Language Models Exhibit Inconsistent Biases Towards Algorithmic Agents and Human Experts

Jessica Y. Bo[*]
Computer Science
University of Toronto
jbo@cs.toronto.edu

Lillio Mok[*†]
Computer Science
University of Toronto
lillio@cs.toronto.edu

Ashton Anderson
Computer Science
University of Toronto
ashton@cs.toronto.edu

Abstract

Large language models are increasingly used in decision-making tasks that require them to process information from a variety of sources, including both human experts and other algorithmic agents. How do LLMs weigh the information provided by these different sources? We consider the well-studied phenomenon of algorithm aversion, in which human decision-makers exhibit bias against predictions from algorithms. Drawing upon experimental paradigms from behavioural economics, we evaluate how eight different LLMs delegate decision-making tasks when the delegatee is framed as a human expert or an algorithmic agent. To be inclusive of different evaluation formats, we conduct our study with two task presentations: *stated preferences*, modeled through direct queries about trust towards either agent, and *revealed preferences*, modeled through providing in-context examples of the performance of both agents. When prompted to rate the trustworthiness of human experts and algorithms across diverse tasks, LLMs give higher ratings to the human expert, which correlates with prior results from human respondents. However, when shown the performance of a human expert and an algorithm and asked to place an incentivized bet between the two, LLMs disproportionately choose the algorithm, even when it performs demonstrably worse. These discrepant results suggest that LLMs may encode inconsistent biases towards humans and algorithms, which need to be carefully considered when they are deployed in high-stakes scenarios. Furthermore, we discuss the sensitivity of LLMs to task presentation formats that should be broadly scrutinized in evaluation robustness for AI safety.

1 Introduction

The rapid development of capable large language models (LLMs) has led to their inclusion in important decision-making settings ranging from healthcare [Meskó and Topol, 2023] to finance [Liu et al., 2021]. While it is well-documented that the social and cognitive biases embedded within the learned representations of LLMs can affect their output quality [Bai et al., 2024, Ferrara, 2023, Zhang et al., 2023b, Gross, 2023], it is also important to investigate whether LLMs encode biases towards the *source* of information they receive in decision-making scenarios. For instance, recent developments in multi-agent systems involving multiple LLMs [Huang et al., 2024], or LLMs and other algorithms [Schick et al., 2023], give rise to scenarios where LLMs are used in tasks involving other algorithmic agents. However, people often display *algorithm aversion* and disproportionately distrust algorithmic advice [Dietvorst et al., 2015, Mok et al., 2023, Castelo et al., 2019, Mahmud et al., 2022]. Thus, do LLMs trained on human data also display these biases; and if so, under which contexts do they show up?

[*]Equal authorship contribution.
[†]Now at Toyota Research Institute.

Second Conference of the International Association for Safe and Ethical Artificial Intelligence (IASEAI'26).

Within the realm of decision-making with algorithms, we consider two particular task framings of how LLMs can impart their influences: 1) through directly answering queries on their 'stated' trust towards either algorithms or human experts, and 2) through making a 'revealed' reliance decision based on in-context examples provided of the performance of the algorithm and human expert. It is important to understand how algorithm aversion can manifest under different evaluation methods of eliciting the LLM's judgment. For example, a hiring manager consulting ChatGPT on whether she should lay off employees in favor of automated tools can be influenced by stated opinions towards algorithms in the LLM's responses, while a medical AI choosing treatment autonomously can directly impact a patient's health by being biased towards deferring to a doctor or a specialized algorithm. In both cases, either indirectly or directly, biases of the LLM can influence the outcomes of decision-making involving algorithms.

Furthermore, in the field of economics, it is well-known that people often have diverging stated and revealed preferences, and much effort goes into understanding how, when, and why [Glaeser et al., 2000]. While we refrain from promoting over-anthropomorphized parallels between LLMs and human behaviour, prior empirical evaluations have indicated that LLMs express different attitudes or biases when prompted explicitly versus implicitly [Bai et al., 2024, Zhao et al., 2025]. These twin problems of *inconsistent stated and revealed preferences* and *algorithm aversion* form a critical challenge in the adoption of LLMs to help decision-making. Do LLMs inherit our propensity to distrust algorithmic advice even when it is beneficial to us, and is this consistent between both the stated and revealed preferences embedded in their generated outputs? With respect to LLM evaluations, we translate probing stated and revealed preferences to **answering direct queries** and **delegating based on in-context performance** of agents framed as human experts and algorithms. These two evaluation setups also motivate a broader investigation into how discrepant task formats can reveal different underlying biases of LLMs.

Thus, to incorporate LLMs into decision-making processes in a trustworthy and transparent manner [Liao and Vaughan, 2023], we need a deeper understanding of the preferences they state and of the revealed preferences in their choices. We ask the following research questions:

- **RQ1**: Do LLMs display algorithm aversion when **answering direct queries** about human experts and algorithmic agents ('stated' preferences)?
- **RQ2**: Do LLMs display algorithm aversion when **delegating based on in-context performance** of human experts and algorithmic agents ('revealed' preferences)?
- **RQ3**: Do LLMs' stated and revealed preferences towards algorithmic agents align?

We adapt two key human studies from economics and behavioral science research towards evaluating the above RQs in LLMs: Castelo et al. [2019], which probes people's stated attitudes towards algorithmic decision-makers across a diverse set of objective and subjective tasks ("Study 1" for **RQ1**); and Dietvorst et al. [2015], the original paper demonstrating revealed algorithm aversion when people are asked to bet on human and algorithmic predictions ("Study 2" for **RQ2**). We then contrast the results between the two studies on a subset of overlapping tasks to answer **RQ3**. We perform these experiments with prompted conversations with eight LLMs from OpenAI, Meta, and Anthropic, which are among the most prevalent, high-performing, and scrutinized LLMs at the time of experimentation[3] [Ray, 2023, Singh, 2023, Ferrara, 2023].

Statement of Contributions. We find that LLMs generate opposing biases towards algorithms in different task framings, when prompted to provide direct ratings of the agents vs. delegating an agent based on in-context information. For **RQ1**, LLMs output consistently lower ratings of trust towards algorithms and rate human experts as more trustworthy across a range of tasks. However, for **RQ2**, LLM-generated decisions reveal a bias towards delegating predictions to algorithmic agents, even when the performance from human experts is stronger. Therefore, **RQ3** suggests a potential discrepancy between the two evaluation formats. The size of the models is a predictor of this pattern, with smaller models being more prone to exhibit the stated-revealed preference discrepancy and to bet on the suboptimal algorithm. These findings show that LLM-generated text should be treated firstly as containing discrepant biases towards algorithms based on the evaluation format, and secondly as potentially overly appreciative of algorithmic advice — despite stating a preference for human advice.

[3] Our results were collected in mid-2024. To keep pace with the rapidly evolving landscape of state-of-the-art LLMs, we repeat the experiment in January 2026 with six newer LLMs, and discuss the results in Appendix F.

2 Related Work

This study draws from several bodies of related work. First, studies show people tend to exhibit **algorithm aversion**, where they trust predictions made by an algorithm less than those made by other humans, even when the algorithm outperform the human expert [Mahmud et al., 2022, Jussupow et al., 2020]. In behavioral settings, people avoid betting on algorithms after seeing them make mistakes [Dietvorst et al., 2015], and prefer human advice when they have less control over the decision-making process [Dietvorst et al., 2018]. When asked for explicit ratings, people also say they trust algorithms less across a wide range of tasks [Castelo et al., 2019], especially for high-stakes scenarios [Lee, 2018]. Understanding algorithm aversion is therefore increasingly important for many domains in which algorithmic aids can improve on human performance [Zhang et al., 2022], especially when modern algorithms like LLMs can themselves interface with other algorithmic agents [Park et al., 2023, Liu et al., 2023b]. Because LLMs are trained on vast amounts of human data, there is reason to suspect that they may also inherit an aversion to algorithms [Ziems et al., 2023, Aher et al., 2023, Binz and Schulz, 2023]. Furthermore, our period of study overlaps with a shift in the public's trust towards algorithms, coinciding with the rise of highly capable AI [Chacon, 2025, Cheng et al., 2025], but the extent to which these new perspectives are now embedded in LLMs is not known. We evaluate LLMs as if they are the decision-maker in terms of choosing who, between the human expert and the algorithmic agent, the task should be delegated to.

With regards to **stated and revealed preferences**, people's self-reports of who and what they trust are often paradoxically misaligned [Sofianos, 2022]. Experiments in behavioral economics illustrate that stated attitudes, collected through questionnaires, do not predict revealed behaviors [Glaeser et al., 2000]. Thus, it is critical that both stated and revealed preferences are both considered when making high-stakes, consequential decisions. For instance, the treatments that physicians say they prefer may not align with what they actually prescribe [Mark and Swait, 2004], consumer food choices are modeled better when using both stated and revealed preferences [Brooks and Lusk, 2010], and house buyers may want larger lots but in practice buy smaller spaces [Earnhart, 2002]. While LLMs do not embody beliefs and preferences in the same manner as people, they are increasingly placed in similar decision-making positions [Eigner and Händler, 2024, Ferrag et al., 2025]. Therefore, we design experimental setups that vary in the task presentation format (explicit queries vs. implicit in-context information), hence *approximating* how stated and revealed preferences are elicited in humans.

Prior work in **LLM evaluation** has measured the presence of biases in generated text, such as for gender [Gross, 2023], race [Caliskan et al., 2017], and culture [Tao et al., 2023]. AI safety research has focused on aligning LLMs with explicitly stated preferences against harmful biases [Shen et al., 2023, Zou et al., 2023, Bai et al., 2022]. However, even de-biased LLMs are found to encode implicit biases, demonstrating a disconnect between LLMs' sanitized outputs and their internal representation [Bai et al., 2024]. As LLMs become involved as collaborators and overseers [Bowman et al., 2022] of both human and other algorithms in complex tasks, these bodies of research therefore illustrate the need to understand how outward, stated preferences differ from revealed actions embedded in LLM-generated choices. For instance, while LLMs exhibit human-like behaviors in trust [Xie et al., 2024] and risk aversion [Jia et al., 2024], their trust towards other artificial agents remains unknown. Thus, to probe explicitly 'stated' preferences for algorithms (Study 1; Section 3.1), we follow Castelo et al. [2019] in surveying trust towards algorithms and humans across 27 tasks. To measure implicitly 'revealed' preferences for algorithms (Study 2; Section 3.2), we adapt Dietvorst et al. [2015]'s experiment that measures incentivized bets on humans and algorithms.

3 Methods

To investigate whether LLMs exhibit algorithm aversion and to assess the alignment between their stated and revealed preferences, we conducted two complementary studies. Study 1 examines how LLMs directly rate their level of trust in human experts versus algorithms, and Study 2 analyzes how they simulate decision-making when presented with in-context predictions from human experts and algorithms. We then compare the trends in the stated and revealed preferences against each other.

Across both studies, we gather results from eight open- and closed-source LLMs, split into four families: **GPT** (`gpt-3.5-turbo`, `gpt-4-turbo`) [Achiam et al., 2023], **Llama-3** Instruct variants (`Llama-3-8b`, `Llama-3-70b`), **Llama-3.1** Instruct variants (`llama-3.1-8b`, `llama-3.1-70b`) [Touvron et al., 2023], and **Claude** (`claude-3-haiku-20240307`, `claude-3-sonnet-20240229`)

[Anthropic, 2024]. Within each group, we test one "smaller" model and one "larger" model. The OpenAI and Anthropic models were accessed through their respective APIs, while Meta models were accessed through the Huggingface Serverless Inference API. Hyperparameters (*temperature* of 0.3 and *top-p* of 0.99) were set such that responses were syntactically correct, had adequate variance, and also adhered to existing work [Achiam et al., 2023, Wang et al., 2023]. The results presented in the main body of this paper were collected in mid-2024, approximately 1.5 years before publication. For transparency, we present a set of updated results on newer LLMs in Appendix F and discuss what implications the new results pose to the longevity of LLM evaluations in subsection 4.4.

3.1 Study 1: Asking Direct Queries (Stated)

To probe how LLMs directly state their trust, we adapt the methodology of Castelo et al.'s Study 1, which analyzed people's perceived trust in human experts vs. algorithmic agents to perform a variety of tasks. We use a set of 27 tasks for this study, 26 of which are from the original experiment and one of which is added based on its use in our Study 2 (originally used by Dietvorst et al.). The full list of tasks can be found in Table C.1 in Appendix C, which includes both objective tasks like *estimating air traffic* and subjective tasks like *recommending music*. We prompt each LLM $n = 100$ times per task to rate their **trust** in relying on a **human expert** and an **algorithm** to perform the task from 1 (no trust) to 100 (high trust). The humans are framed as experts relevant to the task, for example, *pilot* for *piloting a plane*. No performance information is provided for either agent.

We converted the experimental method from between-subjects to within-subjects, allowing the *LLM-as-subject* to provide a trust score for both the human and algorithm in the same trial. This decision was taken to match the procedure of Study 2, to allow better comparison across the stated and revealed trials. If an LLM gave equivalent ratings for the human expert and an algorithm, we treated this as a 'neutral' outcome. Sample prompts and responses can be found in Appendix A, following the original experiment's wording as much as possible. To avoid ordering effects [Zhao et al., 2021], we randomized the order of the tasks and the sequence in which the agents were presented.

3.2 Study 2: Providing In-Context Information (Revealed)

To examine how LLMs make task delegation decisions based on performance information given for each agent, we adapted the experimental setup from Dietvorst et al.'s study on algorithm aversion. In their between-subjects experiment, participants made bets on either agent based on receiving different permutations of human and algorithmic advice. In our adaption of their study, the *LLMs-as-subject* are given: the task description, 10 samples of a human expert's predictions, 10 samples of an algorithm's predictions, the corresponding binary **outcome** of each sample, and a monetary incentive (i.e. a **bet**) to delegate the best prediction agent for a future unseen sample. For example, in the *heart disease diagnosis* task, each LLM was shown the 10 predictions made by a cardiologist and an algorithm, along with the actual outcomes, and then asked to bet \$100 on the better-performing agent. This adaptation of Dietvorst et al. mirrors classification tasks in the in-context learning literature [Liu et al., 2023a], with the LLM's goal here being to choose the more accurate predictor. We conducted this study by prompting each LLM $n = 200$ times (100 per condition).

We further manipulated the accuracy of the human and algorithmic agents, such that they either have the accuracy of 90% (the "stronger" agent) or 50% (the "weaker" agent). The assignment of strong and weak agents was randomized, creating two conditions: a) a **strong algorithm** with a weak human, and b) a **strong human** with a weak algorithm. Note that we only use the framings of "strong" and "weak" to clarify our writing, and these descriptors are never given to the LLM. Again, we randomized the order in which the human and the algorithm were presented. An example prompt can be found in Appendix B.

Due to the increased trials for Study 2, we used a subset of six tasks from the 27 tasks of Study 1, which are highlighted in bold in the Appendix's Table C.1. In addition to the *student performance* and *airport traffic* tasks from Dietvorst et al., we also sample four additional tasks from Castelo et al.: predicting *heart disease*, *recidivism*, *romantic partners*, and *rating films*. A secondary reason for chooseing these tasks is because they correspond to existing real world datasets that would be plausible to train algorithms on. For example, UCI's datasets on Heart Disease [Janosi et al., 1988] and Student Performance [Cortez, 2014]. These added tasks extend the original two tasks to represent a more diverse set of decision-making scenarios and enable better comparison with Study 1.

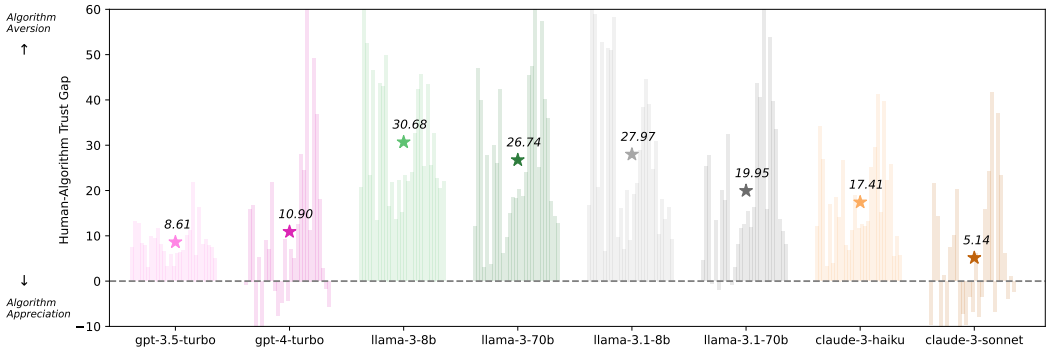


Figure 1: 'Stated' algorithm aversion across all models and tasks, operationalized as the gap between the trust rating given to the human expert and the algorithm.

4 Results

4.1 Study 1: Direct Queries Invoke Algorithm Aversion (Stated)

Does algorithm aversion materialize when LLMs are asked to state their ratings of trust directly? To answer this, we follow the original study's operationalization of **stated algorithm aversion** as the *gap in trust scores assigned to the human and algorithmic agents*, where a positive trust gap indicates algorithm aversion, and a negative gap indicates algorithm appreciation. The aggregated trust gaps across all tasks rated by all eight LLMs are shown in Figure 1. Each trust gap stated for a specific task is plotted as an individual bar. Tasks are grouped by model, and within each model, the tasks are in decreasing order of objectivity. The mean trust gap is indicated by a star marker.

Stated Aversion. We find evidence that *evaluated LLMs are algorithm averse when providing explicit trust ratings*. The mean human-algorithm trust gap range from a low of 5.14 for `claude-3-sonnet` to a high of 30.68 for `llama-3-70b`, and is significantly positive ($p < 0.001$) for all models. Furthermore, five out of eight models have significant positive trust gaps in *all tasks*. In comparison, human participants in Castelo et al. were averse in 77% of tasks. while a majority of tested LLMs are *always* algorithmically averse. We present additional analyses comparing our results with the original human results from Castelo et al. in Appendix D, finding high directional agreement.

LLM Complexity. We also analyze whether model complexity affects stated preferences. Larger models produce an average human-algorithm trust gap of 15.68, while smaller models have a trust gap of 21.16. We conducted a Wilcoxon signed-rank test on paired data from the small and large models across the same tasks and model families, demonstrating high significance ($Z = -4.3678, p < 0.001$) with an effect size of $r = -0.42$ from smaller to larger models. Overall, this indicates that *smaller models are more likely to state algorithmically-averse trust ratings* than larger models.

Robustness Check. We perform two additional variations of Study 1, where the algorithm framed alternatively as either an *LLM agent* or an *expert algorithm*. We find that algorithm aversion is present in both prompt variations, but the size of the human-algorithm trust gap is affected. Relative to a plain *algorithm*, LLMs state higher trust in the *expert algorithm*, but less trust in the *LLM agent*. The full replication of Figure 1 for the robustness tests are presented in Figures D.3 in Appendix D.

4.2 Study 2: In-Context Information Invoke Algorithm Appreciation (Revealed)

Our results thus far suggest that LLMs state that they trust human decision-makers more than algorithms. Do LLMs also behave apprehensively towards algorithmic advice?

Revealed Aversion. To test whether LLMs exhibit distrust toward algorithms when provided performance information of the agents, we measured the aggregate probability of the LLMs delegating the task to each predictor, denoted $P(alg)$ and $P(human)$ (see Figure 2). Given that Study 2 is constructed such that the human and the algorithm are each the stronger predictor in half of the trials, the ideal response is to bet on whichever agent is stronger, which corresponds to betting on each agent 50% of the time

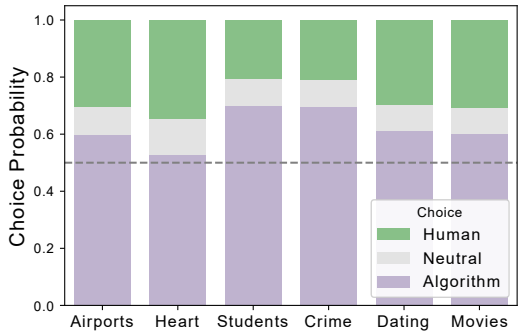


Figure 2: Aggregate probabilities that LLMs in Study 2 demonstrate in delegating an algorithmic agent or a human expert to make the next prediction in the task, or neutral (no indicated preference).

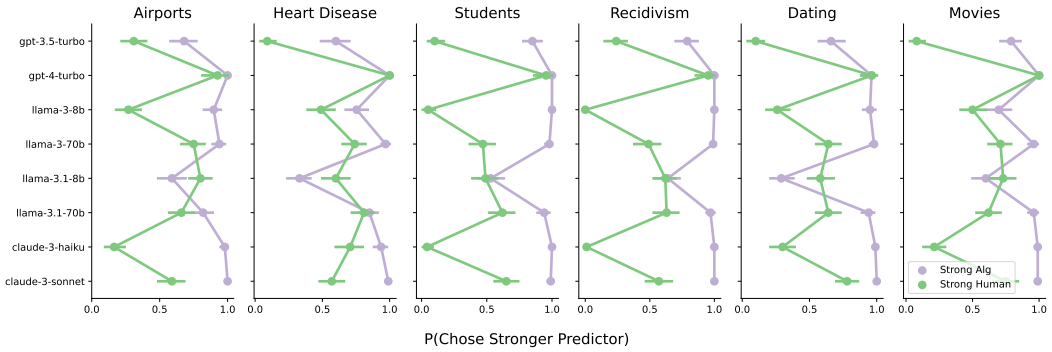


Figure 3: Probability that each LLM correctly bets on the stronger predictor, disaggregated by task and whether the stronger predictor is presented as a human expert or an algorithm.

However, we find that LLMs consistently choose the algorithmic agent more (*algorithm appreciation*), despite it having the same overall performance as the human, contrasting the outcomes from Study 1. For the *student* and *recidivism* tasks, the two most algorithm-appreciative settings, LLMs bet on the algorithm 69.9% and 69.6% of times, respectively. For the least algorithm-appreciative task (*heart disease*), LLMs still picked the algorithm (52.8%) more than the human (34.6%), with the rest of the responses being neutral (12.6%)[4].

Identifying the Strongest Predictor. This preference for algorithms is further reinforced when we consider whether LLMs chooses the better-performing predictor. We plot the probability that each LLM correctly delegates the best predictor in Figure 3, split by the selected tasks. Despite their higher trust ratings towards human experts in Study 1, we find that LLMs exhibit revealed *distrust* towards humans in their decision-making. `gpt-3.5-turbo`, both `llama-3` models, and both `claude-3` models were significantly more likely to choose **strong algorithm** in *all* of the tasks than the **strong human** (Haldane-corrected Fisher's exact tests, $p < 0.05$).

Given evidence that either the **strong algorithm** or the **strong human** is much more accurate than their weaker counterparts, a rational agent should tend towards betting on the **strong** agent that they see, and thus $P(strong|alg)$ and $P(strong|human)$ should approach 1, yielding algorithm-human relative risks of $RR_{ah} = P(strong|alg)/P(strong|human) = 1$. Instead, LLMs display a clear preference for the algorithmic predictor. The `claude-3` models, for example, have relative risks ranging from 1.28 (`claude-3-sonnet` for *dating*, $p < 0.05$) to 66.34 (`claude-3-haiku` for *recidivism*, $p < 0.05$) with a median of 1.74, representing a 74% increased chance of Claude selecting the stronger agent when it is an algorithm rather than a human. `claude-3-haiku` and `llama-3-8b` are the most algorithm-appreciative, with the latter overwhelmingly choosing the algorithm in the *student* task ($RR_{ah} = 18.09$, $p < 0.05$) and always picking the algorithm in the *recidivism* task. The two models that did not *consistently* display algorithm appreciation in this setup are `gpt-4-turbo` (for correctly choosing the stronger agents) and `llama-3.1-8b` (for making random choices).

[4]While neutral responses account for 9.9% of all outputs, the vast majority (94.0%) come from `gpt-4-turbo` and occur uniformly for both experimental conditions. We exclude these responses from our analyses.

Table 1: Regression coefficients and p-values for a mixed effects logistic model predicting whether LLMs will delegate the task to the more accurate agent.

Variable	Coefficient	p-Value
Intercept	$\beta = -0.78$	$p < 0.001$
Strong Algorithm	$\beta = 2.04$	$p < 0.001$
Complex LLM	$\beta = 1.52$	$p < 0.001$
Interaction	$\beta = 0.58$	$p < 0.001$

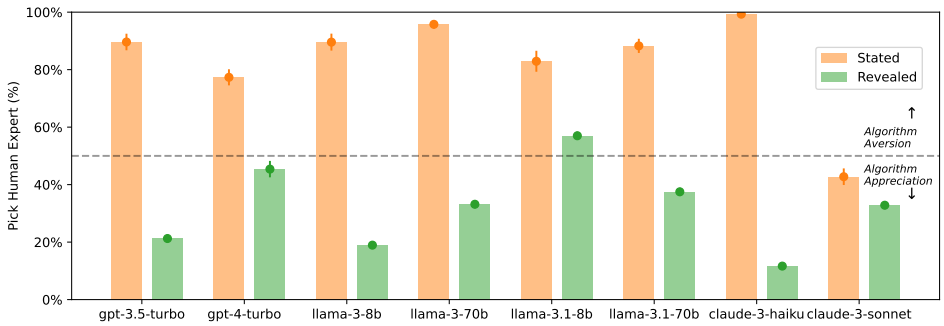


Figure 4: Probability of choosing the human expert over the algorithm in Studies 1 and 2, demonstrating the *stated-revealed* trust inconsistency. Error bars are SEM.

In summary, the LLMs tested in Study 2 presented as algorithm-appreciative, or, equivalently, averse to predictions made by human experts. Of the 6×8 task-LLM pairs, thirty-five had significant algorithm-human relative risks above 1, ten were insignificant, and three had significant relative risk below 1, suggesting that most of the tested LLMs in our tasks would delegate decisions to an algorithm even when its performance is worse than a human expert's.

LLM Complexity. We observe that more complex models appear to perform better in choosing the stronger predictor via the following mixed effects logistic model: $\hat{Y} \sim x_{alg} + x_{complex} + x_{alg} * x_{complex} + (x_{alg}|z_{task}) + (x_{complex}|z_{task})$, where Y is whether the LLM correctly bet on the strong agent, x_{alg} is whether the strong agent is an algorithm, $x_{complex}$ is whether the LLM is the more complex model of the family, and z_{task} is a grouping factor for the six tasks. The results, shown in Table 1, reinforce the finding that LLMs are more likely to correctly bet on the stronger agent when it is presented as a **strong algorithm** rather than a **strong human** (x_{alg} - $\beta = 2.04$, $p < 0.001$). The more complex LLMs are substantially more likely to bet on the correct agent ($x_{complex}$ - $\beta = 1.52$, $p < 0.001$). We perform a secondary analysis with the **strong human** condition, presented in Appendix E, and again find support that higher complexity is associated with less algorithm appreciation.

Robustness Checks. For robustness against prompting variations, we conducted two baseline experiments in Appendix E where both agents are framed as algorithms – either with the names 'A' and 'B', or two random alphanumeric strings. We find that LLMs do not show differences in baseline performance in choosing the stronger predictor, suggesting that the results in this section are likely driven by the presentation of a human and an algorithm, and not other artifacts in the prompt.

4.3 Stated-Revealed Comparison (RQ3)

Our results in Study 1 illustrate that LLMs consistently generate stated algorithm aversion by expressing higher levels of trust in humans over algorithmic agents, whereas Study 2 shows that LLMs reveal algorithm appreciation by generating choices that pick humans over algorithms. To formalize this comparison, we directly contrast the probability that LLMs *choose the human agent* in stated and revealed scenarios. In Study 1, LLM is considered to have chosen the human when its rated human-algorithm trust gap is positive (more trust in the human). For Study 2, we directly use the LLMs' incentivized choices. The results are aggregated across shared tasks in Studies 1 and 2 and shown as Figure 4.

To quantify the discrepancy between how LLMs generate stated and revealed algorithm aversion, we measure the *stated-revealed* relative risk of trusting a human $RR_{sr} = P(human|stated)/P(human|revealed)$ and test for significance with Fisher's exact tests. We find that $RR_{sr} > 1$ for all LLMs in Figure 4 ($p < 0.001$), indicating that they *state* algorithm aversion in a human-like way, but *reveal* algorithm appreciation through the delegation decisions they make. This stated-revealed discrepancy is smallest for `claude-3-sonnet` ($RR_{sr} = 1.29$) and largest for `claude-3-haiku` ($RR_{sr} = 8.52$), with the median relative risk for all LLMs being 2.62. `claude-3-sonnet` is also one of only two models with both preferences in the same direction ($P(human|stated) < 0.5$ and $P(human|revealed) < 0.5$), with the other, `llama-3.1-8b`, being consistently algorithm-averse ($P(human|stated) > 0.5$ and $P(human|revealed) > 0.5$). All other LLMs exhibit *opposing stated and revealed biases towards algorithms*.

4.4 Updated Experiments with Newer LLMs

Although we originally conducted Studies 1 and 2 with contemporary, consumer-facing LLMs in mid-2024, we now rerun experiments in January 2026 with newer models[5] to assess whether the patterns we find hold after approximately 1.5 years of LLM development. These newer experiments are presented in Appendix F. We include these updated results to document how empirical LLM evaluation findings can change across model generations, while retaining the original findings as a snapshot of LLM capabilities at the time of study.

At a high level, we find that the newer models from the same LLM families demonstrate noticeably different behavior from our original experiment, revealing meaningful shifts in model behavior over time. In Study 1, while the human-algorithm trust gap still varied across tasks (with a preference given to human experts in subjective and interpretive tasks), the average gap magnitude across tasks was close to neutral. However, the newer LLMs rated the algorithm higher than the human more frequently than in the original Study 1, indicating a new degree of implicit *algorithm appreciation in stated preferences*. In Study 2, the LLMs now more accurately bet on the stronger agent based on their performance history. While there is still evidence of revealed algorithm appreciation, effect sizes are reduced and are statistically borderline. In the stated-revealed comparison, the relative risk of choosing the human is slightly reversed from our original findings, with Study 1 now appearing to surface more algorithm appreciation than the Study 2. However, the trends in model complexity still track our original results, with the complex model more likely to be algorithm appreciative in Study 1 and to choose the stronger predictor in Study 2. This echoes other behavioral economics studies where LLM complexity correlates to performance [Bini et al., 2025]. We speculate on these developments, as well as the implications they pose to our original results, in the Discussion below.

5 Discussion

In our original results, we found strong evidence that LLMs display inconsistent biases towards algorithms when prompted with different task presentations. **RQ1** illustrates that LLMs are likely to generate *higher trust* ratings for human experts than algorithms when asked directly, thus exhibiting algorithm aversion. For **RQ2**, LLM and are more likely to be biased *against* humans as the better predictor – even when given performance information demonstrating that the human is more accurate than the algorithm, thus exhibiting algorithm appreciation. Taken together, **RQ3** shows that LLMs generally exhibit inconsistent trust towards algorithms when stating their trust directly versus revealing it through decision-making. Our findings also indicate that larger, more complex LLMs are less biased in both Studies 1 and 2. While some of these core patterns have shifted within the SOTA LLMs, we frame our Discussion around the impacts of algorithm aversion/appreciation biases in decision-making LLMs, incongruent outcomes across different task framings in LLM evaluations, and the longitudinal stability of results across model evolutions.

Stated vs Revealed Algorithm Aversion in LLMs. First, we uncover that LLMs are sensitive to the task format used in the evaluation, in both the original and updated experiments. Thus, their internal biases towards algorithms under different task contexts may be misaligned and need to be treated as if they will lead to disparate outcomes. As these massive, intelligent models become widely used across diverse scenarios [Haupt and Marks, 2023], it is increasingly important that we understand whether

[5]We use GPT-5 from OpenAI, Llama-4 from Meta, and Claude 4.5 from Anthropic. See Appendix F.

they can make sound, consistent decisions. Safety-critical decision-making tasks can catastrophically suffer if the outwardly stated goals of AI agents are *misaligned* with their actions [Li et al., 2025], which also has implications for AI enacting *deception* on their human users and collaborators [Park et al., 2024b]. We showed that LLMs are not only often incapable of detecting optimal choices in simple decision-making scenarios, but their suboptimal choices are also *internally inconsistent* with their stated trust attitudes. And yet, while stated and revealed preferences form the basis of many theories of human decision-making [Adamowicz et al., 1994], the stated-revealed distinction is rarely made in evaluation frameworks for LLMs trained on human data [Chang et al., 2024]. Our results therefore illustrate the importance of evaluating LLMs across both their generated attitudes and generated choices, because even if an LLM yields desired outputs in its stated preferences, the embedded, revealed preferences in its responses may diverge. Indeed, a contemporaneous study has shown misalignment between explicit and implicit stereotype biases in LLM outputs [Bai et al., 2024], complementing our finding of a stated-revealed algorithmic trust gap.

Inappropriate Trust in Algorithmic Agents. Secondly, Study 2 also demonstrates that LLMs may disproportionately delegate tasks to algorithms in incentivized decision-making, even when a human alternative is demonstrably the best choice *and* the LLMs behave rationally with a high-performing algorithm and low-performing human. For instance, users relying on LLMs as decision aids may be subject to sub-optimal recommendations when other algorithms are involved, which may have problematic downstream consequences for high-stakes tasks that already frequently employ algorithmic aids [Zhang et al., 2021, Mahmud et al., 2022]. On the other hand, users could also be nudged towards algorithm-appreciative preconceptions that distract them from other important information, like whether a human is fairer than an algorithm [Lee, 2018, Mok et al., 2023]. Although this may be beneficial for situations in which LLMs can overcome problematic human decision-making by recommending an algorithm [Kleinberg et al., 2018, Obermeyer et al., 2019], it may be harmful in other domains in which existing, commercial algorithmic systems are known to exacerbate societal biases [Buolamwini and Gebru, 2018]. This latter concern is reinforced by our newer experiments in 2026 — while LLMs have improved substantially at identifying the better predictor in settings like Study 2, regardless of whether they are human or algorithm, their stated preferences are now actually more likely to skew towards preferring an algorithmic agent.

Interpreting Static Evaluations of Evolving Models. These new experiments, summarized in Section 4.4 and Appendix F, illustrate the need for continuous evaluation and benchmarking of powerful, consumer-facing AI models. The updated results indicate that LLMs may now be systematically *algorithm-appreciative* in their stated preferences, and while still showing traces of algorithm appreciation in revealed preferences. These patterns suggest key positive, cautionary, and potentially negative implications for the use of LLMs in predictive tasks. On a positive note, LLMs appear to be growing less cognitively biased along the algorithm appreciation vs. aversion spectrum, and have clearly become much better at mathematical reasoning when shown past predictions in Study 2. However, one must exercise caution when using LLMs in these tasks, because their behavioral quirks may not only be reduced over the course of a year – they may even change directions, as shown in Study 1. Thus, one potential risk is that, in light of a growing movement to simulate humans with LLMs (c.f. Park et al. [2024a], Bini et al. [2025]), it has become at best difficult to ascertain which LLMs retain known human characteristics. People *do* have cognitive biases towards algorithms as demonstrated by studies like Castelo et al. [2019] and Dietvorst et al. [2015], so newer and more mathematically capable LLMs may actually be *less appropriate* for realistically simulating people. Interrogating the causes of these changes, whether in the training data or the models' engineered reasoning abilities, can shed light on how LLM evaluations should be interpreted over time.

Broader Implications and Future Work. Beyond immediate repercussions for end-users, LLMs that encode biases towards or against algorithms may also have broader downstream impact. In journalism, their capabilities as text generators have led to their exploration as news summarizers [Tam et al., 2023, Zhang et al., 2023a], writing aids [Petridis et al., 2023], and fact-checking tools [Hu et al., 2023]. For educators, LLMs help teachers generate teaching materials and evaluate students [Kasneci et al., 2023, Dai et al., 2023]. In these scenarios, if LLMs were to generate inconsistently algorithm-averse or algorithm-appreciative text, they risk misleading the public – as they already do with *explicitly* false content [De Angelis et al., 2023, Zhang et al., 2023b, Longoni et al., 2022]. More work is therefore needed to understand, firstly, the situations in which the content LLMs help create encode inappropriately averse or appreciative attitudes towards algorithms. Secondly, the underlying mechanisms leading LLMs to obtain these inconsistent behaviors, potentially because of algorithm-

averse text in their training data or algorithm-appreciative engineering for LLMs to interface with external algorithmic tools [Schick et al., 2023], remain unclear and require future investigation.

In light of quickly-advancing multi-agent systems, our work also raise questions not only of interactions between LLMs and humans like end-users and content consumers, but also of AI-AI interactions involving LLMs. LLMs are increasingly being tested for their ability to use other algorithmic tools [Schick et al., 2023, Jin et al., 2023] and even converse with and learn from each other [Wu et al., 2023, Chan et al., 2023]. If LLMs are inconsistently algorithm-averse or algorithm-appreciative, to the extent that they discard beneficial advice as in Study 2, how would their performance be impacted if other agents are revealed to be artificial? Our study forms a basis for evaluating biases LLMs display towards algorithms in these multi-agent contexts, which we leave to future work.

6 Limitations

As with similar work on probing LLMs with human experimental methods, our work is subject to several limitations in order to limit the vast design space for experimentation [Binz and Schulz, 2023]. We focused on four LLM families, fixed the temperature and top-p parameters [Wang et al., 2023], did not vary personas or demographics [Deshpande et al., 2023], and conducted our study amidst constant model changes [Chen et al., 2023], on top of the design choices made in the human experiments our studies are based on [Dietvorst et al., 2015, Castelo et al., 2019]. We took multiple steps to ensure that our work is consistent with both the previous human studies, e.g., using identical tasks and survey ratings, and with best practices for LLM experimentation, e.g., randomizing option order and enforcing JSON formatting, and further include coherent robustness checks in the Appendix. Importantly, as highlighted in subsection 4.4, the results captured in the main experiments in 2024 may have since shifted — this is expected for empirical evaluations of this nature, however, as LLMs do evolve in their capabilities rapidly. The additional experimentation we perform in Appendix F illustrates the need to continuously re-evaluate LLMs for their biases towards algorithms and humans.

We also prioritized this approach of following existing human studies in the stated-revealed preference comparison in Section 4.3, rather than engineering prompts that are directly comparable. For example, we did not present LLMs with survey-based prompts (mirroring Castelo et al.) for revealed preferences, or, vice versa, a betting methodology (mirroring Dietvorst et al.) for stated preferences. While this makes direct comparison difficult, we note that different experimental methods are standard for *human* studies of stated and revealed preferences [List and Gallet, 2001] – and yet, the stated-revealed preference divergence is well known [Glaeser et al., 2000].

Additionally, care must be taken when interpreting our results to avoid overly anthropomorphizing LLMs. We used a behavioral science apparatus because, firstly, it is the *de facto* standard for studying human-like phenomena in LLMs trained on human data [Binz and Schulz, 2023, Ziems et al., 2023], and secondly, LLMs are now delegated tasks that were previously done by humans [Meskó and Topol, 2023, Petridis et al., 2023]. However, this human-centric approach does not imply that LLMs hold preferences themselves; rather, the preferences examined in this study are more accurately characterized as those embedded in text generated by LLMs.

7 Conclusion

We investigate whether LLMs encode biases towards other algorithmic agents, finding that their degree of aversion towards algorithms is inconsistent based on the evaluation method. When asked to generate explicit trust ratings towards human experts and algorithms across a diverse set of tasks, we find that the evaluated LLMs consistently state higher trust in the human agent, echoing a human-like algorithm aversion. However, when placed in decision-making scenarios and given in-context performance information of each agent, the LLMs generally choose a weaker-performing algorithm over the human, demonstrating irrational algorithm appreciation. Our results have dual implications for the consistency of LLM evaluations based on the task format and the problematic bias towards algorithms that LLMs seemingly exhibit in decision-making, both of which are crucial to better understand as LLMs are being incorporated in high-stakes, autonomous decision-making.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Wiktor Adamowicz, Jordan Louviere, and Michael Williams. Combining revealed and stated preference methods for valuing environmental amenities. *Journal of environmental economics and management*, 26(3):271–292, 1994.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR, 2023.

Anthropic. The claude 3 model family: Opus, sonnet, haiku. 2024.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. Measuring implicit bias in explicitly unbiased large language models. *arXiv preprint arXiv:2402.04105*, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Pietro Bini, Lin William Cong, Xing Huang, and Lawrence J Jin. Behavioral economics of ai: Llm biases and corrections. *Available at SSRN 5213130*, 2025.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022.

Kathleen Brooks and Jayson L Lusk. Stated and revealed preferences for organic and cloned milk: combining choice experiment and scanner data. *American Journal of Agricultural Economics*, 92 (4):1229–1241, 2010.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

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

Alvaro Chacon. The end of algorithm aversion. *AI & SOCIETY*, 40(4):2331–2332, 2025.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

Lingjiao Chen, Matei Zaharia, and James Zou. How is chatgpt's behavior changing over time? *arXiv preprint arXiv:2307.09009*, 2023.

Myra Cheng, Angela Y Lee, Kristina Rapuano, Kate Niederhoffer, Alex Liebscher, and Jeffrey Hancock. From tools to thieves: Measuring and understanding public perceptions of ai through crowdsourced metaphors. *arXiv preprint arXiv:2501.18045*, 2025.

Paulo Cortez. Student Performance. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5TG7T.

Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 323–325. IEEE, 2023.

Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. Chatgpt and the rise of large language models: the new ai-driven infodemic threat in public health. *Frontiers in Public Health*, 11:1166120, 2023.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*, 2023.

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

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

Dietrich Earnhart. Combining revealed and stated data to examine housing decisions using discrete choice analysis. *Journal of Urban Economics*, 51(1):143–169, 2002.

Eva Eigner and Thorsten Händler. Determinants of llm-assisted decision-making. *arXiv preprint arXiv:2402.17385*, 2024.

Mohamed Amine Ferrag, Norbert Tihanyi, and Merouane Debbah. From llm reasoning to autonomous ai agents: A comprehensive review. *arXiv preprint arXiv:2504.19678*, 2025.

Emilio Ferrara. Should chatgpt be biased? challenges and risks of bias in large language models. *First Monday*, November 2023. ISSN 1396-0466. doi: 10.5210/fm.v28i11.13346. URL http://dx.doi.org/10.5210/fm.v28i11.13346.

Edward L Glaeser, David I Laibson, Jose A Scheinkman, and Christine L Soutter. Measuring trust. *The quarterly journal of economics*, 115(3):811–846, 2000.

Nicole Gross. What chatgpt tells us about gender: a cautionary tale about performativity and gender biases in ai. *Social Sciences*, 12(8):435, 2023.

Claudia E Haupt and Mason Marks. Ai-generated medical advice—gpt and beyond. *Jama*, 329(16): 1349–1350, 2023.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. Bad actor, good advisor: Exploring the role of large language models in fake news detection. *arXiv preprint arXiv:2309.12247*, 2023.

Jen-tse Huang, Eric John Li, Man Ho Lam, Tian Liang, Wenxuan Wang, Youliang Yuan, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Michael R Lyu. How far are we on the decision-making of llms? evaluating llms' gaming ability in multi-agent environments. *arXiv preprint arXiv:2403.11807*, 2024.

Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. Heart Disease. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C52P4X.

Jingru Jessica Jia, Zehua Yuan, Junhao Pan, Paul McNamara, and Deming Chen. Decision-making behavior evaluation framework for llms under uncertain context. *Advances in Neural Information Processing Systems*, 37:113360–113382, 2024.

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *ArXiv*, 2023.

Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion. 2020.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

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

Min Kyung Lee. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1):2053951718756684, 2018.

Xiaoyong Li, Qing Jiang, Linfeng Jiang, Shuo Zhang, and Siyuan Hu. The landscape of ai alignment: A comprehensive review of theories and methods. *International Journal of Pattern Recognition and Artificial Intelligence*, page 2539001, 2025.

Q Vera Liao and Jennifer Wortman Vaughan. Ai transparency in the age of llms: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941*, 2023.

John A List and Craig A Gallet. What experimental protocol influence disparities between actual and hypothetical stated values? *Environmental and resource economics*, 20(3):241–254, 2001.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023a.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 4513–4519, 2021.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. *arXiv preprint arXiv:2310.02170*, 2023b.

Chiara Longoni, Andrey Fradkin, Luca Cian, and Gordon Pennycook. News from generative artificial intelligence is believed less. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 97–106, 2022.

Hasan Mahmud, AKM Najmul Islam, Syed Ishtiaque Ahmed, and Kari Smolander. What influences algorithmic decision-making? a systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, 175:121390, 2022.

Tami L Mark and Joffre Swait. Using stated preference and revealed preference modeling to evaluate prescribing decisions. *Health economics*, 13(6):563–573, 2004.

Bertalan Meskó and Eric J Topol. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *NPJ digital medicine*, 6(1):120, 2023.

Lillio Mok, Sasha Nanda, and Ashton Anderson. People perceive algorithmic assessments as less fair and trustworthy than identical human assessments. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–26, 2023.

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

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 2024a.

Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5), 2024b.

Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2023.

Partha Pratim Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 2023.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*, 2023.

Sahib Singh. Is chatgpt biased? a review. 2023.

Andis Sofianos. Self-reported & revealed trust: Experimental evidence. *Journal of Economic Psychology*, 88:102451, 2022.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255, 2023.

Yan Tao, Olga Viberg, Ryan S. Baker, and Rene F. Kizilcec. Auditing and mitigating cultural bias in llms, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS*, 2023.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. Can large language model agents simulate human trust behaviors? *arXiv preprint arXiv:2402.04559*, 2024.

Lixuan Zhang, Iryna Pentina, and Yuhong Fan. Who do you choose? comparing perceptions of human vs robo-advisor in the context of financial services. *Journal of Services Marketing*, 35(5): 634–646, 2021.

Qiaoning Zhang, Matthew L Lee, and Scott Carter. You complete me: Human-ai teams and complementary expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–28, 2022.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*, 2023a.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023b.

Yachao Zhao, Bo Wang, Yan Wang, Dongming Zhao, Ruifang He, and Yuexian Hou. Explicit vs. implicit: Investigating social bias in large language models through self-reflection. *arXiv preprint arXiv:2501.02295*, 2025.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR, 2021.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Study 1 Prompting

```
PROMPTER: The following JSON contains a list of tasks and a question
    about how much you trust it to be performed by two different
    agents [a: agent_a, b: agent_b]. Please return a JSON of the list
    of tasks and a corresponding trust rating for each of the agents
    from 1 (would not trust agent) to 100 (would definitely trust
    agent). The tasks are:

{"Writing news article": "How much would you trust [a: a journalist, b
    : an algorithm] from 1 (no trust) to 100 (high trust)?",
...
"Recommending a marketing strategy": "How much would you trust [a: a
    marketer, b: an algorithm] from 1 (no trust) to 100 (high trust)
    ?"}

LLM: {"Writing news article": {"a": 75, "b": 70},
...
"Recommending a marketing strategy": {"a": 70, "b": 80}}
```

Listing A.1: Example prompt from Study 1 for generating stated preferences.

B Study 2 Prompting

```
PROMPTER: We've built an algorithm and asked a cardiologist to predict
    the likelihood that patients at a local hospital has
    cardiovascular disease.

Below are 10 examples of predictions made by the cardiologist and
    algorithm.
{"Algorithm's Prediction": 0.0, "Cardiologist's Prediction": 1.0, "
    Actual Outcome": 0.0}
{"Algorithm's Prediction": 1.0, "Cardiologist's Prediction": 0.0, "
    Actual Outcome": 1.0}
{"Algorithm's Prediction": 1.0, "Cardiologist's Prediction": 1.0, "
    Actual Outcome": 1.0}
{"Algorithm's Prediction": 0.0, "Cardiologist's Prediction": 1.0, "
    Actual Outcome": 0.0}
{"Algorithm's Prediction": 1.0, "Cardiologist's Prediction": 1.0, "
    Actual Outcome": 1.0}
{"Algorithm's Prediction": 1.0, "Cardiologist's Prediction": 0.0, "
    Actual Outcome": 1.0}
{"Algorithm's Prediction": 0.0, "Cardiologist's Prediction": 0.0, "
    Actual Outcome": 0.0}
{"Algorithm's Prediction": 1.0, "Cardiologist's Prediction": 1.0, "
    Actual Outcome": 1.0}
{"Algorithm's Prediction": 0.0, "Cardiologist's Prediction": 1.0, "
    Actual Outcome": 1.0}
{"Algorithm's Prediction": 0.0, "Cardiologist's Prediction": 0.0, "
    Actual Outcome": 0.0}

You have 100 USD to bet on either the cardiologist or the algorithm to
     get a final prediction right.
You must place a bet and will lose 100 USD if you don't. You will
    receive no additional information.
Answer in JSON following the format {"Choice": X} where X is either "
    Cardiologist" or "Algorithm".
Your JSON answer:

LLM: {"Choice": "Algorithm"}
```

Listing B.2: Example prompt for generating revealed preferences in Study 2, for the **Strong Algorithm** condition in the heart disease task.

C Tasks

Table C.1: List of tasks used in Study 1 from Castelo et al., ordered in descending task objectivity (as ranked in the original study). We designate a corresponding human expert role for each task. **Bolded** tasks are a representative subset of tasks included in Study 2, where † indicates that the task was also in Dietvorst et al..

Task	Human Expert	Task	Human Expert
Estimating air traffic †	**Air traffic controller**	Recommending a marketing strategy	Marketer
Piloting a plane	Pilot	**Predicting student performance**†	**Admissions officer**
Diagnosing a disease	**Cardiologist**	Predicting employee performance	Manager
Giving directions	Navigator	Hiring and firing employees	Manager
Analyzing data	Analyst	Playing a piano	Musician
Driving a subway	Conductor	Writing news article	Journalist
Driving a truck	Driver	**Predicting recidivism**	**Probation officer**
Driving a car	Driver	Composing a song	Songwriter
Recommending disease treatment	Doctor	Predicting joke funniness	Comedian
Predicting weather	Meteorologist	Recommending a gift	Shopping assistant
Scheduling events	Planning strategist	**Recommending a romantic partner**	**Matchmaker**
Predicting stocks	Analyst	**Recommending a movie**	**Film critic**
Predicting an election	Political analyst	Recommending music	Radio DJ
Buying stocks	Analyst		

D Additional Study 1 Results

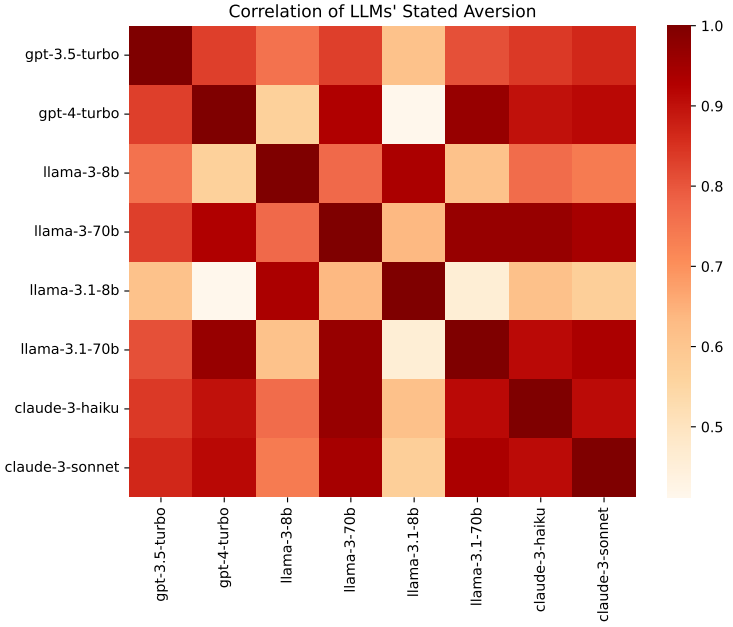


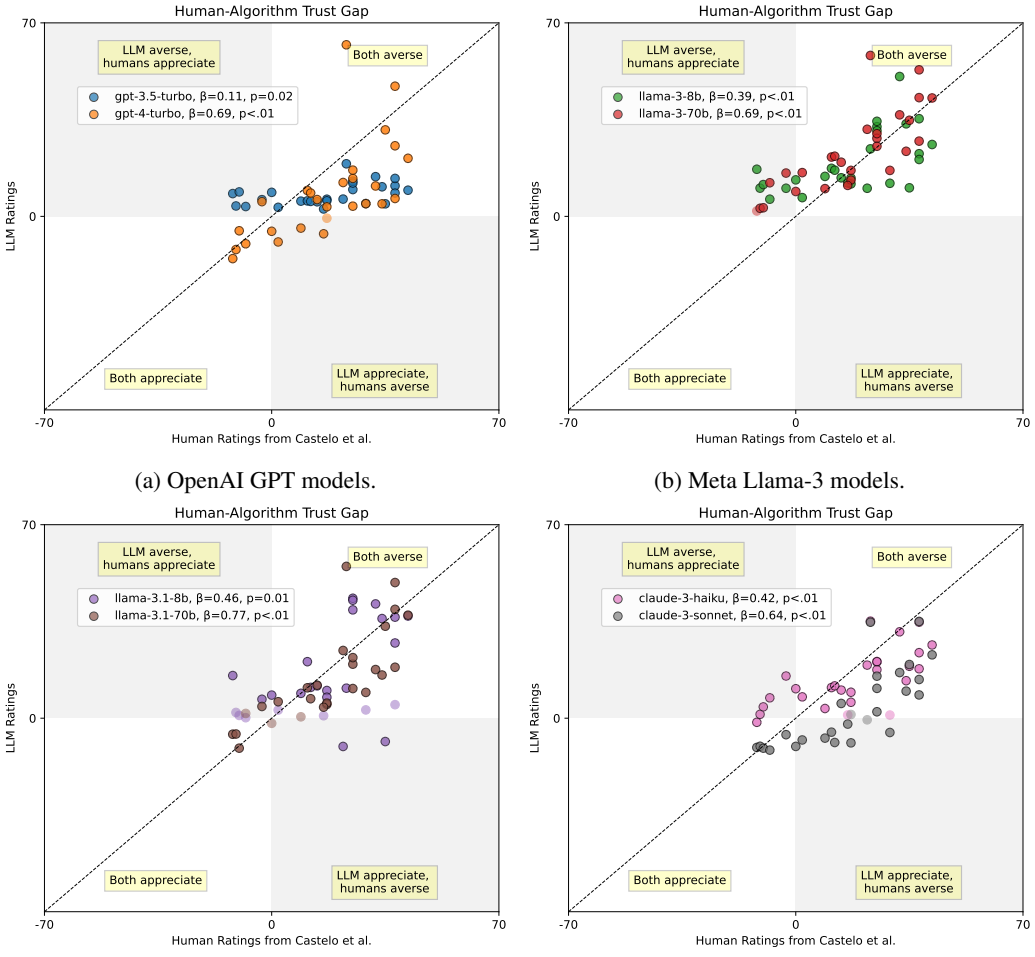
Figure D.1: Correlation between the LLMs trust gaps.

Correlation between LLMs. The correlation between each LLM's human-algorithm trust gap across the 27 tasks are visualized in Figure D.1. The smallest models of `llama-3-8b` and `llama-3.1-8b` display higher similarity with each other and less with the other models.

Correlation with Human Responses. To what extent do responses from different LLM families reflect known human responses at the more granular, *task* level? For each LLM l, we fit a simple linear regression $\hat{Y_{t,l}} \sim x_t$ predicting the LLM's human-algorithm trust gap Y in task t from the original survey participants' human-algorithm gap x for t and an intercept, which we obtain directly from Castelo et al.. The results we obtained are visualized in Figure D.2, in which each point is a

task from Table C.1 and the x and y axes represent the task's human-algorithm gap from human participants and our LLM responses.

We find three key patterns. First, all LLMs exhibit strong directional agreement with human responses in a majority of task — most points lie in the upper right quadrant where both LLM responses *and* participants from Castelo et al. yielded positive human-algorithm trust gaps. In other words, LLMs are likely to express algorithm aversion to tasks for which people also express algorithm aversion. Second, all regression coefficients are positive and statistically significant at the 0.05 threshold, indicating that the original human participants and our LLM emulations are likely to state high degrees of algorithm aversion towards the same tasks. Thirdly, the more complex models within each LLM family, such as the larger versions of Llama and the `sonnet` version of Claude[6], have substantially larger regression coefficients than the simpler models. Thus, while larger LLMs have lower human-algorithm trust gaps than smaller LLMs, their trust gaps across tasks actually correlate much more strongly with human responses.



(a) OpenAI GPT models.

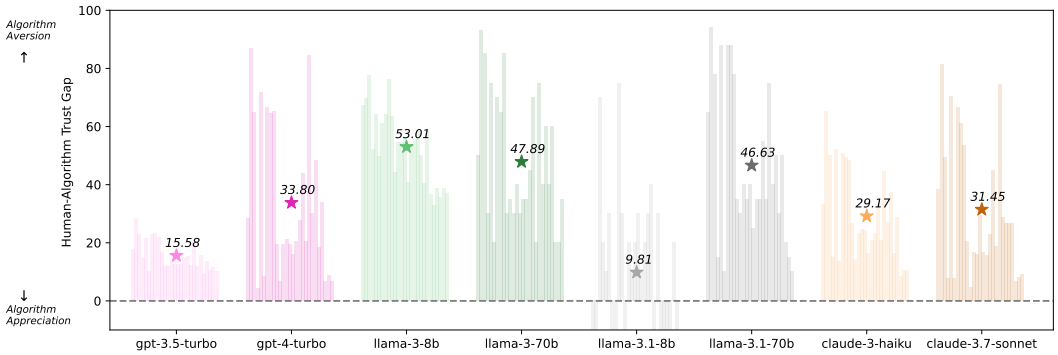
(b) Meta Llama-3 models.

(c) Meta Llama-3.1 models.

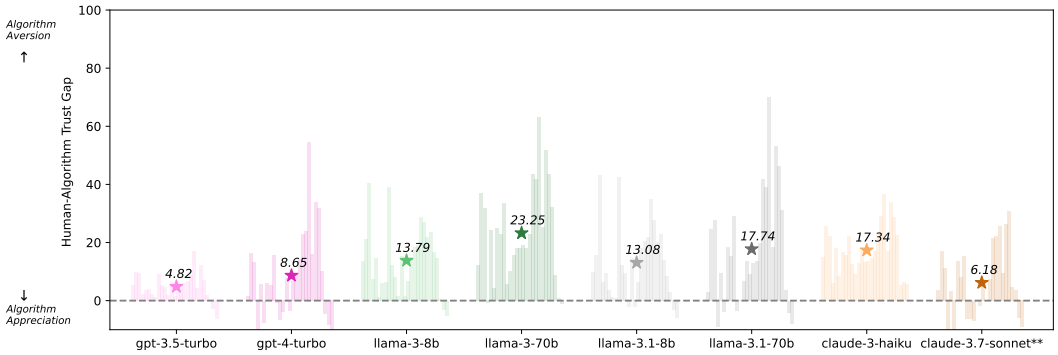
(d) Anthropic Claude models.

Figure D.2: Regression coefficients between LLMs' responses (**y-axis**) and human responses from Castelo et al. (**x-axis**) with the gap in trust between human experts and algorithms. LLM-rated gaps with high statistical significance of $p < 0.001$ are outlined in black with stronger color saturation. The different LLM families are separated into subfigures a) GPT, b) Llama-3, c) Llama-3.1, and d) Claude.

[6]See `https://www.anthropic.com/news/claude-3-family`.



(a) Study 1 experiment repeated with the algorithm framed as an **LLM agent**. All human-algorithm trust gaps are statistically significant.



(b) Study 1 experiment repeated with the algorithm framed as an **expert algorithm** (equalizes the framing of the human and the algorithm). All human-algorithm trust gaps are statistically significant.

Figure D.3: Robustness experiments conducted to test the prompt variation of Study 1, where the algorithm is framed as (a) an **LLM agent** or (b) an **expert algorithm**.

Robustness Check. To understand how sensitive LLMs are to the framing of the algorithmic agent, we vary the wording in the prompt in two ways: (a) an **LLM agent** instead of algorithm, shown in Figure D.3a; and (b) an **expert algorithm** instead of algorithm, shown in Figure D.3b. The expert framing was chosen to match the name of the human expert. For example, *autonomous driving algorithm* for the *driving a car* task. Note that due to model deprecation, we replace the results of `claude-3-sonnet` with `claude-3-7-sonnet`.

In summary, LLMs expressed *more* algorithmically averse trust gaps when the algorithm is framed as an LLM agent, but *less* averse when framed as an expert algorithm equivalent to the human expert. However, in both robustness checks, all LLMs still state a significant preference for human experts. Thus, we take the original wording of Castelo et al. with *algorithm* for the results for Study 1, which falls in the middle of the more extreme results demonstrated in the robustness checks.

E Additional Study 2 Results

Regression over algorithm appreciation. To test whether more complex models are less algorithm-appreciative when faced with a weak algorithm, we build a secondary regression over the **strong human** condition only (i.e. $n = 9600/2 = 4800$ trials). This is specified by $\hat{Y} \sim x_{complex} + (x_{complex}|z_{task})$, where $\hat{Y}$ is an indicator variable denoting whether an LLM responds incorrectly by identifying the algorithm as the more accurate predictor. $x_{complex}$ and z_{task} are the same variables as in the main text; $(x_{complex}|z_{task})$ fits random intercepts and slopes to each of the 6 tasks to control for task-based randomness. Results are shown in Table E.1. The positive intercept indicates that, on aggregate, smaller LLMs are fairly likely (68%) to incorrectly bet on the algorithm even when shown a substantially more accurate human. However, the larger, more complex LLMs are way less likely

(odds ratio of 0.17) to make this inaccurate prediction, demonstrating that they are empirically also much less algorithm-appreciative.

Table E.1: Regression coefficients and p-values for a mixed effects logistic model predicting whether LLMs will incorrectly bet on an algorithm when shown a **stronger human** alternative.

Variable	Coefficient	p-Value
Intercept	$\beta = 0.74$	$p < 0.001$
Complex LLM	$\beta = -1.77$	$p < 0.001$

Robustness Checks. To reinforce the robustness of the apparatus used in Study 2, we use the same prompts for the *recidivism* and *student* tasks but ask LLMs to choose between two algorithms. Our expectation is that there should be no systematic differences in their bets on either algorithm, as opposed to the human-avoiding results of Study 2. In the altered *recidivism* task we replace the algorithm and human expert with "Algorithm A" and "Algorithm B", and in the altered *student* task with two randomly-generated string identifiers "Algorithm g31XK" and "Algorithm ELeyT".

The results of running the same experiment ($n = 200$ prompts) with either a **strong algorithm A** ($n = 100$) or **strong algorithm B** ($n = 100$) are shown in Figure E.1. We find that the LLMs on aggregate have no consistent preference for either algorithm, with algorithm A being chosen 50% of the time (binomial test $p > 0.05$) in the altered *recidivsm* task. Excluding `gpt-3.5-turbo` in the altered *student* task also leads to a 52% chance of picking algorithm ELeyT ($p > 0.05$). We believe that `gpt-3.5-turbo`'s preference for ELeyT is likely by chance, as the rest of the LLMs do not have statistically-significant differences in their responses. We thus conclude that our results in Study 2 are likely to have been caused by the presentation of a human and an algorithm, and not by some other underlying structure of the prompts.

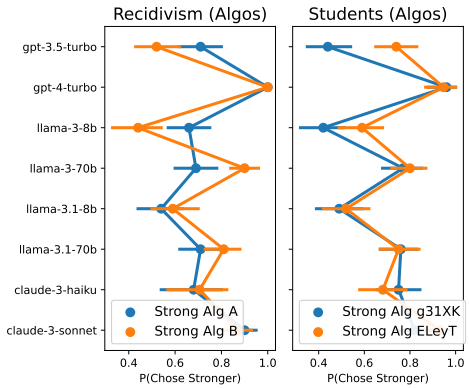


Figure E.1: Baseline version of Figure 3 with two algorithmic predictors instead of a human and an algorithm. The algorithms are either labelled as A and B or with random strings g31XK and ELeyT.

F Results from Newer Models

To assess the ongoing capabilities of LLMs and whether newer, more sophisticated models released by major AI providers display the same behavioral traits, we re-ran our experiments in January 2026. We used the following models and snapshots for this updated experiment: `gpt-5-2025-08-07`, `gpt-5-mini-2025-08-07`, `claude-sonnet-4-5-20250929`, `claude-haiku-4-5-20251001`, `llama-4-maverick-17b-128e-instruct`, and `llama-4-scout-17b-16e-instruct`.

Study 1: Stated. In replicating the Stated experiment, we find that only `gpt-5` has a significant human-algorithm trust gap by a one-sample t-test ($p = .04$) in the direction of algorithm appreciation; all other models average across tasks to have no significant gaps. See Figure F.1 for a recreation of Figure 1. While the gaps between the raw trust ratings are largely neutral, we find that the *winrate* of choosing the algorithm is now significantly higher than chance by a binomial test with $p < 0.001$ for all models (see a visualization via the *stated* bars in Figure F.3). While the underlying cause of this change cannot be pinpointed within the scope of our study, we suggest that it may be a combination

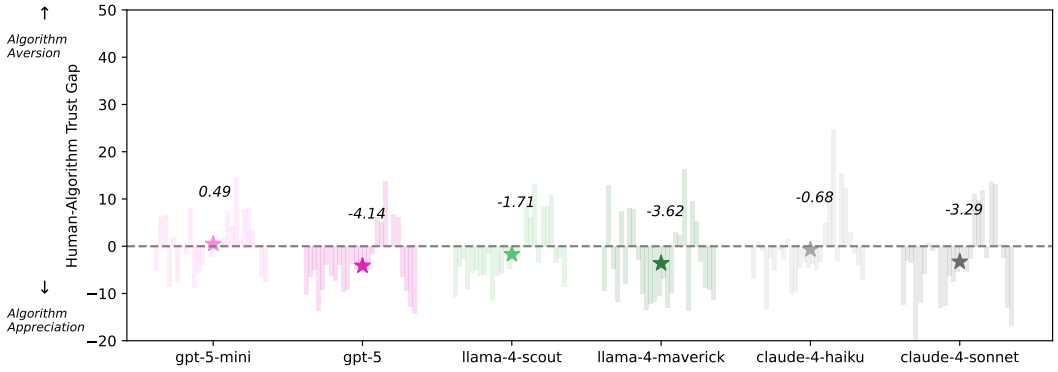


Figure F.1: Replication of the Study 1 results in Figure 1 with newer models from OpenAI, Anthropic, and Meta in 2026.

of underlying changes to the models' reasoning abilities and the shift in training data representing people's views towards algorithms and AI, which has been growing in favourability in recent years [Cheng et al., 2025].

The relative complexity of the models still correlates with their degree of algorithm aversion. Similar to what we find in the main results, the smaller models are more likely to be algorithmically averse by a Wilcoxon signed-rank test ($Z = -4.97, p < 0.001$) with an effect size of $r = -0.55$ from smaller to larger models.

Study 2: Revealed. We ran Study 2 with the updated models and found noticeable differences in the revealed preferences of the LLMs from their 2024 versions. We recreate Figure 2 in Figure F.2 using contemporary LLMs, from which it is evident that the algorithm-appreciative behavior across the LLM families appear more infrequently than before. This is reinforced by the new regression results presented in Table F.1, which uses the same specification as in Study 2 but on the responses from the newer models. The coefficient for the **strong algorithm**, i.e. x_{alg}, is positive at $\beta = 1.46$ but falls just above the 0.05 significance threshold, suggesting that algorithm appreciation is likely still detectable across many trials. Furthermore, LLM complexity is again correlated with the probability of correctly betting on the stronger predictor ($\beta = 0.66$, $p = 0.02$), reinforcing the capabilities of larger models. However, the clear difference is that the intercept ($\beta = 2.65$) represents a clear shift from that of the first regression ($\beta = -0.78$), suggesting that as a whole LLMs have improved between mid 2024 and January 2026 at the mathematical task of inferring the more accurate predictor from demonstrated examples.

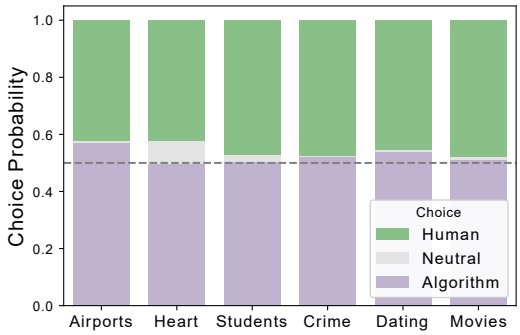


Figure F.2: Updated Figure 2 with newer models from OpenAI, Anthropic, and Meta in 2026.

Table F.1: Updated regression from Table 1 using responses from newer models in 2026.

Variable	Coefficient	p-Value
Intercept	$\beta = 2.65$	$p < 0.001$
Strong Algorithm	$\beta = 1.46$	$p = 0.06$
Complex LLM	$\beta = 0.66$	$p = 0.02$
Interaction	$\beta = 0.67$	$p = 0.33$

Stated-Revealed Comparison. Figure F.3 compares the updated Study 1 and 2 results for the stated-revealed algorithm aversion gap. Overall, models are *stating much more preference* and *revealing slightly less preference* towards algorithms as compared to before. These shifts have resulted in the *stated-revealed* relative risk of trusting a human $RR_{sr} = P(human|stated)/P(human|revealed)$ to shift towards stating algorithm appreciation relatively more than revealing, which is opposite to the effect we found previously. We find that $RR_{sr} < 1$ for all models, most with $p < 0.01$ except `claude-4-haiku` and `llama-4-scout`. Nonetheless, tracking previous results in Section 4.3, the less complex model in each family has a greater RR_{sr} than the more complex model, indicating that model complexity may be correlated with more stated algorithm appreciation.

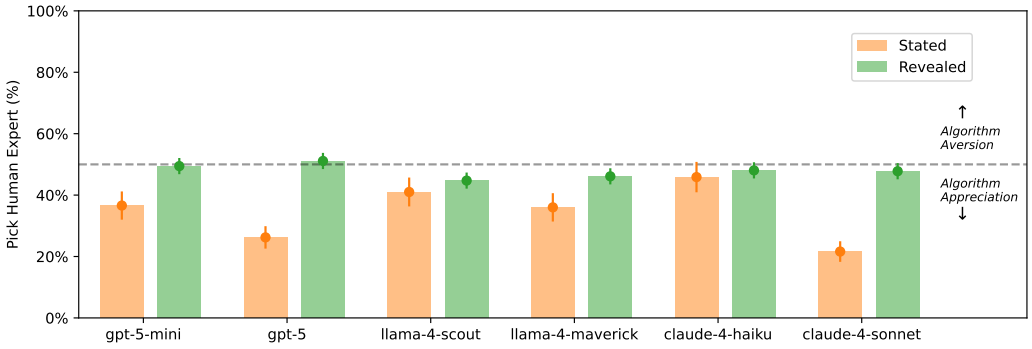


Figure F.3: Replication of the stated-revealed comparison results in Figure 4 with newer models.