(0.105)	(0.171)	(0.118)	(0.175)
Observations	29,157	9,679	4,678	4,986	4,644	5,170
Participants	307	102	102	102	103	103
Wald χ^2	133.7	165.9	150.9	151.9	194.1	230.3
p-value	0.000	0.000	0.000	0.000	0.000	0.000

Notes. Robust standard errors clustered at the participant level are reported in parentheses. The sample includes only trials with an algorithmic advisor in which the participant's initial "left–right" decision differs from the advisor's recommendation (disagreement trials). In columns (1) and (2), we include treatment-condition fixed effects. All specifications additionally control for series and round. "Post-Advice Success" is a dummy equal to 1 if the participant second "left–right" decision is correct, otherwise 0. "Pre-Advice Success" is a dummy equal to 1 if the participant first "left–right" decision is correct, otherwise 0. "Algorithm" is a dummy equal to 1 if the advisor is an algorithm, and 0 otherwise. "Switch" is a dummy equal to 1 if the participant follows the advice, and 0 otherwise. "Advisor more confident" is a dummy equal to 1 if the advisor's confidence level is higher than the participant's confidence level, and 0 otherwise. Meta-d' and Calibration are standardized to make their coefficients comparable across specifications. Higher values indicate higher metacognitive ability. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table 14: Random-effects panel logit regressions of "Switch" to algorithmic advice with controls.

Dependent variable	Switch					
	All		Well-calibrated		Overconfident	
	ALL		Weak Disc.	Strong Disc.	Weak Disc.	Strong Disc.
	(1)	(2)	(3)	(4)	(5)	(6)
Advisor more confident	1.814***	1.869***	1.496***	2.471***	1.223***	2.214***
	(0.114)	(0.111)	(0.156)	(0.212)	(0.231)	(0.192)
Meta-d'	-0.298**	-0.592***	-0.402	-0.473	-0.803***	-0.807***
	(0.101)	(0.145)	(0.330)	(0.291)	(0.203)	(0.206)
Advisor more confident		0.384***	0.423	0.504**	0.473**	0.283
× Meta-d'		(0.111)	(0.243)	(0.186)	(0.177)	(0.191)
Calibration	-0.057	-0.191	-0.140	-0.108	-0.350	-0.350
	(0.112)	(0.127)	(0.270)	(0.298)	(0.199)	(0.199)
Advisor more confident		0.178	0.218	0.434	0.007	0.051
× Calibration		(0.098)	(0.183)	(0.291)	(0.142)	(0.142)
Series HL	-0.425*	-0.453*	-0.701	-1.431**	0.203	0.639
	(0.208)	(0.211)	(0.597)	(0.469)	(0.458)	(0.515)
NARS	-0.010	-0.004	0.044	-0.028	0.049	0.013
	(0.020)	(0.021)	(0.040)	(0.040)	(0.043)	(0.049)
CRT score	-0.070	-0.073	-0.095	-0.092	-0.115	-0.166
	(0.065)	(0.067)	(0.116)	(0.114)	(0.111)	(0.129)
Age	-0.114	-0.114	0.163	0.732*	-0.863	-1.016
	(0.242)	(0.243)	(0.381)	(0.369)	(0.505)	(0.566)
Male	0.034	0.065	-0.317	-0.555	0.598	0.806
	(0.216)	(0.217)	(0.391)	(0.439)	(0.380)	(0.436)
Years of study	0.114	0.127	0.296	0.073	0.493	0.846**
	(0.101)	(0.103)	(0.184)	(0.198)	(0.261)	(0.314)
Round	0.002***	0.002***	0.007**	0.003	0.003	-0.001
	(0.001)	(0.001)	(0.003)	(0.002)	(0.002)	(0.002)
Constant	-2.733***	-2.830***	-3.079***	-1.146	-2.976***	-3.016**
	(0.385)	(0.373)	(0.783)	(0.613)	(0.784)	(1.061)
Observations	14,755	14,755	2,353	2,496	2,404	2,717
Participants	154	154	51	51	53	53
Wald χ^2	329.5	329.5	128.2	281.6	79.79	271.7
p-value	0.000	0.000	0.000	0.000	0.000	0.000

Notes. Entries are logit coefficients. Robust standard errors clustered at the participant level are reported in parentheses. The sample includes only trials with an algorithmic advisor in which the participant's initial decision differs from the advisor's recommendation (disagreement trials). Columns (1) and (2) additionally include treatment-condition fixed effects (shown as condition indicators). "Switch" is a dummy equal to 1 if the participant follows the advice, and 0 otherwise. "Advisor more confident" is a dummy equal to 1 if the advisor's confidence level is higher than the participant's confidence level, and 0 otherwise. Meta-d' and Calibration are standardized. Controls include NARS, CRT score, age, gender, education, series, and round. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table 15: Random-effects panel logit regressions of "Post-Advice Success" from an algorithmic advisor with controls.

Dependent variable	Post-Advice Success					
	All		Well-calibrated		Overconfident	
	ALL		Weak Disc.	Strong Disc.	Weak Disc.	Strong Disc.
	(1)	(2)	(3)	(4)	(5)	(6)
Pre-Advice Success	1.204***	1.535***	1.486***	1.406***	0.961**	0.498
	(0.175)	(0.281)	(0.306)	(0.403)	(0.309)	(0.356)
Switch	1.058***	1.039***	0.774***	1.557***	0.760***	0.832***
	(0.074)	(0.120)	(0.124)	(0.266)	(0.101)	(0.157)
Meta-d'	0.087***	0.050	0.081*	0.123*	0.162**	0.111
	(0.024)	(0.039)	(0.036)	(0.052)	(0.061)	(0.072)
Switch		-0.049	-0.117	0.228	-0.154	0.015
× Meta-d'		(0.066)	(0.097)	(0.147)	(0.095)	(0.086)
Calibration	0.038	0.075*	0.082	0.207*	0.131***	0.142*
	(0.021)	(0.034)	(0.050)	(0.087)	(0.039)	(0.058)
Switch		-0.149	-0.226	-0.060	-0.138*	-0.329***
× Calibration		(0.090)	(0.120)	(0.143)	(0.069)	(0.057)
Series HL	0.035	0.066	0.193	0.062	0.041	-0.050
	(0.042)	(0.083)	(0.175)	(0.144)	(0.149)	(0.182)
NARS	-0.009*	-0.005	-0.010	-0.011	-0.012	0.008
	(0.005)	(0.010)	(0.009)	(0.008)	(0.011)	(0.012)
CRT score	-0.002	-0.012	-0.046	0.034	-0.010	0.012
	(0.014)	(0.022)	(0.034)	(0.035)	(0.026)	(0.039)
Age	-0.012	-0.045	0.114	0.060	-0.044	-0.145
	(0.047)	(0.077)	(0.097)	(0.101)	(0.078)	(0.098)
Male	0.006	-0.018	0.001	0.012	0.137	0.040
	(0.046)	(0.053)	(0.120)	(0.132)	(0.082)	(0.146)
Years of study	-0.014	0.044	-0.026	-0.142**	-0.061	0.021
	(0.024)	(0.036)	(0.054)	(0.048)	(0.043)	(0.059)
Round	0.000	0.000	0.002	0.001	0.001	0.000
	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
Constant	-0.543***	-0.853***	-1.061***	-0.313	-0.483	-0.147
	(0.119)	(0.174)	(0.295)	(0.257)	(0.332)	(0.319)
Observations	14,755	4,785	2,353	2,496	2,404	2,717
Participants	154	50	51	51	53	53
Wald χ^2	372.5	122.8	81.77	100.6	117.6	229.9
p-value	0.000	0.000	0.000	0.000	0.000	0.000

Notes. Robust standard errors clustered at the participant level are reported in parentheses. "Post-Advice Success" is a dummy equal to 1 if the participant's second decision is correct, and 0 otherwise. "Pre-Advice Success" is a dummy equal to 1 if the participant's first decision is correct, and 0 otherwise. "Switch" is a dummy equal to 1 if the participant follows the advice, and 0 otherwise. Meta-d' and Calibration are standardized. Controls include NARS, CRT score, age, gender, education, series, and round. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

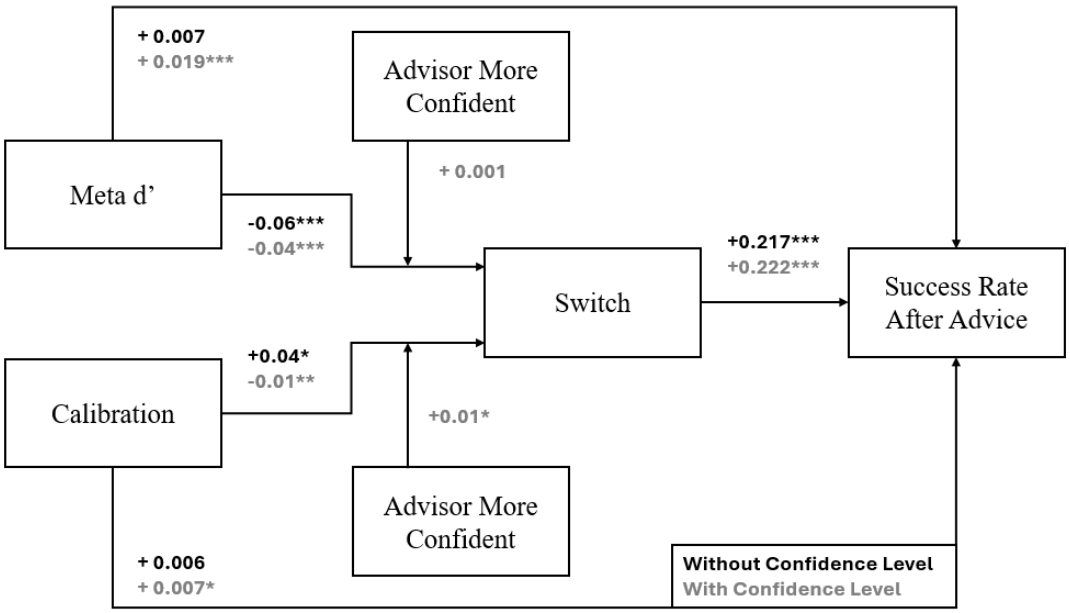


Figure 17: **Causal path linking metacognitive ability, switch, and Success Rate after the Advice.** The figure summarizes the estimated mediation structure on disagreement trials, i.e., cases where the participant's initial "left–right" choice differs from the advisor's recommendation. All relationships are estimated using pooled logit models with standard errors clustered at the participant level. Path coefficients are reported as average marginal effects (AME), and the Switch link is reported as a risk difference (RD) from predicted probabilities at Switch = 1 vs. Switch = 0. Total, indirect, and direct effects (TE/NIE/NDE) for a +1 SD change in meta-d' or calibration are obtained by G-computation with a participant-cluster bootstrap (1000 repetitions). In the Without confidence condition, the estimated total effects are $TE_{\text{meta-}d'} = -0.006$ (n.s.) and $TE_{\text{calibration}} = 0.010$ ($p < 0.05$). In condition with a confidence level, they are $TE_{\text{meta-}d'} = 0.004$ (n.s.) and $TE_{\text{calibration}} = -0.010$ ($p < 0.05$). The outcome, "post-advice success rate", is regressed on "pre-advice success ", Switching (Switch = 1 if the participant follows the advisor, 0 otherwise), participants' standardized calibration and meta-d' scores, "Advisor type" (Algo = 1 for algorithm, 0 for human), and controls for series order and round. In the "Without confidence" baseline condition, switching is modeled as a function of the same covariates as the outcome equation, excluding 'pre-advice success ". In the With confidence conditions, we additionally include an "Advisor More Confident" indicator (= 1 if the advisor reports higher confidence than the participant, 0 otherwise) as a mediator between calibration/meta-d' and switching, capturing whether the advisor's stated confidence exceeds the decision maker's confidence. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

DOCUMENTS DE TRAVAIL GREDEG PARUS EN 2026
GREDEG Working Papers Released in 2026

2026-01 Estelle Malavolti & Frédéric Marty
L'équilibre économique des aéroports secondaires européens à l'épreuve des dynamiques de concurrence fiscale

2026-02 Charlie Joyez
A New Index of Export-Import Proximity: Conceptual Foundations and Global Patterns

2026-03 Patrice Bougette & Frédéric Marty
The SCP Paradigm Revisited: What Structuralism Really Contributed to U.S. Antitrust

2026-04 Imen Bouhlel, Nathalie Lazaric & Paolo Zeppini
Competitive Diffusion and Sustainability Transitions: The Case of Plastics Recycling Technologies

2026-05 Maxime Menuet
Fractional Replicator Dynamics

2026-06 Ilona Dielen, Patrice Bougette & Christophe Charlier
A (Green) Switch in Time Saves Nine: Assessing the Environmental Damage of the European Truck Cartel

2026-07 Thibault Schrepel & Godefroy de Boiscuillé
Questioning the Digital Markets Act's Legality

2026-08 Michael Finus & Paolo Zeppini
Green Lifestyles and Social Tipping Points

2026-09 Mathieu Chevrier & Sébastien Massoni
When Does Advisor Confidence Improve Decisions? Evidence from Human and Algorithmic Advice