University of Cyprus
Department of Economics

Working Paper 05-2025

Understanding the Source of Algorithmic Aversion: An Experimental Approach

Elia Antoniou

Department of Economics, University of Cyprus, P.O. Box 20537, 1678 Nicosia, Cyprus
www.ucy.ac.cy/econ/?lang=en

Understanding the Source of Algorithmic Aversion: An Experimental Approach[*]

Elia Antoniou[†]

November 25, 2025

Abstract

This paper examines the source of algorithmic aversion, defined as the unwillingness to accept advice or decisions made by algorithms. Algorithmic aversion is conceptualized as a form of individual partiality, driven either by belief-based factors attributed to differences in perceived ability, or by preference-based factors which reflect a disamenity associated with selecting algorithms. To empirically test the predictions of the model, a preregistered online experiment was conducted, where participants evaluated answers to objective and subjective economic questions, with varying information on whether the source was human or algorithm. The results provide no evidence of algorithmic aversion in the evaluation task: participants did not systematically favor human-generated answers over algorithm-generated ones. These results suggest that algorithmic aversion may not be as robust or uniform as previously assumed.

Keywords: Algorithmic Aversion; Human-AI interaction; Experimental Economics.

JEL Classification: D83, D90, C91

[*] This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement number: 857636 - SInnoPSis - H2020-WIDESPREAD-2018-2020/H2020-WIDESPREAD-2018-04. This project has also received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No [101079196] (Twinning for Excellence in Management and Economics of Research and Innovation [TWIN4MERIT]). This study has been approved by the Cyprus National Bioethics Committee (CNBC DECISION2023.01.104), and registered as https://doi.org/10.17605/OSF.IO/M9UD7. I am grateful to Zacharias Maniadis, Philippos Louis, as well as attendees of the Behavioral and Experimental Economics Lab in Florence, Italy, and the 2025 International Meeting on Experimental and Behavioral Social Sciences in Valencia, Spain for their insightful feedback.

[†] University of Cyprus, SInnoPSis. antoniou.elia@ucy.ac.cy

1 Introduction

Since the widespread adoption of algorithms, Artificial Intelligence (AI) systems and large language models (LLMs), people often have negative attitudes when taking advice, delegating tasks to, or accepting outcomes performed by algorithms.[1] This biased assessment against algorithms, which results in negative behaviors and attitudes toward algorithms when compared to human agents, is referred to as 'Algorithmic aversion' (Dietvorst, Simmons, and Massey, 2015).

Algorithmic technologies can provide substantial benefits to individuals, such as increased efficiency, improved decision-making, and access to expertise in many domains (Chak, Croxson, D'Acunto, Reuter, Rossi, and Shaw, 2022). Realizing the full benefits of AI systems requires not only access to these tools, but also a willingness to engage with and utilize them. Technological progress requires not only the development of new technologies, but also their adoption, as both are key drivers of technological advancement and economic growth (Romer, 1990). Fully understanding the concept of algorithmic aversion and identifying its sources can provide valuable insights into how people perceive and respond to advice from algorithms. It could also provide insights into how challenging it is to change people's perceptions of algorithms.

Although this topic has been explored in the fields of psychology, information systems, and computer science, the source of algorithmic aversion remains unclear (Jussupow, Benbasat, and Heinzl, 2020). Germann and Merkle (2023) were the first to conduct an incentivised experiment within economics, reporting no evidence of algorithmic aversion. They find that participants exhibited no strong preference regarding whether the financial intermediary was human or algorithm, focusing primarily on investment returns. In behavioral and experimental economics, Ivanova-Stenzel and Tolksdorf (2024) show that providing information about algorithmic performance increases participants' willingness to rely on algorithms, although algorithms remain systematically underutilized. Buchanan and Hickman (2024) find that when authorship is disclosed, participants express equal

[1] The term "algorithm" is used broadly to refer to algorithmic systems, AI models, and large language models (LLMs).

skepticism toward both human and AI writers. The authors comment that trust in content generated by AI seems to be context-dependent. Finally, in the context of delegation, Gogoll and Uhl (2018) demonstrate that people are hesitant to delegate moral decisions to a machine.

The preceding evidence highlights substantial heterogeneity across domains and experimental designs. The lack of a unified theoretical framework makes it challenging to gain a comprehensive understanding of the concept, including its characteristics, preceding factors, and implications (Mahmud, Islam, Ahmed, and Smolander, 2022). Building on this literature, the present study adopts an experimental economics framework to investigate the source of algorithmic aversion.

The primary aim of the paper is to address this gap by providing a theoretical framework that explores the underlying sources of algorithmic aversion. Within a formal economic framework, algorithmic aversion is conceptualized as a form of individual partiality. The model assumes two possible sources of algorithmic aversion, defined as belief-based partiality and preference-based partiality. The former is driven by individuals' beliefs about the average ability of algorithms, while the latter is modeled as a disamenity term in the utility function. By varying key parameters of the model, the effects on algorithmic aversion differ, allowing for the identification of its underlying source. The research question of this paper is what is the underlying source of algorithmic aversion and under what conditions it arises.

The paper includes five hypotheses, three of which are derived from the theoretical predictions of the model. The first hypothesis proposes that algorithmic aversion arises when there is uncertainty about ability, leading individuals to rely on subjective beliefs about the ability of algorithms. The second hypothesis suggests that algorithmic aversion decreases as information about quality becomes more precise, since individuals rely less on subjective beliefs. The third hypothesis predicts that if algorithmic aversion is observed under uncertainty, when subjective beliefs play a larger role, but disappears when information about ability becomes more objective, then this aversion is driven by belief-based factors related to differences in perceived ability. The remaining two hypotheses are based on prior empirical evidence. The fourth and fifth hypotheses examines whether trust in algorithms and demographic factors such as gender, age, and education increases evaluations of algorithms.

To empirically test the predictions of the model, a preregistered online experiment was conducted using a between-subject design. By generating a micro-level dataset, this study helps fill a gap in the literature of technology adoption (Comin and Mestieri, 2014). In the experiment, participants acted as evaluators, assessing the correctness of five economic answers that were either objective or subjective. Participants were informed that the answers had been provided by either human economists or AI Chatbots, and were asked to judge whether each answer was correct or incorrect. This task allows for a direct test of the model's predictions by linking the level of subjectivity in judgment to the source of algorithmic aversion.

The experimental results suggest that, in the sample, participants did not systematically favor answers provided by humans as opposed to those provided by algorithms. One possible explanation for this finding is that participants may perceive humans and algorithms as equally capable, indicating the absence of belief-based partiality, and may also lack a clear preference for responses provided by humans, suggesting an absence of preference-based partiality.[2]

The paper proceeds as follows. Section 2 provides a literature review, Section 3 presents the theoretical model, Section 4 the experimental design and Section 5 the experimental results, while Section 6 concludes. Proofs are in Appendix A.

2 Literature Review

An algorithm is defined as a set of rules or instructions, usually in the form of computer code, designed to process input and generate output, either through rule-based or machine learning methods (Jussupow et al., 2020). The term 'Algorithmic aversion' was first popularized by Dietvorst et al. (2015). It describes a concept of people being

[2] Theoretically, as shown in Section 3 of the paper, belief-based and preference-based partiality could cancel each other out, but the data provide no evidence of such an effect.

averse to algorithmic forecasters after observing their performance, even when these algorithms outperform human forecasters. A broader definition of algorithmic aversion was later introduced, suggesting that people are unwilling to use superior algorithms that are perceived as imperfect (Dietvorst, Simmons, and Massey, 2018). Algorithmic aversion is also defined as "the tendency of individuals to discount computer-based advice more heavily than human advice, although the advice is identical otherwise" as proposed by Joe, Commerford, Dennis, and Wang (2019).

While early studies report consistent evidence of algorithmic aversion (Castelo, Bos, and Lehmann, 2019; Dietvorst and Bharti, 2020; Dietvorst et al., 2015; Jussupow et al., 2020), later studies report mixed or null effects (David and Sade, 2018; Jussupow et al., 2020). For instance, Schaap, Bosse, and Hendriks Vettehen (2024) reported little evidence of algorithmic aversion in the finance and dating domains. Conversely, some papers report the opposite of algorithmic aversion, referred to as 'algorithmic appreciation', which is defined as positive behaviors and attitudes toward algorithms compared to human agents (Holzmeister, Holmén, Kirchler, Stefan, and Wengström, 2023; Logg, Minson, and Moore, 2019).

The occurrence of algorithmic aversion is shaped by several factors. The lack of trust is one of the primary reasons people are unwilling to rely on algorithmic advice (Buchanan and Hickman, 2024; Chak et al., 2022; David and Sade, 2018). Trust in content generated by artificial intelligence (AI) is context-dependent, and people's willingness to rely on or trust AI can be affected by factors such as framing and perception of autonomy (Buchanan and Hickman, 2024). Algorithmic aversion can also be explained by factors related to both algorithmic and human characteristics, such as algorithmic agency or performance,[3] as well as human expertise and social distance (Jussupow et al., 2020).[4] Task characteristics also matter, as people are less willing to trust algorithms for subjective tasks, perceiving them as lacking human-like abilities (Castelo et al., 2019; Dijkstra, Liebrand, and Timminga, 1998).[5] In contrast, trust in algorithms increases as

[3] Algorithmic agency is the capacity of algorithms to operate and make decisions autonomously.

[4] Social distance is the perceived degree of closeness between individuals, which influences their willingness to interact.

[5] Objective tasks can be measured and quantified (e.g., buying stocks), while subjective tasks are open

tasks become more objective (Castelo et al., 2019). Task subjectivity therefore plays a key role, contributing to a decreased use of algorithms in tasks that are perceived as subjective (Mahmud et al., 2022).

Performance and error sensitivity influence how individuals interact with and assess algorithmic systems. People are less likely to rely on algorithmic forecasters after witnessing them making mistakes, even when the algorithm outperforms humans (Dietvorst et al., 2018). They often overreact and abandon the algorithm after it errs, even when its mistakes are similar or fewer than those made by humans (Dietvorst et al., 2015). People's rejection of algorithms may also originate from their reduced sensitivity to forecasting errors (Dietvorst and Bharti, 2020). Dietvorst and Bharti (2020) suggest that there is a concave relationship between the magnitude of errors and people's sensitivity to them. As errors increase, sensitivity diminishes, leading people to favor human forecasters over algorithms, despite the higher risk and lower accuracy. This effect is more pronounced when people believe the algorithm will not provide a near-perfect answer.

The drivers of algorithmic aversion identified in prior research can be linked to the two types of partiality in the theoretical framework of this paper. Belief-based partiality captures biases arising from perceptions of algorithmic ability, such as trust deficits, sensitivity to errors, and lower acceptance in subjective tasks. Preference-based partiality reflects a general dislike or discomfort that persists even when performance is objectively equivalent. This mapping connects existing empirical findings to the mechanisms in the model.

While the above theories suggest that people engage in conscious reflective processes when they evaluate algorithms, other theories focus on implicit biases, such as prejudice, which shape behavior unconsciously (Dovidio, Kawakami, and Gaertner, 2002; Serenko and Turel, 2021). Using implicit association tests (IATs), Turel and Kalhan (2023) found that, on average, people have an implicit bias against AI, perceiving it as untrustworthy. Empirical evidence suggests that this bias is more pronounced when AI is compared to human experts than with individuals similar to oneself.

to interpretation and based on opinion or intuition (e.g., recommending jokes).

The adoption of algorithms can yield significant benefits, such as mitigating the impact of heuristics and biases in human decision-making. For example, robo-advice in loan repayment decisions has been shown to improve repayment outcomes and reduce inequality, especially among financially vulnerable households (Chak et al., 2022). Therefore, the accessibility and scalability of such tools make them important for improving decision quality.

Despite extensive research in psychology, information systems, and computer science, algorithmic aversion has received limited attention in economics, with only a few studies addressing it (Chak et al., 2022; Gogoll and Uhl, 2018). This paper represents one of the first attempts to examine algorithmic aversion from an economic angle. The paper contributes to the literature by applying an economic model to conceptualize algorithmic aversion as a form of individual partiality, which can be belief-based or preference-based. The paper further contributes to the literature by assessing the broader applicability of the theoretical model in a different context. The aim of the paper is to enhance both the theoretical and empirical understanding of algorithmic aversion.

3 Methodology

The theoretical model considers evaluators and workers.[6] In this model, evaluators assess the output of workers after observing their group identity and a signal about quality.

3.1 Model

Worker The model considers a set of finite workers. Each worker has an observable group identity, denoted by $g \in \{H, A\}$, where H is human and A is an algorithm. The worker has an unobservable ability level, denoted by $\alpha \sim N(\mu_g, 1/\tau_a)$ with $\mu_g \in R$ and precision $\tau_a > 0$. The mean of ability depends on the worker's group identity, while the variance is the same.[7] Each worker completes a sequence of tasks $t = 1, 2, \ldots, n$, where

[6] The theoretical model of this paper is an extension of the model developed by Bohren, Imas, and Rosenberg (2019) to study gender discrimination.

[7] This ensures that the distinction between the two identity groups is determined solely by differences in the mean ability, and not by differences in the variance. By keeping τ_a constant any observed difference

each task has a hidden quality, denoted by $q_t = \alpha + \epsilon_t$, with $\epsilon_t \sim N(0, 1/\tau_\epsilon)$ an independent random shock with precision $\tau_\epsilon > 1$. Ability is fixed over time for each individual.

Evaluator There is a set of evaluators who assess the performance of the workers. To keep things simple, assume that there is only one evaluator per task who reports a binary evaluation, denoted by $v_t \in \{0, 1\}$. For simplicity, evaluators are indexed by tasks, such that index t denotes both the task and the evaluator assigned to it.

Before reporting the evaluation for each task, the evaluator observes the worker's group identity $g \in \{H, A\}$, and a signal of quality for the current task denoted by $s_t = q_t + \eta_t$, where $\eta_t \sim N(0, 1/\tau_\eta)$ is an independent random shock with precision $\tau_\eta > 0$. The precision of the signal reflects the evaluator's subjectivity in judgment, with lower signal precision indicating greater subjectivity in judgment. Specifically, lower signal precision increases the evaluator's uncertainty regarding the worker's ability and, consequently, the quality of the worker's output. This uncertainty leads to an increase in the subjectivity of the judgment, as the evaluator relies more on his subjective beliefs due to less precise information.

Each evaluator has a subjective prior belief about the average ability of workers in group g, represented by $\hat{\mu}_g^t$. In addition, each evaluator has a type-specific taste parameter, denoted by c_g^t, which can be positive or negative depending on the evaluator's preferences toward group g. If this taste parameter is positive, the evaluator has a disamenity value associated with tasks performed by a worker with identity g. For example, when $c_A^t > 0$ the evaluator has a disamenity value from tasks performed by algorithms. Formally, the type of evaluator is defined as $\theta_t \equiv (\hat{\mu}_g^t, c_g^t)$, which summarizes both the evaluator's subjective prior beliefs and preference parameters.

Evaluations Evaluations are binary and the evaluator can downvote or upvote a task, $v_t \in \{0, 1\}$. If the evaluator downvotes the task, the valuation is defined as $v_t = 0$, whereas if the evaluator upvotes the task, the valuation is defined as $v_t = 1$. The evaluator must

in the evaluation of the worker can be attributed to either the evaluator's beliefs about average ability or their preferences. A common value of τ_a across groups reflects a fixed level of uncertainty with regards to the abilities of both group identities.

choose one of the two options for each task. Under these conditions, the evaluator receives a payoff of $q_t - c_g^t$ when upvoting a task performed by a worker with identity g and quality q. In contrast, the evaluator receives a payoff of 0 from downvoting a task. Formally, the evaluator's payoff is defined as:

$$v_t(q_t, c_g^t) = \begin{cases} q_t - c_g^t, & \text{if } v_t = 1 \\ 0, & \text{if } v_t = 0 \end{cases}$$

The type-specific taste parameter (c_g^t) enters the payoff function with a negative sign, so that a positive value represents a disamenity toward group g, whereas a negative value reflects a positive utility from tasks performed by that group. Hence, this parameter can take either positive or negative values.

Definitions An evaluator of type θ_t exhibits belief-based partiality when his evaluation is influenced by subjective beliefs about the distribution of ability across identity groups. In this case, the evaluator has belief-based partiality toward humans if he or she believes that the average ability of humans is higher than the average ability of algorithms.[8]

Definition 1 (Belief-Based Partiality): An evaluator of type θ_t has belief-based partiality toward humans if $\hat{\mu}_H^t > \hat{\mu}_A^t$.

Preference-based partiality arises when an evaluator's assessment is influenced by personal preferences or disamenities toward a specific group. In other words, the evaluator may exhibit a preference for one identity group while experiencing a disamenity associated with tasks performed by the other group. An evaluator has preference-based partiality toward humans if the taste parameter for algorithms is higher than the one for humans.

Definition 2 (Preference-Based Partiality): An evaluator has preference-based partiality toward humans if $c_A^t > c_H^t$.

[8] Belief-based partiality is modeled through the evaluator's beliefs about the average ability. Evaluators may form stereotypes using heuristics in probability judgment, particularly through the use of the representativeness heuristic (Bordalo, Coffman, Gennaioli, and Shleifer, 2016; Tversky and Kahneman, 1983).

Aggregate analogues of partiality are taken with respect to the average beliefs and preferences of evaluators. There is aggregate belief-based partiality toward humans if the average belief about human ability exceeds the average belief about the ability of algorithms, i.e., $E[\hat{\mu}_H^t] > E[\hat{\mu}_A^t]$. There is aggregate preference-based partiality toward humans if the expected taste parameter for algorithms is higher than that for humans, i.e., $E[c_A^t] > E[c_H^t]$.

Subjectivity of Judgment The precision of the signal reflects the level of subjectivity in judgment involved in the evaluation of a worker. An increase in the precision of the signal generates lower uncertainty about quality. Specifically, since the signal provides information about quality ($s_t = q_t + \eta_t$ with $\eta_t \sim N(0, 1/\tau_\eta)$), as τ_η increases, the variance of the signal decreases, thereby reducing uncertainty about the signal.

3.1.1 Optimal Evaluation

An evaluator of type θ_t chooses his optimal evaluation by maximizing his expected payoff with respect to his posterior belief about quality. Let $v_t(s_t, g)$ be the optimal evaluation and $\hat{E}_t[q_t - c_g^t|s_t, g]$ the evaluator's expectation in the case of an upvote. The optimal evaluation conditional on the group identity of the worker and the observed signal, is defined as:

$$v_t(s_t, g) \equiv \arg \max_{v_t \in \{0,1\}} \hat{E}_t[q_t - c_g^t|s_t, g]. \tag{1}$$

Decision Rule The evaluator maximizes his expected payoff by choosing to upvote the task ($v_t = 1$) if and only if his expectation of the posterior distribution of quality conditional on the signal and the group identity (s_t, g), is higher than his type-specific taste parameter associated with the specific worker group. In other words, when:

$$\hat{E}_t[q_t|s_t, g] \geq c_g^t, \tag{2}$$

where the expectation is taken with respect to the posterior distribution of quality. It is important to note that $\hat{E}[q|s, g]$ is strictly increasing in the signal, since $f_{s|q}$ satisfies the monotone likelihood ratio property (MLRP) with respect to quality. This property implies that as quality increases there is higher probability of observing a higher signal.

3.1.2 Algorithmic Aversion

Suppose the evaluator is at the first task ($t = 1$), and has no information on prior evaluations.[9] The evaluator has subjective prior beliefs about the average ability of workers $(\hat{\mu}_H^t, \hat{\mu}_A^t)$, and type-specific taste parameters (c_A^t, c_H^t) that capture preferences toward each group. Lastly, the evaluator observes a signal of quality.

The evaluator's prior belief about quality is normally distributed with mean $\hat{\mu}_g^t$ and precision $\tau_q \equiv \tau_a\tau_\epsilon/(\tau_a + \tau_\epsilon)$, expressed as $q_1 \sim N(\hat{\mu}_g^t, 1/\tau_q)$. The signal follows a conditional distribution, expressed by $s_1/q_1 \sim N(q_1, 1/\tau_\eta)$. Given the prior belief and signal distribution, the evaluator's posterior belief about quality conditional on observing the signal, is also normally distributed, denoted by $q_1/s_1 \sim N(\frac{\tau_q\hat{\mu}_g^t+\tau_\eta s_1}{\tau_q+\tau_\eta}, \frac{1}{\tau_q+\tau_\eta})$. All derivations are provided in Appendix A.1. Using Equation 2 the evaluator reports an upvote on the first task if the expected quality conditional on the signal and the group identity of the worker, is greater than the type-specific taste parameter:

$$\hat{E}_t[q_1|s_1, g] = \frac{\tau_q\hat{\mu}_g^t + \tau_\eta s_1}{\tau_q + \tau_\eta} \geq c_g^t. \tag{3}$$

Algorithmic aversion is defined as the difference in evaluations between a human and an algorithm, at the same signal, defined as:

$$AlgoAversion_t(s_t) \equiv v_t(s_t, H) - v_t(s_t, A) > 0 \tag{4}$$

Definition 3 (Algorithmic Aversion): An evaluator exhibits algorithmic aversion if $\text{AlgoAversion}_t(s_t) > 0$. At the aggregate level, there is algorithmic aversion if $\text{AlgoAversion}(s_t) > 0$.

[9] Although the derivation is presented for the first task, for simplicity the notation retains the general index t to emphasize that the expressions apply to any task.

Combining Equations 3 and 4, the aggregate formal expression for algorithmic aversion is given by:

$$AlgoAversion(s_t) = (\frac{\tau_q}{\tau_q + \tau_\eta})(\hat{\mu_H} - \hat{\mu_A}) + (c_A - c_H). \tag{5}$$

There is algorithmic aversion ($AlgoAversion(s_t) > 0$) if the evaluator exhibits either belief-based partiality, i.e., $\hat{\mu_H} > \hat{\mu_A}$, or preference-based partiality, i.e., $c_A > c_H$, or both. In other words, Equation 5 is positive when the evaluator believes that the average ability of humans is higher than the average ability of algorithms, or when the disamenity value associated with tasks performed by algorithms is higher than the one for humans, or when both conditions hold. See Appendix A.2 for the derivations.

In contrast, there is algorithmic appreciation when ($AlgoAversion(s_t) < 0$). This occurs when the evaluator believes that the average ability of algorithms is higher than the one of humans $\hat{\mu_A} > \hat{\mu_H}$, or when the disamenity value associated with tasks performed by humans is higher than that for algorithms $c_H > c_A$, or when both conditions hold.[10]

No algorithmic aversion ($AlgoAversion(s_t) = 0$) may occur in three possible conditions: (i) when beliefs and preferences are equal across groups ($\hat{\mu_H} = \hat{\mu_A}, c_H = c_A$), (ii) when there is belief-based partiality toward humans but preference-based partiality toward algorithms ($\hat{\mu_H} > \hat{\mu_A}, c_A > c_H$), and (iii) when there is belief-based partiality toward algorithms but preference-based partiality toward humans ($\hat{\mu_A} > \hat{\mu_H}, c_H > c_A$). In cases (ii) and (iii), belief-based and preference-based partialities offset each other, resulting in no algorithmic aversion.

Source of Algorithmic Aversion The level of subjectivity in judgment (τ_η), modeled as the precision of the signal of quality, is an important parameter in identifying the source of algorithmic aversion. Changes in this parameter affect algorithmic aversion differently, depending on whether algorithmic aversion arises from preference-based or belief-based partiality. Focusing on algorithmic aversion as established in Equation 5, if there is an increase in the precision of the signal, in other words as τ_η increases, then the signal provides more precise information about quality. As the evaluator relies less on his

[10] Algorithmic appreciation may occur when either one of these effects exists, or when both exist simultaneously, as long as their combined effect is large enough to make the total expression negative.

subjective beliefs about average ability, these beliefs have a smaller impact on evaluations. As a result, algorithmic aversion will decrease, if it is driven by belief-based partiality since the precision (τ_η) is multiplied with the difference in beliefs about the average ability of the two groups. As the signal becomes perfectly objective (i.e., $\tau_\eta \to \infty$) then the first term in Equation 5 is eliminated and algorithmic aversion converges to the preference parameters (i.e., $AlgoAversion(s_t) \to (c_A - c_H)$). In this case, there is algorithmic aversion only if the evaluator has preference-based partiality and the disamenity value towards algorithms is positive and higher than that for humans (i.e, $c_A > c_H$).

Observation 1: If algorithmic aversion is caused by belief-based partiality, it decreases as the precision of the signal increases, otherwise it remains constant with respect to the precision of the signal. When the signal becomes perfectly objective, and quality is fully observable, there is algorithmic aversion only if the source is preference-based partiality.[11]

4 Experimental Design and Procedures

4.1 Experimental Design

The experiment is designed to evaluate the predictions of the model by testing whether algorithmic aversion exists and, if so, determining its underlying source. The experiment is preregistered on the Open Science Framework, under the number https://osf.io/m9ud7.

The experiment consisted of three parts. In the first part, participants evaluated the correctness of answers to five economic questions. In the second part, participants indicated their level of trust in algorithms. In the third part, participants provided information about their individual characteristics and answered questions related to financial and computer literacy. The full set of experimental screens can be found in Appendix C.

Participants in the main task of the experiment acted as evaluators, assessing answers to either objective or subjective economic problems. To match the empirical model of the paper, evaluations were binary, and therefore participants were asked to vote whether an

[11] Observation 1 holds under the assumption that evaluators update their beliefs according to Bayes' rule.

answer was correct or incorrect. The experiment followed a 2x3 between-subject design. The first experimental factor was the type of question (subjective, objective). Participants were randomly assigned to one of the two treatments, where in Treatment 1, evaluations involved only objective questions, while in Treatment 2, evaluations involved only subjective questions. The second experimental factor was the identity of the answer provider (human, algorithm, control). Participants were provided with information about the source of the answers based on their assigned treatment. They could read the content of all available answers and observe the selected answer. In the control treatment, participants were not given any information about the source. This group was used to assess the effect of revealing the provider's identity. Each participant evaluated content from only one type of answer provider. The answers shown to participants were identical across all three source treatments.

Table 1 shows the experimental allocation of participants across treatments. Randomization was implemented at the individual level. The order in which questions were presented to participants was also randomized, to prevent any anchoring effects from initial impressions formed by the first answer. Additionally, to avoid the formation of reputation-based beliefs, participants were informed that the questions were selected from a set of responses previously provided by economists or generated by algorithms, depending on their assigned treatment.

Table 1: Experimental Treatment

Treatment 1 (Objective)	Treatment 2 (Subjective)
Human	Human
Algorithm	Algorithm
Control	Control

At the start of the experiment, participants received on-screen instructions about the parts of the experiment and how they can earn money. They were informed that their task was to evaluate whether an answer was correct or incorrect, and they were instructed that their goal was to accurately make evaluations. To prevent learning or feedback effects, participants in the experiment did not receive feedback on their performance. The experiment was incentivised, and participants' final payoff included a fixed participation fee of £1.50 and an additional payoff based on their evaluation. Specifically, participants could earn an additional fee of £0.50 for each correct evaluation on the correctness of an answer. The total payoff and the correct answers were displayed on the screen at the end of the experiment.

The multiple-choice questionnaire used in the experiment was developed and validated by Alysandratos, Boukouras, Geōrganas, and Maniadis (2020).[12] The questionnaire included questions related to economic issues within a real-world context. The authors propose that, in general, people should be able to infer the correct answer using economic reasoning rather than a concrete mathematical formula. For the purposes of the experiment, the original questionnaire was separated into two questionnaires, with one containing five objective questions and the other containing five subjective questions. Objective questions have clear mathematical answers and are more objective by nature, while subjective questions require judgment under uncertainty and involve policy or opinion-based outcomes. Each question has only one correct answer, which was validated based on a 70% consensus among academic economists. See Appendix B.1 for the questionnaires used in each treatment.

In both the objective and subjective treatments, three answers shown to participants were correct and two were incorrect. This distribution was identical across the human and algorithm conditions. For the human condition, answers were drawn from responses provided by academic economists, as reported by Alysandratos et al. (2020). For the algorithm condition, answers were selected from a pool of responses generated by AI Chatbots (such as ChatGPT-3, Google Bard, Claude, You.com, and Microsoft Copilot) to match the human distribution. Both correct and incorrect answers were included. Appendix B.2 outlines the procedure used to determine the selected answers.

[12] The authors used this questionnaire to empirically investigate the persuasiveness of expert advice compared to populist advice.

Connection with the Theoretical Model In the experiment, algorithmic aversion is reflected in the voting behavior of participants and it is measured as the difference in net votes between human and algorithm generated answers. To facilitate this, a new variable called 'Net Votes' is constructed, representing the difference between the number of answers evaluated as correct and those evaluated as incorrect. Participants drew inferences from the worker's group identity and additionally observe the content of each question along with the possible answers. The content of the answer and the related question served as the signal of quality.

Participants were expected to have subjective beliefs about the average ability of algorithms, captured by the parameter $\hat{\mu}^t_A$, as suggested by the theoretical model. These beliefs reflect participants' perceptions of the internal workings and capabilities of algorithms. Likewise, participants were expected to have subjective beliefs about the average ability of humans and the accuracy of their answers, denoted with $\hat{\mu}^t_H$. As suggested in the literature on trust and task subjectivity (Castelo et al., 2019; Mahmud et al., 2022), subjective tasks involve higher uncertainty about quality. This leads participants to rely more heavily on their subjective beliefs, which are shaped by the worker's group identity. On the contrary, objective task reduce uncertainty about quality, thereby increasing the objectivity of judgment.

Research Hypotheses

All hypotheses for this study were preregistered. Hypotheses 1 to 3 are derived from the theoretical model of the paper. Hypothesis 1 is driven by the fact that evaluations on subjective answers involve greater uncertainty about quality, which increases the subjectivity of judgment. Participants are more likely to rely on their subjective beliefs about the average ability of the answer provider, which depends on the group identity, either human or algorithm. In the treatment with subjective questions and answers, algorithms are expected to receive fewer 'Net Votes' than humans, indicating the presence of algorithmic aversion.

Hypothesis 1: Algorithms will receive significantly fewer 'Net Votes' than humans in the treatment with the questions that have subjective answers.

On the contrary, evaluations on objective questions and answers, involve lower uncertainty about quality, as it is easier to infer the correct answer. Therefore, evaluations in this context are less likely to depend on the identity of the answer provider and more likely to reflect the actual correctness of the response. This type of questions reduce the subjectivity of judgment, thus, no algorithmic aversion is expected to occur on objective questions. Hypothesis 2, using an equivalence test (Lakens, 2017), suggests that there will be no meaningful difference in 'Net Votes' received by humans and algorithms in the objective treatment.

Hypothesis 2: There is no difference in 'Net Votes' between algorithms and humans in the treatment with the questions that have objective answers. Namely, the difference lies within a practically negligible range ($\Delta = 1.5$),[13] indicating equivalence.

The first two hypotheses capture the direct difference in 'Net Votes' between human and algorithm answers within each treatment. In contrast, Hypothesis 3 employs a pooled regression with an interaction term to test whether the human-algorithm gap differs between subjective and objective questions. Thus, Hypothesis 3 tests whether the human advantage in 'Net Votes' is significantly larger for subjective questions.

Hypothesis 3 aligns with Observation 1 and Equation 5, which suggest that if algorithmic aversion stems from belief-based partiality, then algorithmic aversion should decrease as the precision of the signal increases. In the experiment, the signal corresponds to the content of each questionnaire, either objective or subjective. As the precision of the signal increases, we move from the subjective treatment, characterized by higher uncertainty and greater subjectivity in judgment, to the objective treatment where uncertainly is reduced and judgment is more objective.

The purpose of this hypothesis is to identify the source of algorithmic aversion. Hy-

[13] Since participants evaluate five questions in total, the dependent variable 'Net Votes' takes values from -5 to +5. A difference of 1.5 in 'Net Votes' between group identities implies that, on average, one identity group is judged to have 1-2 additional answers evaluated as 'correct' than the other.

pothesis 3 builds on the theoretical framework, which predicts that if algorithms receive fewer 'Net Votes' than humans in subjective questions (as stated in Hypothesis 1), and there is a reduction in this difference in objective questions (as stated in Hypothesis 2), then algorithmic aversion stems from belief-based partiality. On the contrary, if the difference in 'Net Votes' remains constant across both treatments, indicating persistent algorithmic aversion despite the increase in the precision of the signal, then the source of aversion can be attributed to preference-based partiality.[14] This is consistent with Equation 5, which shows that the type-specific taste parameter remains unaffected by changes in the precision of the signal, even when the signal becomes perfectly objective.

Hypothesis 3: The human advantage in 'Net Votes' will be significantly larger on subjective answers compared to objective answers.

The remaining two hypotheses arise from the existing literature on algorithmic aversion. Hypothesis 4 explores how trust in algorithms affects people's evaluations in the experiment. Empirical evidence shows that unwillingness to rely on algorithmic advice comes from the lack of trust in this type of advice (Buchanan and Hickman, 2024; Chak et al., 2022). Also, demand for robo-advice is positively related to people's trust in such advice (Buchanan and Hickman, 2024; Chak et al., 2022; Oksanen, Savela, Latikka, and Koivula, 2020). Studies further report that algorithms are perceived as less trustworthy due to their lack of intuition and their subjective evaluation (Castelo et al., 2019; Lee, 2018). Hypothesis 4 examines whether trust in algorithms is significantly associated with a higher number of 'Net Votes' for algorithm-generated answers.

Hypothesis 4: General trust in algorithms is expected to be significantly positively correlated with the 'Net Votes' for the algorithm.

Hypothesis 5 is based on the existing literature on the relationship between individuals' demographic factors and algorithmic aversion. Several studies report a connection between characteristics, such as age, education, and gender, with peoples' resistance to

[14] Observing no algorithmic aversion is both treatments suggests that either beliefs and preferences are equal across groups, or that belief-based and preference-based partialities offset each other.

the use of algorithms. For example, females have been found to perceive algorithms as less useful in specific domains (Araujo, Helberger, Kruikemeier, and De Vreese, 2020). Females have stronger aversion towards robo-investment and report a preference for human decision making, compared to males (David and Sade, 2018; Niszczota and Kaszás, 2020; Sunstein and Reisch, 2023). These findings are consistent with other studies documenting gender differences on investment behavior (Barber and Odean, 2001; Lusardi and Mitchell, 2008). Age has also been negatively associated with the perceived usefulness of algorithmic decision-making, with older individuals tending to view such decisions as less useful (Araujo et al., 2020). Additionally, David and Sade (2018) report a nonlinear relationship between age and willingness to pay for algorithmic advice. Ozkes, Hanaki, Vanderelst, and Willems (2024) find no significant relationship between demographic characteristics and algorithmic behavior. This suggests that individual factors may not consistently influence attitudes toward algorithms. Hypothesis 5 explores whether females, older individuals, and those with lower levels of education report lower evaluations for algorithms.

Hypothesis 5: Being a female, older, and less educated is expected to be significantly correlated with a lower number of 'Net Votes' for algorithms.

5 Experimental Results

The experiment was programmed using oTree (Chen, Schonger, and Wickens, 2016) and conducted online via the platform 'Prolific' in English. The comprehensive set of instructions can be found in Appendix C. Experimental sessions took place in March 2025. Participants were randomly recruited from the UK population. The average duration of the experiment was approximately 11 minutes. The average earnings were £2.95, including a fixed participation fee of £1.5. The total number of participants was 349. A total of 11 participants were removed from the analysis, since they did not complete all of parts of the experiment. The number of participants for Treatment 1 and Treatment 2 were 175 and 174, respectively.[15] Table 2 provides a summary of the number of subjects in

[15] The sample size was determined based on the sample used by Bohren et al. (2019), as this paper extends their theoretical model.

each treatment.

Table 2: Number of Participants

Treatments	Human	Algorithm	Control
Treatment 1 (Objective Questions)	69	68	38
Treatment 2 (Subjective Questions)	67	68	39

A summary of key demographic characteristics of participants across treatment groups is presented in Table 3. The table illustrates the proportion of participants who are females, have income higher than £59,999, and hold at least a bachelor degree. It also includes the average age of participants within each group. In addition, the table reports the proportion of individuals who have previously used an algorithmic advisor, those who have reported a trust level in algorithms greater than 50% (in part 2 of the experiment), and the proportion of participants who did not evaluate all responses as "correct" in the main task. There are a few differences in demographic characteristics across treatment groups. Treatment 1 includes a slightly higher proportion of females and participants with a bachelor's degree, whereas Treatment 2 has a lower proportion of individuals who evaluate all responses as "correct". Equivalence tests confirm that these differences are not negligible, whereas income, prior experience with algorithmic advisors, and reported trust appear balanced across groups.[16] Table D.1 in Appendix D.1 provides information about the variables used throughout the paper for a more comprehensive understanding.

Table 3: Summary of key demographic characteristics across treatment groups

Treatments	Females	High Income	Age	Bachelor Degree	Used Algo-advisor	> 50% Trust	Not all Correct
Treatment 1	0.549	0.194	45.0	0.469	0.086	0.354	0.983
Treatment 2	0.489	0.172	43.8	0.391	0.126	0.368	0.810

OLS regression results regarding Hypotheses 1 to 3 are presented in Table 4. The variable 'Net Votes' is constructed for each individual by calculating the number of answers

[16] Two one-sided tests for proportions were conducted with $\epsilon = 0.10$ and $\alpha = 0.05$. Equivalence was not supported for gender ($\Delta = 0.060$, 90% CI [-0.028, 0.148]), university education ($\Delta = 0.078$, 90% CI [-0.009, 0.165]), and "Not all correct" responses($\Delta = 0.173$, 90% CI [0.121, 0.224]).

evaluated as correct minus the number of answers evaluated as incorrect. This measure reflects participants' evaluations rather than the actual correctness of their judgments. Column 1 examines the number of 'Net Votes' given to humans, versus algorithms, in the treatment with questions that have subjective answers, which is Treatment 2. The coefficient for human is not statistically significant ($p = 0.418$), suggesting that participants did not evaluate human and algorithm answers differently in this treatment. This finding does not support Hypothesis 1.

Column 2 reports the results of an equivalence test comparing 'Net Votes' between humans versus algorithms, in the treatment with questions that have objective answers, which is Treatment 1. Using an equivalence margin of $\Delta = 1.5$ the results show statistical equivalence, thus providing support for Hypothesis 2. It also shows exploratory OLS results for the objective treatment. The coefficient for human is not statistically significant ($p = 0.746$). This result is considered exploratory since only the equivalence test was preregistered for Hypothesis 2.

Column 3 reports pooled regression results from Treatment 1 and 2 within the same model. The interaction term between the human indicator and the subjective treatment is not statistically significant ($p = 0.440$), indicating that there is no human advantage on questions that have subjective answers. Therefore, Hypothesis 3 is not supported. The coefficient on the subjective variable is statistically significant ($p = 0.001$), indicating that subjective answers received higher 'Net Votes' than objective ones. Additional robustness checks that account for the attention check,[17] along with probit regression results, yield similar findings. Appendices D.2 - D.3 provide the related tables. Taken together, the regression results in Table 4 provide no evidence of algorithmic aversion in the sample.

Figure 1 shows the distribution of 'Net Votes' across treatments. In the objective treatment (Treatment 1), 'Net Votes' are centered around zero and the distributions appear identical for humans and algorithms. In the subjective treatment (Treatment 2), there is a slight shift of the distribution to the right and a small increase in positive 'Net Votes' for

[17] The final question of the experiment served as an attention check. Specifically, participants were asked: *"Who provided the answers to the five economic questions in the main task of this experiment?"* The response options were "Algorithms", "Human Economists", and "I do not recall".

Table 4: OLS Regression Results - Hypotheses H1 to H3

	Net Votes		
OLS Regression	**Treatment 2**	**Treatment 1**	**Both**
	(1)	**(2)**	**(3)**
Human	0.263	-0.123	-0.123
	(0.326)	(0.382)	(0.382)
Subjective			1.147***
			(0.346)
Human × Subjective			0.387
			(0.502)
Intercept	0.676***	-0.470*	-0.470*
	(0.220)	(0.267)	(0.267)
Observations	135	137	272
R-squared	0.005	0.001	0.098
Equivalence Test (TOST)	Equivalent	Equivalent	–
Hypothesis	H1	H2	H3

Notes: Robust standard errors in parentheses. *p<0.10, **p<0.05, ***p<0.01. $\Delta = 1.5$.
Net Votes: number of answers evaluated as correct minus those evaluated as incorrect.

humans. Specifically, in the subjective treatment, about 7% of participants in the human condition evaluated all answers as correct, whereas only 4% did so in the algorithmic condition. However, this difference is not statistically significant, and the distributions for humans and algorithms remain almost identical. Results from a Kolmogorov–Smirnov test for both treatments indicate no statistically significant difference in the distributions of 'Net Votes' between the two groups.[18] This result suggests that participants evaluated algorithmic and human answers similarly in both treatments, consistent with the absence of algorithmic aversion in the sample.

Table 5 reports the results from OLS regressions examining the relationship between trust in algorithms and 'Net Votes'. The analysis includes only participants who evaluated answers from algorithms only, across Treatments 1 and 2. The table shows results from two different measures of trust in algorithms: 'Average Trust' and 'Self-reported Trust'. The variable 'Average Trust' was constructed using responses from the second part of

[18] KS statistic for Treatment 1 is 0.025 ($p = 1.000$) and KS statistic for Treatment 2 is 0.063 ($p = 0.997$).

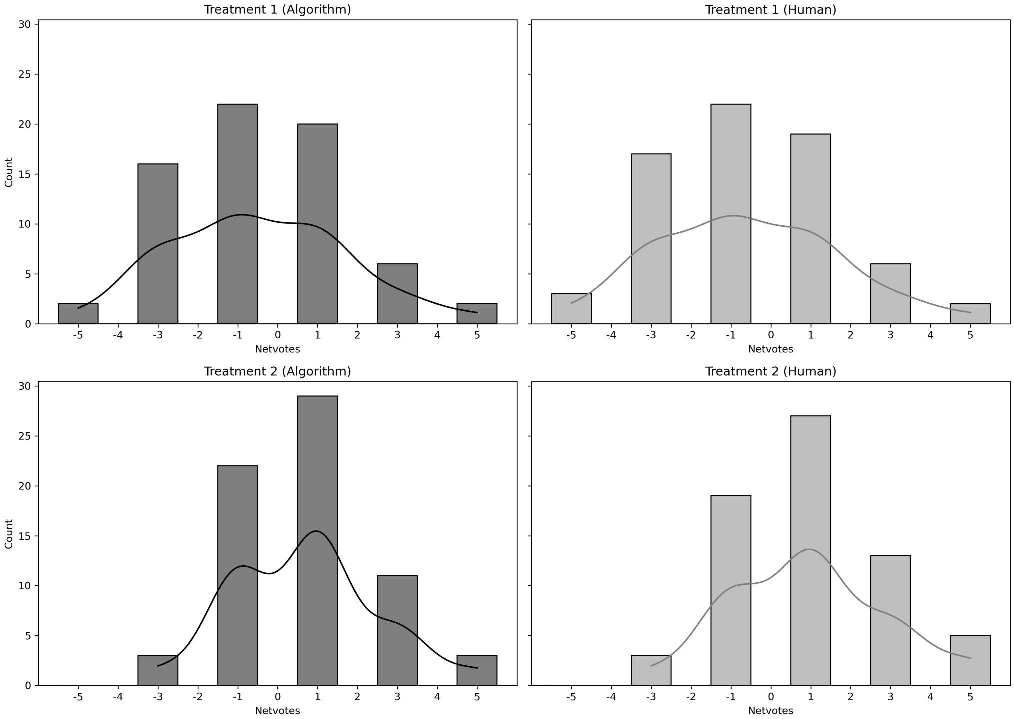


Figure I: Distribution of Net Votes

the experiment, where participants indicated their level of trust for algorithms, compared to humans, to perform specific tasks.[19] The variable was computed by averaging the trust scores reported by each participant. The coefficient on average trust in Column (1) of Table 5 is not statistically significant ($p = 0.206$). The variable 'Self-reported Trust' is statistically significant ($p = 0.006$),[20] indicating that higher self-reported trust in algorithms is associated with higher 'Net Votes' for algorithm-generated answers. Similar results are obtained when participants who did not pass the attention check are excluded from the analysis (see Appendix D.2, Table D.2.1).

The relationship between the two measures was examined using Spearman's rank-order correlation. The results indicate a statistically significant and moderately strong positive correlation ($\rho = 0.525, p < 0.001$). The difference in the results of the two measures may

[19] This task is based on the work of Castelo et al. (2019). Participants indicated how much they trust either an algorithm or a human on a scale from 0 ("not at all") to 100 ("completely") to perform six tasks. The tasks varied in their level of subjectivity, with three tasks being subjective and the remaining three being objective.

[20] The variable 'Self-reported Trust' in Column (2) is a self-reported categorical measure of trust derived from the experimental question: "On a scale of 1 to 5, how much do you trust algorithms to make accurate decisions?". The possible answers were: 'Not Trustworthy', 'Low Trust', 'Moderate Trust', 'Significant Trust' and 'Complete Trust'.

Table 5: OLS Regression Results – Hypothesis H4

	Net Votes	
OLS Regression	**Algorithm Treatment**	**Algorithm Treatment**
	(1)	**(2)**
Average Trust	0.011	
	(0.009)	
Self-reported Trust		0.586***
		(0.214)
Intercept	-0.409	-1.699**
	(0.431)	(0.656)
Observations	136	136
R-squared	0.009	0.043
Hypothesis	H4	H4

Notes: Robust standard errors in parentheses. *p<0.10, **p<0.05, ***p<0.01.

arise from how they are constructed. Specifically, 'Average Trust' is calculated as the mean of participants' ratings across six hypothetical scenarios, in which they used a bar scale to indicate how much they trusted an algorithm or a human. By contrast, 'Self-reported Trust' reflects a single, general self-assessment. As a result, the first measure may introduce more variability, whereas the second one captures a more stable attitude towards trust in algorithms.

Figure 2 presents box plots of the distribution of 'Net Votes' for algorithm versus human-generated answers across objective and subjective questions, separately for two different levels of trust using the 'Self-reported Trust' variable. Participants were divided into 'Low Trust' and 'High Trust' groups based on their responses to the question: "On a scale of 1 to 5, how much do you trust algorithms to make accurate decisions?" Those who selected 'Not Trustworthy', 'Low Trust', or 'Moderate Trust' were classified as 'Low Trust', while those who selected 'Significant Trust' or 'Complete Trust' were classified as 'High Trust'.

Figure 2 shows no statistically significant differences in 'Net Votes' between human and algorithmic answers for either trust group across treatments. Among low-trust participants, algorithmic answers received slightly fewer 'Net Votes' than human answers in the

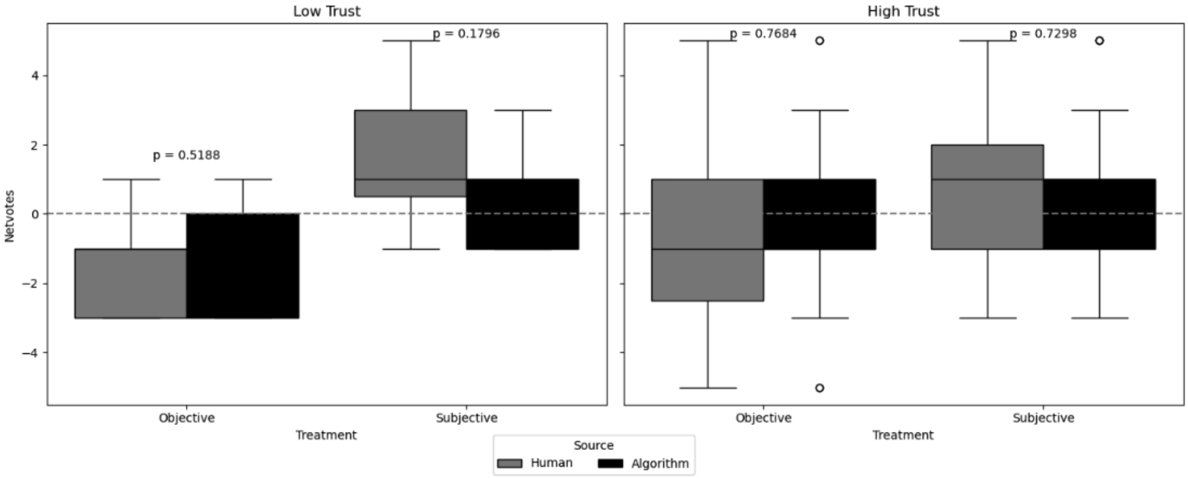


Figure 2: Net Vote Distributions by Trust Level

subjective treatment, although the difference is not statistically significant ($p = 0.179$). For high-trust participants, evaluations of human and algorithm-generated answers were nearly identical in both treatments ($p = 0.768, p = 0.729$).

Finally, Table 6 presents results for Hypothesis 5. Column (1) examines whether women assign fewer 'Net Votes' to answers provided by algorithms, compared to men, across both treatments. Although the coefficient is negative, the result is not statistical significant ($p = 0.308$), indicating no evidence that being female is associated with a significantly lower evaluation of algorithmic responses.

Similarly, Columns (2) and (3) show that neither age ($p = 0.424$) nor education ($p = 0.578$) are significantly correlated with a lower number of 'Net Votes' for algorithms.[21] Lastly, Column (4), which includes all covariates, also shows no support for Hypothesis 5. Results are similar when participants who did not respond correctly to the attention check question are excluded (see Appendix D.2, Table D.2.3).

To examine the effect of revealing the selector's identity, OLS regressions are performed comparing evaluations of human or algorithm answers to those from the control treatment, where no information about the source was provided. As shown in Table 7, human answers receive significantly higher 'Net Votes' than the control group in objective questions (Treatment 1), but not in subjective questions (Treatment 2). A similar pattern

[21] Education is coded into 4 categories: 0 = other, 1 = high school, 2 = bachelor's degree, 3 = master's degree, 4 = doctorate.

Table 6: OLS Regression Results – Hypothesis H5

	Net Votes			
	Algorithm Treatment			
OLS Regression	**(1)**	**(2)**	**(3)**	**(4)**
Female	-0.362			-0.324
	(0.356)			(0.355)
Age		-0.011		-0.008
		(0.014)		(0.014)
Education			0.122	0.073
			(0.211)	(0.209)
Intercept	0.292	0.585	-0.139	0.505
	(0.234)	(0.645)	(0.460)	(0.779)
Observations	136	136	136	136
R-squared	0.008	0.005	0.003	0.012
Hypothesis	H5	H5	H5	H5

Notes: Robust standard errors in parentheses. *p<0.10, **p<0.05, ***p<0.01.

is observed when the algorithm treatment is compared to the control treatment. These results suggest that revealing the identity of the answer source increases evaluations only in objective questions, possibly due to the nature of the task. In contrast, in subjective questions, where uncertainty is higher, participants may default to positive evaluations regardless of the source, making group identity less significant.

The preregistered hypotheses of this paper were evaluated jointly to provide a comprehensive assessment of algorithmic aversion across treatments, trust measures, and demographic characteristics. Taken together, the experimental results show no evidence of algorithmic aversion. Across all specifications, participants did not systematically favor human generated answers over algorithmic ones, and the coefficients on the key variables remain statistically insignificant.

To summarize, Hypotheses 1 and 3 are not supported, whereas Hypothesis 2 is supported. Hypothesis 4 regarding trust is supported, showing that trust in algorithms is positively associated with ‘Net Votes’. Nevertheless, since the second trust measure did not yield statistically significant results, this finding should be interpreted with caution. Demographic variables like gender, age, and education are not statistically significant. As a result, Hypothesis 5 is not supported.

Table 7: OLS Regression Results - Comparison with Control

Net Votes	
OLS Regression	**(1)**
Human	1.037**
	(0.488)
Algorithm	1.161**
	(0.485)
Subjective	1.965***
	(0.527)
Human × Subjective	-0.430
	(0.641)
Algorithm × Subjective	-0.818
	(0.631)
Intercept	-1.631***
	(0.405)
Observations	349
R-squared	0.131

Notes: Robust standard errors in parentheses.

*p<0.10, **p<0.05, ***p<0.01.

Exploratory Analysis

An OLS regression was performed to compare participants' evaluation accuracy between Treatments 1 and 2 (See Appendix D.4, Table D.4.1). Results showed that participants in the subjective treatment (Treatment 2) correctly evaluated, on average, 1.18 more answers than those in the objective treatment (Treatment 1) ($p = 0.000$). Specifically, participants in Treatments 1 and 2 evaluated on average 2.23 and 2.90 answers as "correct", respectively. Participants in Treatment 2 were slightly more likely to evaluate answers as "correct". However, both treatments included three correct and two incorrect answers.

To examine whether the subjective questions were indeed more subjective than the objective ones, mean evaluations and variances were computed for each question. The mean share of positive evaluations in the objective questions varies from 0.33 to 0.53, with variance across participants around 0.24. For the subjective questions, the mean share of positive evaluations varies between 0.35 and 0.87, and the variance varies from 0.11 to 0.24. This indicates that the subjective treatment included questions with greater heterogeneity in perceived correctness, since some questions were considered more correct than others.

An independent t-test was also performed comparing the within-subject variances across treatments showing no statistically significant difference in voting variability ($t = -1.31$, $p = 0.19$). This indicates that participants were equally consistent in their evaluations in both treatments. Taken together, these findings suggest that although participants in the subjective treatment were more likely to evaluate answers as correct, this pattern does not reflect higher inconsistency or noise in their evaluations. Instead, it reflects greater perceived correctness of subjective questions, which may arise from increased uncertainty under lower signal precision, consistent with the theoretical framework.

To better understand what explains 'Net Votes', an OLS regression analysis was performed using various control variables. Results are reported in Table D.4.2 in Appendix D.4. The results indicate that subjective questions received significantly higher 'Net Votes', while other predictors such as trust, gender, and age are not statistically significant. Education and salary are marginally significant, suggesting a possible role in evaluations.

Specifically, education has a negative coefficient, which may suggest that individuals with higher education were more critical in their evaluations, leading them to evaluate more answers as "incorrect". In contrast, salary has a positive coefficient, implying that participants with higher incomes tended to evaluate more answers as "correct". One possible explanation could be that those individuals were more likely to evaluate an answer as "correct" unless it was clearly wrong, perhaps due to their lower sensitivity to the monetary incentives compared to participants in lower income categories. However, these results should be treated as tentative, as they are based on exploratory analyses and the coefficients are only marginally significant.

Post-hoc power analysis for Hypothesis 1 showed that the observed power to detect a small-to-moderate effect size (Cohen's $d = 0.342$) was approximately 52.3%. Similarly, for Hypothesis 3, the observed power to detect a small-to-moderate interaction (Cohen's $f = 0.202$) was 52.0%. Although the study was designed with a sample size based on prior literature (Bohren et al., 2019), the observed effect sizes are smaller than anticipated. As a result, the lack of statistical significance should be interpreted with caution, as it may reflect limited power rather than the absence of an effect.

Lastly, a supplementary robustness check using only questions with high agreement, showed no meaningful differences between human and algorithm evaluations. These results indicate that evaluations of human and algorithmic answers remain similar even when belief-based uncertainty is low, supporting the interpretation that the overall absence of algorithmic aversion reflects equality of preferences rather than opposing biases that cancel each other out. This is also confirmed by an equivalence test. Refer to Table D.4.3 in Appendix D.4.

To conclude, the exploratory analysis indicates that participants in the subjective treatment were generally more likely to evaluate answers as "correct", and that education and salary influence 'Net Votes'. These findings were not preregistered and should therefore be interpreted with caution.

6 Discussion and Conclusions

The paper provides an economic theoretical framework to better understand the topic of algorithmic aversion. The theoretical model of the paper conceptualizes algorithmic aversion as a form of individual partiality, and by varying key parameters in the model, its source can be distinguished as either belief-based or preference-based partiality.

An online experiment was performed to test the theoretical predictions of the model. Participants in the experiment did not exhibit algorithmic aversion. Specifically, the data do not align with the theoretical predictions of the model, since no statistically significant difference was observed between evaluations of human and algorithmic answers in either treatment condition. One possible explanation for this finding is that participants may perceive humans and algorithms as equally capable, indicating the absence of belief-based partiality, and may also lack a clear preference for responses provided by humans, suggesting an absence of preference-based partiality. Theoretically, belief-based and preference-based partiality could offset each other, but the data provide no evidence of such an effect.

The contribution of this paper is to provide a novel theoretical framework for interpreting algorithmic aversion. As a further step, this paper aimed at testing this framework empirically using a controlled experiment. Results can be interpreted as a test of the theoretical model applied to the emerging topic of algorithmic aversion.

Another potential explanation of the results is that algorithmic aversion might not persist in evaluation tasks. The theoretical model assumed that individuals would differ in their evaluations either due to differences in their beliefs about the ability of the two groups, or due to a disamenity associated with selecting algorithmic workers. The absence of such differences in the evaluation task raises the possibility that algorithmic aversion may be more relevant in other domains or tasks, such as delegation tasks, where individuals entrust decisions to algorithms and may experience regret when outcomes are negative.

Additionally, the selection of economists as a reference group may have influenced the results if trust in them is limited. The absence of algorithmic aversion observed in the data may reflect the particular position that economists hold on the perceived human–algorithm competence spectrum. If people hold specific views about economists that do not necessarily extend to other expert groups or laypeople, future research should verify whether the results generalize to other experts as well as to laypeople. Nonetheless, another possible interpretation of the results is that economists occupy an intermediate position on the perceived competence spectrum between human and algorithmic decision-making.

The general finding of no algorithmic aversion suggests that we may need to re-examine the assumptions around algorithmic aversion. If algorithmic aversion does exist, is it robust across all contexts, or is it contingent on specific environmental or task-related factors? Future research should aim to explore these questions further.

SUPPLEMENTARY MATERIAL

Appendix A.1 Derivations

The worker has an unobservable ability $\alpha \sim N(\mu_g, 1/\tau_a)$ with $\tau_a > 0$ the precision. The worker can complete a sequence of tasks where each task has a hidden quality $q_t = \alpha + \epsilon_t$ where $\epsilon_t \sim N(0, 1/\tau_\epsilon)$ is an independent random shock with precision $\tau_\epsilon > 0$. Here we calculate the evaluator's prior belief about quality and show that the prior belief is normally distributed with mean $\hat{\mu}_g$ and precision $\tau_q \equiv \tau_a \tau_\epsilon/(\tau_a + \tau_\epsilon)$, i.e $q_1 \sim N(\hat{\mu}_g, 1/\tau_q)$.

To find the prior belief of quality, we need to combine the distributions of α and ϵ_t. To find the mean of q_t we use the mean of α which is μ_g and the mean of ϵ_t which is 0. Therefore, the mean of q_t is calculated as follows:

$$E[q_t] = E[\alpha + \epsilon_t] = E[\alpha] + E[\epsilon_t] = \hat{\mu}_g + 0 = \hat{\mu}_g.$$

To find the variance of q_t, we use the variance of a $1/\tau_\alpha$ and the variance of ϵ_t, which is $1/\tau_\epsilon$. Since α and ϵ_t are independent, the variance of q_t is the sum of the two variances:

$$Var(q_t) = Var(\alpha) + Var(\epsilon) = 1/\tau_\alpha + 1/\tau_\epsilon = \frac{\tau_\alpha + \tau_\epsilon}{\tau_\alpha \tau_\epsilon}.$$

Consequently, the precision of quality is defined as $\tau_q \equiv \tau_a \tau_\epsilon/(\tau_a + \tau_\epsilon)$.

To find the evaluator's posterior belief about the quality conditional on observing the initial signal s_1, we will use Bayesian updating. Given the prior belief about q_1 and the distribution of the initial signal s_1, the evaluator's posterior belief about the quality q_1 conditional on observing s_1 is updated using Bayesian inference. We know that the initial signal has conditional distribution $s_1/q_1 \sim N(q_1, 1/\tau_\eta)$. The posterior mean is a weighted average of the prior mean and the observed signal, weighted by their respective precisions, which is $\frac{\tau_q \hat{\mu_g} + \tau_\eta s_1}{\tau_q + \tau_\eta}$. Remember that $\hat{\mu}_g$ is the prior mean, s_1 is the observed signal, τ_q is the precision of the prior mean and $\tau\eta$ is the precision of the signal. The posterior precision is the sum of the prior precision and the precision of the signal noise, which is $\tau_q + \tau_\eta$.

The variance is denoted by $1/\tau_{post}$, where τ_{post} is the posterior precision. Given the prior belief and signal distribution, the evaluator's posterior belief about quality conditional on observing s_1 is also normally distributed, $q_1/s_1 \sim N(\frac{\tau_q \hat{\mu_g} + \tau_\eta s_1}{\tau_q + \tau_\eta}, \frac{1}{\tau_q + \tau_\eta})$.

Appendix A.2: Derivations

Based on Equation 4, algorithmic aversion is defined as the positive difference in evaluations between humans and algorithms, at the same signal:

$$AlgoAversion_t(s_t) \equiv v_t(s_t, H) - v_t(s_t, A)$$

Using Equation 2, the evaluator reports an upvote on the first task if:[22]

$$\hat{E}_i[q_1|s_1, g] = \frac{\tau_q \hat{\mu}_g^t + \tau_\eta s_1}{\tau_q + \tau_\eta} \geq c_g^t.$$

Substituting the above to Equation 4, and assuming a positive vote for both group identities[23], we derive the formal definition:

$$AlgoAversion_t(s_t) = \left(\frac{\tau_q \hat{\mu}_H^t + \tau_\eta s_1}{\tau_q + \tau_\eta}\right) - c_H^t - \left(\frac{\tau_q \hat{\mu}_A^t + \tau_\eta s_1}{\tau_q + \tau_\eta}\right) + c_A^t$$

Rearranging the terms:

$$AlgoAversion_t(s_t) = (\frac{\tau_q}{\tau_q + \tau_\eta})(\hat{\mu}_H^t - \hat{\mu}_A^t) + (c_A^t - c_H^t).$$

Taking expectations across evaluators, and letting $\hat{\mu}_g = E\left[\hat{\mu}_g^t\right]$, $c_A = E[c_A^t]$, the aggregate expression is given by Equation 5:

$$AlgoAversion(s_t) = (\frac{\tau_q}{\tau_q + \tau_\eta})(\hat{\mu}_H - \hat{\mu}_A) + (c_A - c_H),$$

[22] The derivation is shown for the first task ($t = 1$) for notational simplicity, but the result holds for any task t.

[23] This assumption is made to facilitate a clearer understanding of the components of algorithmic aversion.

where $\hat{\mu}_H$ and $\hat{\mu}_A$ denote the average beliefs about human and algorithmic ability, respectively, and c_A and c_H denote the average taste parameter for algorithms and humans.

Appendix B.1. Questionnaire

This section presents the economic questions used in the experiment, with the correct answers indicated in bold, as reported by Alysandratos et al. (2020).

Treatment 1 - Objective Questions

1. A prestigious, merit-based, scholarship for graduate studies from a private institution is awarded to 5% of the applicants. Assume that an undergraduate student is chosen at random and applies for the scholarship. What is the likelihood that (s)he will be a recipient of this scholarship?

a) 0% **b) Less than 5%** c) 5% d) More than 5

Answer shown to participants: b)

2. In Richland at first no Value Added Tax (VAT) or other tax is imposed on fruits. The price of a kilo of apples is 100 Richland Pounds. The government is thinking of imposing a VAT of 24% on all fruits. What will be the price of apples after the market adjusts to the tax?

a) 100 **b) Between 100 and 124** c) 124 d) More than 124

Answer shown to participants: b)

3. The government of Rainland borrows 100 billion from private investors at a 5% interest rate. If it pays back to investors 5 billion per year, how many years will it take to repay its debt?

a) 20 b) 40 **c) It depends on the borrowing terms** d) For ever

Answer shown to participants: c)

4. Robert won a free ticket to see Justin Bieber. But Beyonce is performing on the same night and he can only attend one of the two events. He likes Beyonce and he would pay up to £50 to see her perform, and the tickets for Beyonce's event cost £40. What must be the minimum value of Bieber's performance to Robert so that Robert chooses Bieber over Beyonce?

a) £0 **b) £10** c) £40 d) £50

Answer shown to participants: a)

5. The previous government of Girtonia, a developed country, invested 100 million last year in building a regional airport. The airport is now ready to open its doors and it is expected to generate a total net profit of 75 million for the duration of its use. The current government is reconsidering the project and has found a new location for the airport. The new airport would yield earnings of 150 million for the duration of its use and it would also cost 100 million to build. If the old airport is abandoned it would have 0 value to the government. Should the government go ahead with the new project?

a) No b) Yes c) Both projects are equally profitable d) Insufficient information to answer

Answer shown to participants: b)

Treatment 2 - Subjective Questions

1. In an attempt to address its housing crisis, manifested through rapidly rising rents and house prices, the mayor of Bigcitia, a burgeoning capital in a high income country, announced that the city will impose a freeze for existing renters and restrict rent increases to 1% of the average price in the neighbourhood for new renters. Will this policy alleviate the housing crisis and result to more people finding a house in the next 5 years?

a) Yes b) Maybe Yes **c) No** d) All previous options are equally likely to be correct

Answer shown to participants: c)

2. The government of Freeland, a multiethnic, without a dominant ethnic group, free market, high income economy announces a new law according to which all workers of ethnicity K must receive a 50% higher wage than comparable employees. What do you expect to be the effect of the new law on the probability of finding a job for a random member of ethnicity K that is now entering the labour market for the first time?

a) Positive b) Neutral **c) Negative** d) All previous options are equally likely to be correct

Answer shown to participants: c)

3. After several successful trials a start up from California has announced the commercial licencing of its eagerly awaited autonomous car technology. Market analysts expect it will

take 6 months for taxi companies to obtain regulatory approval and another 6 months to fully deploy the technology. Assuming the analysts' timeline is accurate, what do you expect to be the effect on the employment rate of current taxi drivers 12 months from now?

a) It will increase b) It will be unaffected **c) It will decrease** d) All previous answers are equally likely to be correct

Answer shown to participants: c)

4. Hobson Plc and Thornbush Plc announced on Friday, after the stock market had closed, an unexpected merger of equals. During the weekend the majority of economic analysts and financial media, who were surprised by the news, have spoken against it on the basis that it will be unprofitable. What is the most likely price movement for the stock prices of the two companies over the coming week (Monday to Friday) if they are allowed to continue trading their stocks on Monday and no additional news on the value of the two companies arrives to the markets?

a) Both up b) Hobson up, Thornbush down c) Hobson down, Thornbush up **d) Both down**

Answer shown to participants: a)

5. Following its commitment to cut global warming emissions, the Prime Minister of Richland announced a 10-year guaranteed price scheme, significantly above current market prices, for buying electricity from new installations of wind and solar power farms in the country. Five years after the implementation of this policy, the percentage of electrical power produced from renewable sources will be:

a) Lower b) The same **c) Higher** d) It cannot be determined

Answer shown to participants: d)

Appendix B.2.

To ensure a valid statistical analysis and eliminate the possibility that differences in participants' choices are due to varying answer quality between the different sources, it is important to create an environment where the two groups have the same responses. Therefore, the answers shown to participants were the same across all treatments. To do this, a pool of answers from economists and AI chatbots were used. Specifically, answers were drawn both from the responses of humans and algorithms. Answers for the humans were drawn from the responses of academic economists at universities across Europe, based on the distribution of answers provided by Alysandratos et al. (2020). Answers for the algorithms were drawn from the responses of AI chatbots, such as ChatGPT-3, Google Bard, Claude, You.com and Microsoft Copilot. The answers shown to participants were randomly selected from both pools based on specific criteria.

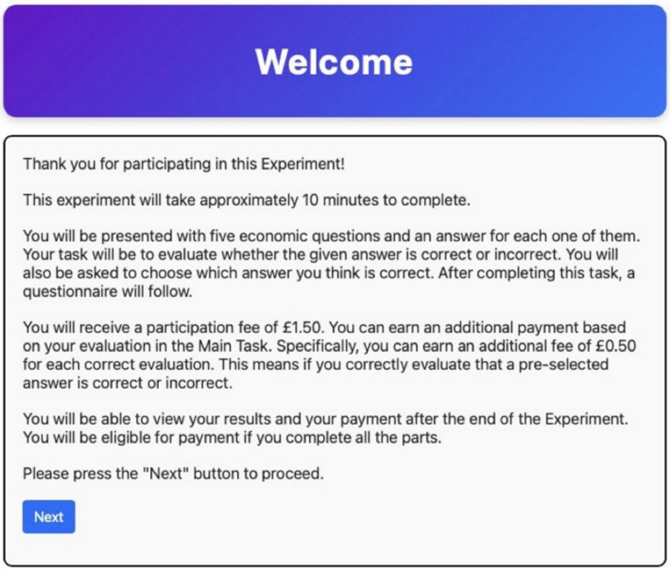
To facilitate this process, a Python script was used to randomly assign answers to the 10 questions, ensuring that the selected answers reflect the distributions of the AI chatbots and the human responses. The script included criteria to balance the number of correct and incorrect answers between the two treatments, with three answers being correct and two being incorrect. The outcome of this process determined the answers shown to participants in the experiment under the algorithm, human, and control treatment.

This approach guaranteed that algorithmic and human responses are drawn from their respective distributions while maintaining an equal number of correct answers for both groups and preserving randomness in response assignment. Participants in the algorithm treatment were informed that the answers to the questions were generated by different AI chatbots. The name of the AI chatbots, along with some information about their workings, were also provided to participants. Participants in the human treatment were informed that the answers to the questions were coming from a set of responses previously provided by different academic economists from European universities. In both treatments, participants were asked to evaluate each question independently.

Appendix C. Experiment Flow

The screenshots presents the objective treatment with human answers.

Experiment: Treatment 1



Welcome

Thank you for participating in this Experiment!

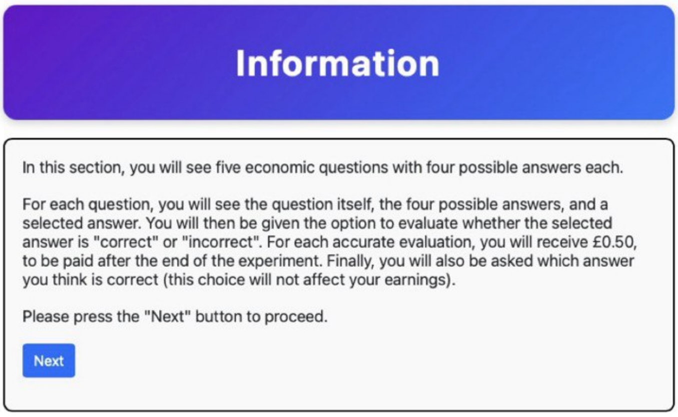
This experiment will take approximately 10 minutes to complete.

You will be presented with five economic questions and an answer for each one of them. Your task will be to evaluate whether the given answer is correct or incorrect. You will also be asked to choose which answer you think is correct. After completing this task, a questionnaire will follow.

You will receive a participation fee of £1.50. You can earn an additional payment based on your evaluation in the Main Task. Specifically, you can earn an additional fee of £0.50 for each correct evaluation. This means if you correctly evaluate that a pre-selected answer is correct or incorrect.

You will be able to view your results and your payment after the end of the Experiment. You will be eligible for payment if you complete all the parts.

Please press the "Next" button to proceed.

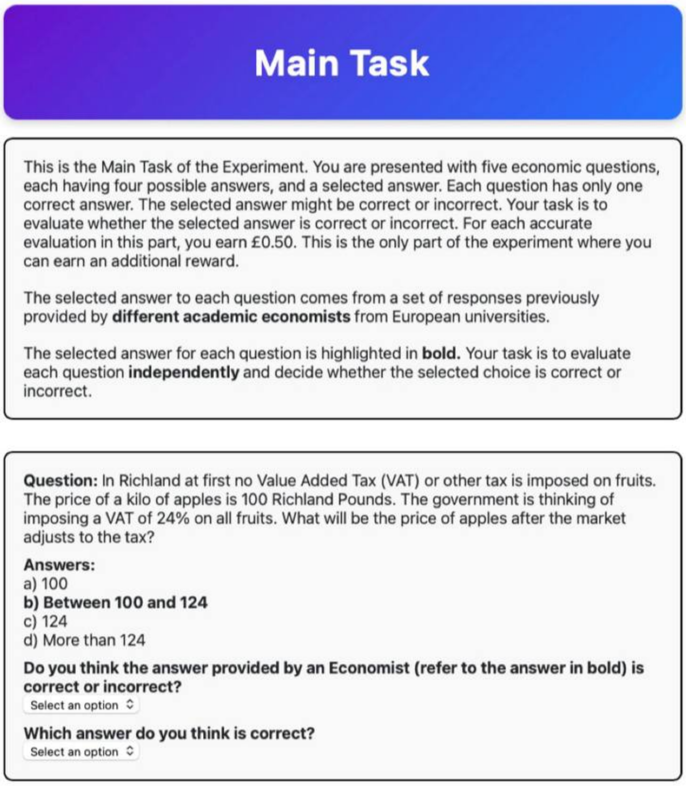


Information

In this section, you will see five economic questions with four possible answers each.

For each question, you will see the question itself, the four possible answers, and a selected answer. You will then be given the option to evaluate whether the selected answer is "correct" or "incorrect". For each accurate evaluation, you will receive £0.50, to be paid after the end of the experiment. Finally, you will also be asked which answer you think is correct (this choice will not affect your earnings).

Please press the "Next" button to proceed.



Main Task

This is the Main Task of the Experiment. You are presented with five economic questions, each having four possible answers, and a selected answer. Each question has only one correct answer. The selected answer might be correct or incorrect. Your task is to evaluate whether the selected answer is correct or incorrect. For each accurate evaluation in this part, you earn £0.50. This is the only part of the experiment where you can earn an additional reward.

The selected answer to each question comes from a set of responses previously provided by **different academic economists** from European universities.

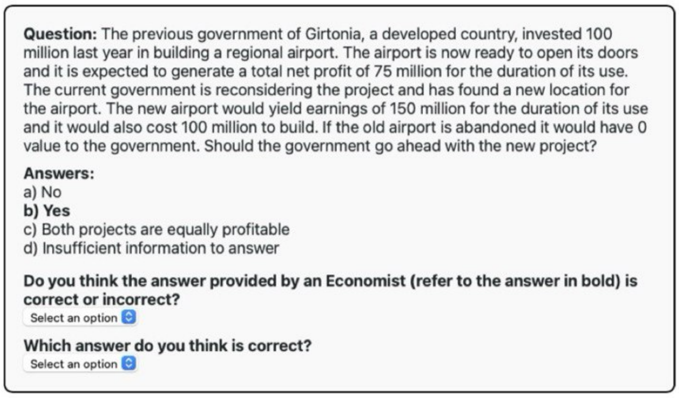
The selected answer for each question is highlighted in **bold.** Your task is to evaluate each question **independently** and decide whether the selected choice is correct or incorrect.

Question: In Richland at first no Value Added Tax (VAT) or other tax is imposed on fruits. The price of a kilo of apples is 100 Richland Pounds. The government is thinking of imposing a VAT of 24% on all fruits. What will be the price of apples after the market adjusts to the tax?

Answers:
a) 100
b) Between 100 and 124
c) 124
d) More than 124

Do you think the answer provided by an Economist (refer to the answer in bold) is correct or incorrect?
Select an option ⇕

Which answer do you think is correct?
Select an option ⇕

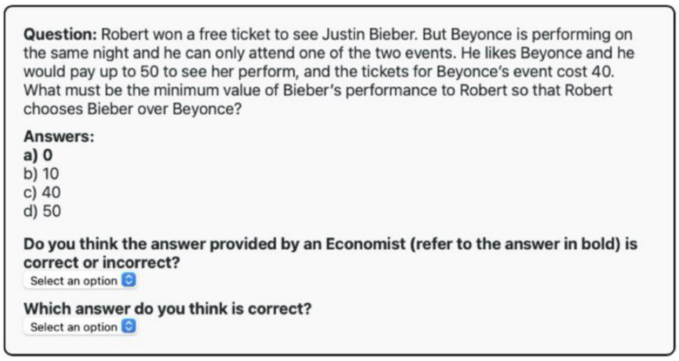


Question: The previous government of Girtonia, a developed country, invested 100 million last year in building a regional airport. The airport is now ready to open its doors and it is expected to generate a total net profit of 75 million for the duration of its use. The current government is reconsidering the project and has found a new location for the airport. The new airport would yield earnings of 150 million for the duration of its use and it would also cost 100 million to build. If the old airport is abandoned it would have 0 value to the government. Should the government go ahead with the new project?

Answers:
a) No
b) Yes
c) Both projects are equally profitable
d) Insufficient information to answer

Do you think the answer provided by an Economist (refer to the answer in bold) is correct or incorrect?
Select an option ⇕

Which answer do you think is correct?
Select an option ⇕

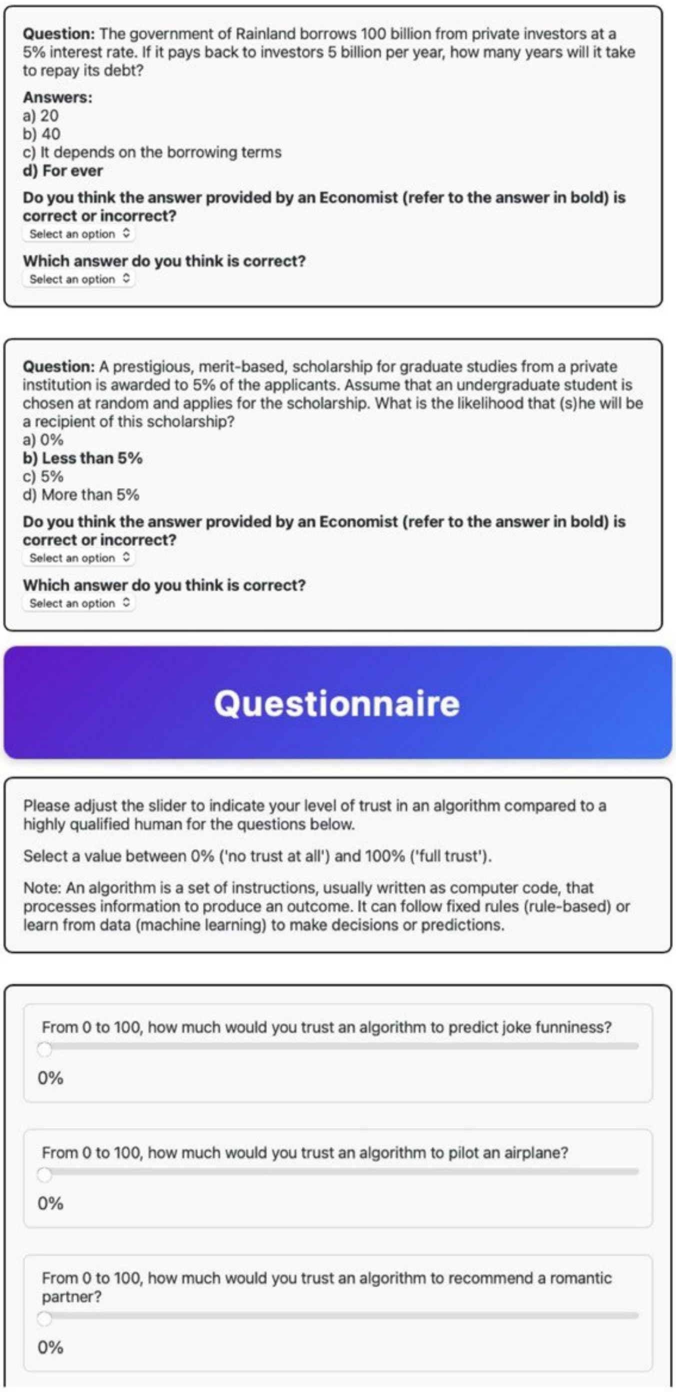


Question: Robert won a free ticket to see Justin Bieber. But Beyonce is performing on the same night and he can only attend one of the two events. He likes Beyonce and he would pay up to 50 to see her perform, and the tickets for Beyonce's event cost 40. What must be the minimum value of Bieber's performance to Robert so that Robert chooses Bieber over Beyonce?

Answers:
a) 0
b) 10
c) 40
d) 50

Do you think the answer provided by an Economist (refer to the answer in bold) is correct or incorrect?
Select an option ⇕

Which answer do you think is correct?
Select an option ⇕



Question: The government of Rainland borrows 100 billion from private investors at a 5% interest rate. If it pays back to investors 5 billion per year, how many years will it take to repay its debt?

Answers:
a) 20
b) 40
c) It depends on the borrowing terms
d) For ever

Do you think the answer provided by an Economist (refer to the answer in bold) is correct or incorrect?
Select an option

Which answer do you think is correct?
Select an option

Question: A prestigious, merit-based, scholarship for graduate studies from a private institution is awarded to 5% of the applicants. Assume that an undergraduate student is chosen at random and applies for the scholarship. What is the likelihood that (s)he will be a recipient of this scholarship?
a) 0%
b) Less than 5%
c) 5%
d) More than 5%

Do you think the answer provided by an Economist (refer to the answer in bold) is correct or incorrect?
Select an option

Which answer do you think is correct?
Select an option

Questionnaire

Please adjust the slider to indicate your level of trust in an algorithm compared to a highly qualified human for the questions below.

Select a value between 0% ('no trust at all') and 100% ('full trust').

Note: An algorithm is a set of instructions, usually written as computer code, that processes information to produce an outcome. It can follow fixed rules (rule-based) or learn from data (machine learning) to make decisions or predictions.

From 0 to 100, how much would you trust an algorithm to predict joke funniness?

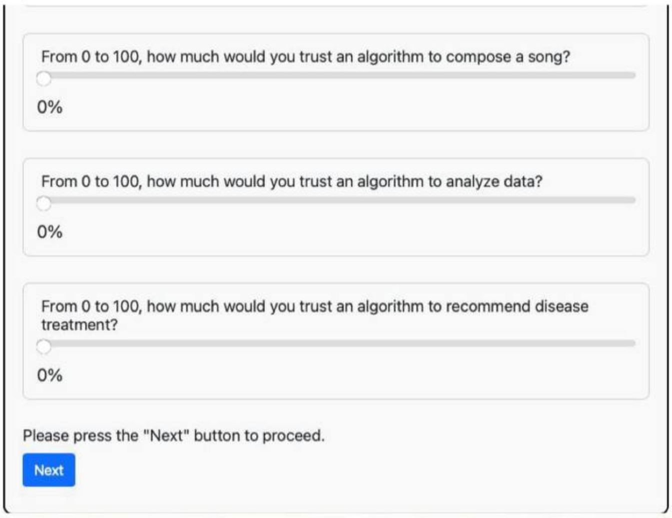
0%

From 0 to 100, how much would you trust an algorithm to pilot an airplane?

0%

From 0 to 100, how much would you trust an algorithm to recommend a romantic partner?

0%



From 0 to 100, how much would you trust an algorithm to compose a song?

0%

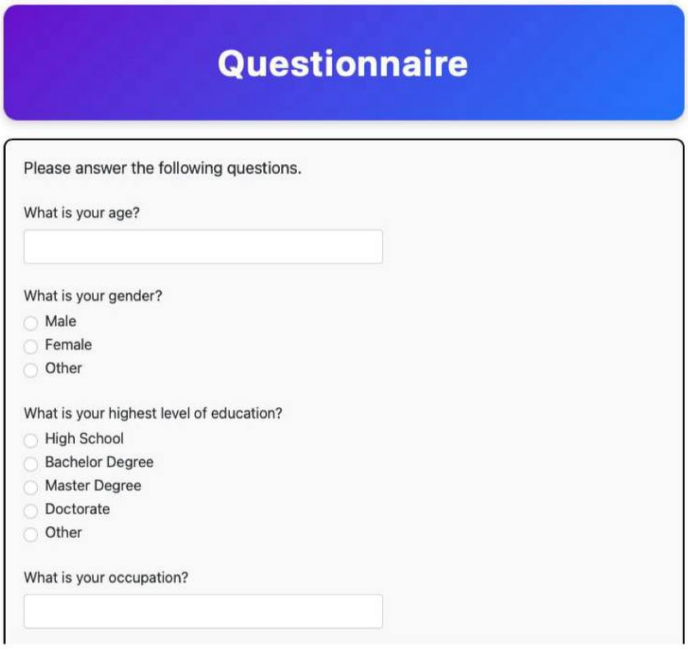
From 0 to 100, how much would you trust an algorithm to analyze data?

0%

From 0 to 100, how much would you trust an algorithm to recommend disease treatment?

0%

Please press the "Next" button to proceed.



Questionnaire

Please answer the following questions.

What is your age?

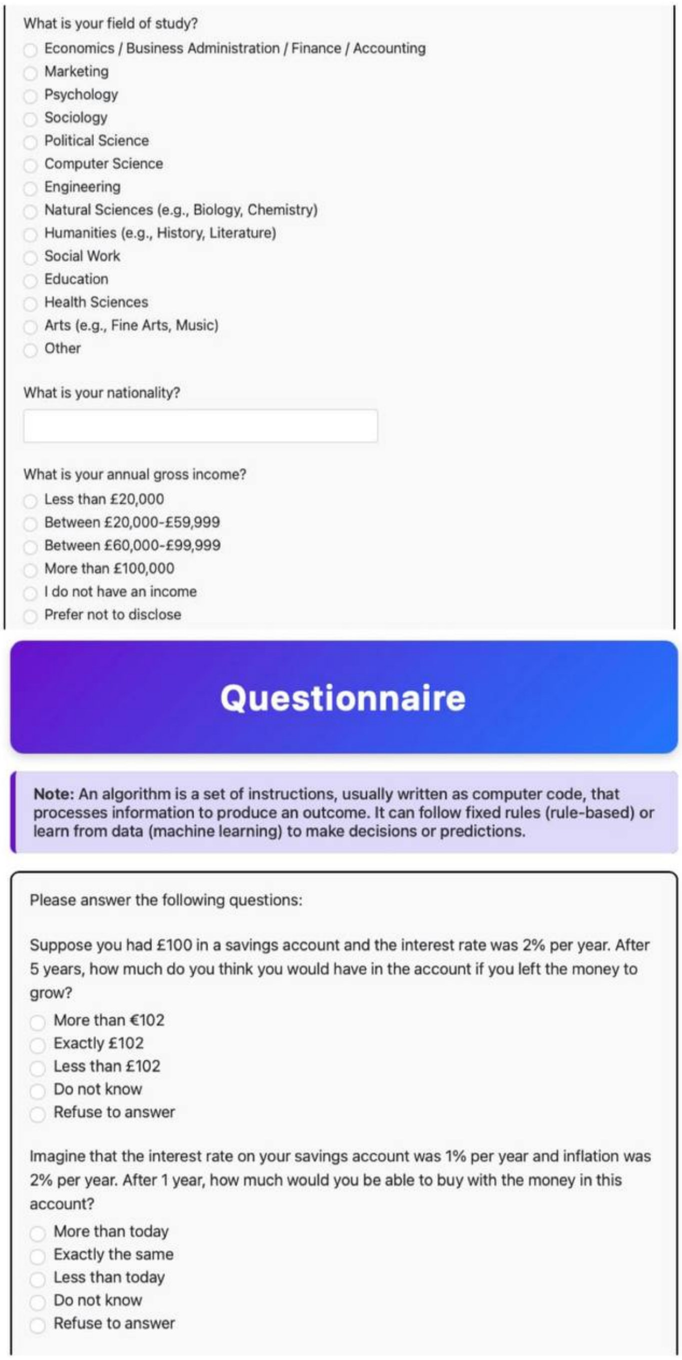
What is your gender?

○ Male
○ Female
○ Other

What is your highest level of education?

○ High School
○ Bachelor Degree
○ Master Degree
○ Doctorate
○ Other

What is your occupation?



What is your field of study?

○ Economics / Business Administration / Finance / Accounting
○ Marketing
○ Psychology
○ Sociology
○ Political Science
○ Computer Science
○ Engineering
○ Natural Sciences (e.g., Biology, Chemistry)
○ Humanities (e.g., History, Literature)
○ Social Work
○ Education
○ Health Sciences
○ Arts (e.g., Fine Arts, Music)
○ Other

What is your nationality?

What is your annual gross income?

○ Less than £20,000
○ Between £20,000-£59,999
○ Between £60,000-£99,999
○ More than £100,000
○ I do not have an income
○ Prefer not to disclose

Questionnaire

Note: An algorithm is a set of instructions, usually written as computer code, that processes information to produce an outcome. It can follow fixed rules (rule-based) or learn from data (machine learning) to make decisions or predictions.

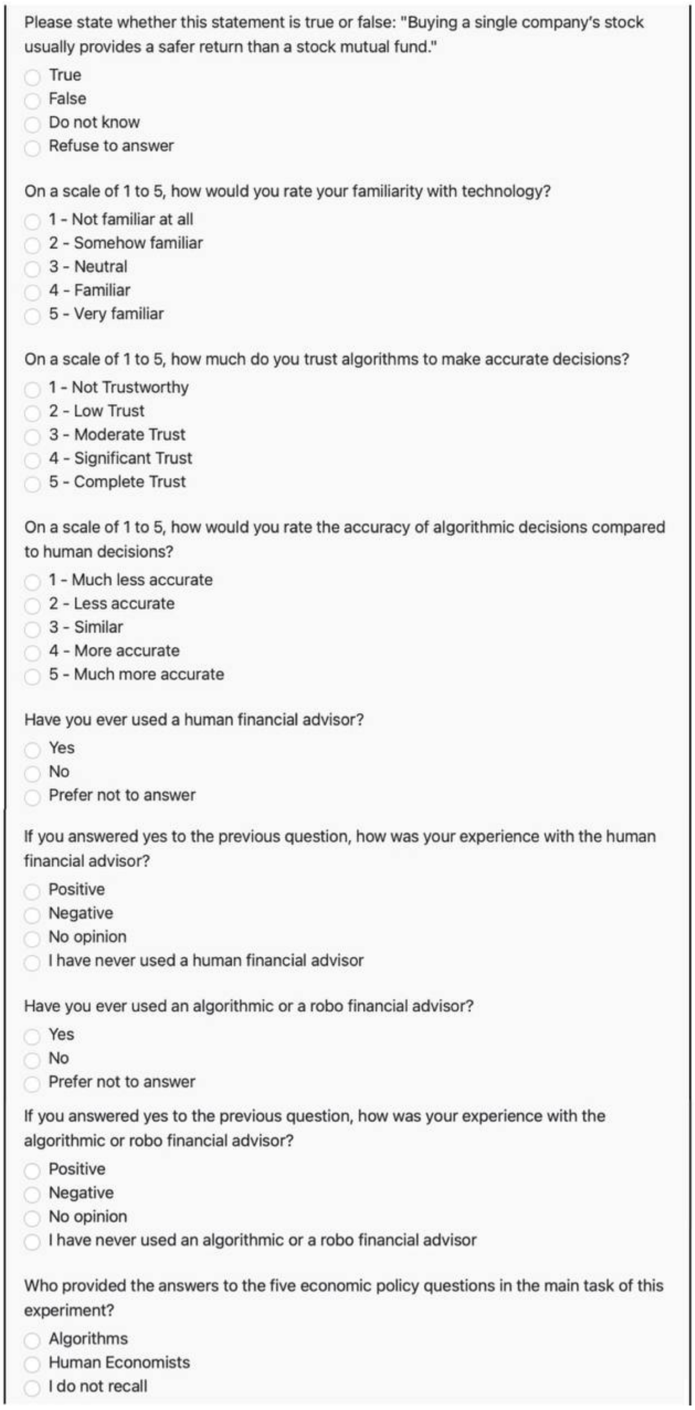
Please answer the following questions:

Suppose you had £100 in a savings account and the interest rate was 2% per year. After 5 years, how much do you think you would have in the account if you left the money to grow?

○ More than €102
○ Exactly £102
○ Less than £102
○ Do not know
○ Refuse to answer

Imagine that the interest rate on your savings account was 1% per year and inflation was 2% per year. After 1 year, how much would you be able to buy with the money in this account?

○ More than today
○ Exactly the same
○ Less than today
○ Do not know
○ Refuse to answer



Please state whether this statement is true or false: "Buying a single company's stock usually provides a safer return than a stock mutual fund."

○ True
○ False
○ Do not know
○ Refuse to answer

On a scale of 1 to 5, how would you rate your familiarity with technology?

○ 1 - Not familiar at all
○ 2 - Somehow familiar
○ 3 - Neutral
○ 4 - Familiar
○ 5 - Very familiar

On a scale of 1 to 5, how much do you trust algorithms to make accurate decisions?

○ 1 - Not Trustworthy
○ 2 - Low Trust
○ 3 - Moderate Trust
○ 4 - Significant Trust
○ 5 - Complete Trust

On a scale of 1 to 5, how would you rate the accuracy of algorithmic decisions compared to human decisions?

○ 1 - Much less accurate
○ 2 - Less accurate
○ 3 - Similar
○ 4 - More accurate
○ 5 - Much more accurate

Have you ever used a human financial advisor?

○ Yes
○ No
○ Prefer not to answer

If you answered yes to the previous question, how was your experience with the human financial advisor?

○ Positive
○ Negative
○ No opinion
○ I have never used a human financial advisor

Have you ever used an algorithmic or a robo financial advisor?

○ Yes
○ No
○ Prefer not to answer

If you answered yes to the previous question, how was your experience with the algorithmic or robo financial advisor?

○ Positive
○ Negative
○ No opinion
○ I have never used an algorithmic or a robo financial advisor

Who provided the answers to the five economic policy questions in the main task of this experiment?

○ Algorithms
○ Human Economists
○ I do not recall

Treatment 2 follows the same structure, but includes subjective questions (See Appendix B.1). Participants assigned to the treatment with answers provided by algorithms, viewed the following: *The selected answer to each question comes from a set of responses previously generated by* ***different algorithms****, such as ChatGPT-3, Google Bard, Claude, You.com and Microsoft Copilot. These are AI chatbots that operate using deep learning algorithms trained on large datasets, allowing users to interact through an online chat interface by entering an input and receiving an output.*

Appendix D.1 Description of variables

Table D.1: Description of variables used in the analysis

Variable	Explanation
Human	1 if Human, 0 if Algorithm
Female	1 if Female, 0 if Male
Age	Continuous variable
Education	Categorical variable 0 = other 1 = high school 2 = bachelor's degree 3 = master's degree 4 = doctorate
Salary	Categorical variable 1 = Less than £20,000 2 = Between £20,000-£59,999 3 = Between £60,000-£99,999 4 = More than £100,000
Subjective	1 if Treatment 2, 0 if Treatment 1
Average Trust	Average score from the trust elicitation task
Self-reported Trust	Categorical variable for the question: *On a scale of 1 to 5, how much do you trust algorithms to make accurate decisions* 1 = Not Trustworthy 2 = Low Trust 3 = Moderate Trust 4 = Significant Trust 5 = Complete Trust

Appendix D.2 Extra Analysis - Attention Check

The final question of the experiment served as an attention check. Specifically, participants were asked: *"Who provided the answers to the five economic questions in the main task of this experiment?"* The response options were "Algorithms", "Human Economists", and "I do not recall". Tables D.2.1 - D.2.3 exclude a total of 58 participants who failed to correctly answer the attention check question, and report OLS regression results corresponding to Hypotheses 1 to 5. The results remain consistent with those reported in Tables 4, 5 and 6 in the main analysis, indicating that the main findings are robust to the exclusion of inattentive participants.

Table D.2.1: OLS Regression Results - Hypotheses H1 to H3 with Attention Check

	Net Votes		
OLS Regression	**Treatment 2**	**Treatment 1**	**Both**
	(1)	**(2)**	**(3)**
Human	0.221	-0.031	-0.031
	(0.393)	(0.439)	(0.439)
Subjective			1.219***
			(0.352)
Human × Subjective			0.252
			(0.589)
Intercept	0.727***	-0.492*	-0.492*
	(0.224)	(0.271)	(0.271)
Observations	105	109	214
R-squared	0.003	0.000	0.094
Equivalence Test (TOST)	Equivalent	Equivalent	–
Hypothesis	H1	H2	H3

Notes: Robust standard errors in parentheses. *p<0.10, **p<0.05, ***p<0.01. $\Delta = 1.5$.

Table D.2.2: OLS Regression Results – Hypothesis H4 with Attention Check

	Net Votes	
OLS Regression	**Algorithm Treatment**	**Algorithm Treatment**
	(1)	**(2)**
Average Trust	-0.402	
	(0.437)	
Self-reported Trust		0.607^{***}
		(0.216)
Intercept	0.011	-1.751^{***}
	(0.009)	(0.662)
Observations	133	133
R-squared	0.009	0.046
Hypothesis	H4	H4

Notes: Robust standard errors in parentheses. *p<0.10, **p<0.05, ***p<0.01.

Table D.2.3: OLS Regression Results - Hypotheses H5 with Attention Check

	Net Votes			
	Algorithm Treatment			
OLS Regression	**(1)**	**(2)**	**(3)**	**(4)**
Female	-0.324			-0.285
	(0.364)			(0.362)
Age		-0.011		-0.008
		(0.014)	0.127	(0.014)
Education			(0.215)	0.082
				(0.212)
Intercept	0.281	0.587	-0.137	0.475
	(0.237)	(0.658)	(0.463)	(0.781)
Observations	133	133	133	133
R-squared	0.006	0.004	0.003	0.010
Hypothesis	H5	H5	H5	H5

Notes: Robust standard errors in parentheses. *p<0.10, **p<0.05, ***p<0.01.

Appendix D.3 Extra Analysis - Probit Regression Results

Probit regression results are presented in Table D.3. The binary depended variable equals 1 if a participant's net vote is positive (i.e., more answers evaluated as correct than incorrect), and 0 otherwise. The table also reports average marginal effects to facilitate interpretation. Consistent with the findings in Table 4, the results show no statistically significant evidence of algorithmic aversion. Specifically, the coefficient for 'human' is not statistically significant across all regressions, and the interaction term in Column (3) between 'human' and 'subjective' is insignificant, as well.

Table D.3: Probit Regression Results - Hypotheses H1 to H3

	Positive Net Votes		
Probit Coefficients	**Treatment 2**	**Treatment 1**	**Both**
	(1)	**(2)**	**(3)**
Human	0.106	-0.052	-0.052
	(0.222)	(0.217)	(0.217)
Subjective			0.561**
			(0.218)
Human × Subjective			0.159
			(0.310)
Intercept	0.338**	-0.223	-0.223
	(0.155)	(0.153)	(0.153)
Observations	135	137	272
Pseudo R-squared	0.001	0.000	0.004
Marginal Effects			
Human	0.039	-0.020	-0.020
	(0.082)	(0.084)	(0.082)
Subjective			0.212***
			(0.079)
Human × Subjective			0.060
			(0.117)
Hypothesis	H1	H2	H3

Notes: Robust standard errors in parentheses. *p<0.10, **p<0.05, ***p<0.01.

Appendix D.4 Exploratory Analysis

Figure 3 shows the distribution of total correct answers for participants in Treatments 1 and 2, excluding those in the control group. The y-axis indicates the number of participants per score level (0–5). Most participants, and specifically 90, have found exactly 3 correct answers. Table D.4.1 reports OLS results comparing the number of correct answers between Treatments 1 and 2. As the table suggests, participants in Treatment 2 evaluated correctly, on average, 1.18 more answers than those in Treatment 1 ($p = 0.000$).

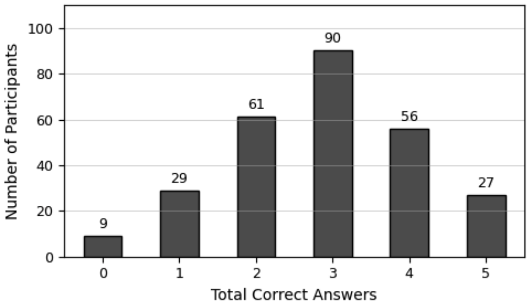


Figure III: Participant Performance Distribution

Table D.4.1: OLS Regression Results – Total Correct Answers per Treatment

Total Correct Answers	
OLS Regression	**(1)**
Treatment 2	1.189***
	(0.132)
Intercept	2.277***
	(0.098)
Observations	272
R-squared	0.231

Notes: Robust standard errors in parentheses.
*p<0.10, **p<0.05, ***p<0.01.

Table D.4.2 presents results from an OLS regression analysis on ‘Net Votes’ using various control variables. The results indicate that subjective questions received significantly higher ‘Net Votes’, while other predictors such as trust, gender, and age are not statistically significant. Education and salary are marginally significant, suggesting a possible role in evaluations.

Table D.4.2: OLS Regression Results - Exploratory Analysis on Net Votes

	Net Votes
OLS Regression	**(1)**
Human	-0.0788
	(0.393)
Average Trust	0.011
	(0.008)
Subjective	1.084***
	(0.370)
Human × Subjective	0.376
	(0.512)
Age	-0.004
	(0.011)
Female	-0.052
	(0.262)
Education	-0.265*
	(0.151)
Salary	0.354*
	(0.200)
Intercept	-1.004
	(0.804)
Observations	259
R-squared	0.124

Notes: Robust standard errors in parentheses.
*p<0.10, **p<0.05, ***p<0.01.

Lastly, to assess whether the finding of no algorithmic aversion arises due to a cancellation of belief-based and preference-based partialities, a test was conducted using only the questions with high participant agreement. This means questions that have been voted as correct by more than 70% of the participants. The specific criterion approximates conditions of high information precision. Under such circumstances, belief-based differences are expected to be minimal, allowing any remaining difference to reflect preference-based partiality. Two questions (Q2 and Q3) met this criterion, both from the subjective treatment. The mean 'Net Votes' are 1.13 for human and 1.24 for algorithm answers, showing a nonsignificant difference of –0.10 ($p = 0.60$). A TOST equivalence test confirms statistical equivalence ($p < 0.001$).

Table D.4.3: OLS and Equivalence Test for High-Agreement Questions

Net Votes	
OLS Regression	**Treatment 2**
	(1)
Human	-0.101
	(0.190)
Intercept	1.235***
	(0.133)
Observations	135
R-squared	0.002
Group Means	
Human	1.134
Algorithm	1.235
Difference	-0.101
Equivalence Test (TOST)	Equivalent

Notes: Robust standard errors in parentheses.

*p<0.10, **p<0.05, ***p<0.01.$\Delta = 1$

References

Alysandratos, T., A. Boukouras, S. Geōrganas, and Z. Maniadis (2020). *The expert and the charlatan: an experimental study in economic advice.* Social Science Research Network.

Araujo, T., N. Helberger, S. Kruikemeier, and C. H. De Vreese (2020). In ai we trust? perceptions about automated decision-making by artificial intelligence. *AI & society 35*(3), 611–623.

Barber, B. M. and T. Odean (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The quarterly journal of economics 116*(1), 261–292.

Bohren, J. A., A. Imas, and M. Rosenberg (2019). The dynamics of discrimination: Theory and evidence. *American economic review 109*(10), 3395–3436.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2016). Stereotypes. *The Quarterly Journal of Economics 131*(4), 1753–1794.

Buchanan, J. and W. Hickman (2024). Do people trust humans more than chatgpt? *Journal of Behavioral and Experimental Economics 112*, 102239.

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

Chak, I., K. Croxson, F. D'Acunto, J. Reuter, A. G. Rossi, and J. Shaw (2022). *Robo-advice for borrower repayment decisions.* Financial Conduct Authority.

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

Comin, D. and M. Mestieri (2014). Technology diffusion: Measurement, causes, and consequences. In *Handbook of economic growth*, Volume 2, pp. 565–622. Elsevier.

David, D. and O. Sade (2018). Robo-advisor adoption, willingness to pay, and, trust: an experimental investigation. Technical report, Working paper.

Dietvorst, B. J. and S. Bharti (2020). People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological science 31*(10), 1302–1314.

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

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

Dijkstra, J. J., W. B. Liebrand, and E. Timminga (1998). Persuasiveness of expert systems. *Behaviour & Information Technology 17*(3), 155–163.

Dovidio, J. F., K. Kawakami, and S. L. Gaertner (2002). Implicit and explicit prejudice and interracial interaction. *Journal of personality and social psychology 82*(1), 62.

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

Gogoll, J. and M. Uhl (2018). Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics 74*, 97–103.

Holzmeister, F., M. Holmén, M. Kirchler, M. Stefan, and E. Wengström (2023). Delegation decisions in finance. *Management Science 69*(8), 4828–4844.

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

imental Economics 112, 102270.

Joe, J., B. Commerford, S. Dennis, and J. Wang (2019). Complex estimates and auditor reliance on artificial intelligence.

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

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science 8*(4), 355–362.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big data & society 5*(1), 2053951718756684.

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

Lusardi, A. and O. S. Mitchell (2008). Planning and financial literacy: How do women fare? *American economic review 98*(2), 413–417.

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

Niszczota, P. and D. Kaszás (2020). Robo-investment aversion. *Plos one 15*(9), e0239277.

Oksanen, A., N. Savela, R. Latikka, and A. Koivula (2020). Trust toward robots and artificial intelligence: An experimental approach to human–technology interactions online. *Frontiers in Psychology 11*, 568256.

Ozkes, A. I., N. Hanaki, D. Vanderelst, and J. Willems (2024). Ultimatum bargaining: Algorithms vs. humans. *Economics Letters 244*, 111979.

Romer, P. M. (1990). Endogenous technological change. *Journal of political Economy 98*(5, Part 2), S71–S102.

Schaap, G., T. Bosse, and P. Hendriks Vettehen (2024). The abc of algorithmic aversion: Not agent, but benefits and control determine the acceptance of automated decision-making. *AI & society 39*(4), 1947–1960.

Serenko, A. and O. Turel (2021). Why are women underrepresented in the american it industry? the role of explicit and implicit gender identities. *Journal of the Association for Information Systems 22*(1), 8.

Sunstein, C. R. and L. Reisch (2023). Do people like algorithms? a research strategy. *A Research Strategy (August 18, 2023).*

Turel, O. and S. Kalhan (2023). Prejudiced against the machine? implicit associations and the transience of algorithm aversion. *Mis Quarterly 47*(4).

Tversky, A. and D. Kahneman (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review 90*(4), 293.