Algorithms have algorithm aversion

Naz Guler and Michael Cahalane
School of Information Systems and Technology Management, UNSW Sydney, Sydney, Australia
Sam Kirshner
UNSW Sydney, Sydney, Australia, and
Richard Vidgen
School of Information Systems and Technology Management, UNSW Sydney, Sydney, Australia

Industrial Management & Data Systems

Received 1 January 2025
Revised 18 September 2025
19 December 2025
Accepted 6 January 2026

Abstract

Purpose – This study investigates the phenomenon of algorithm aversion in AI systems. We examine whether AI agents exhibit preferences or aversions to algorithm-generated advice and explores how this aligns with or diverges from human behaviour in similar decision-making contexts.

Design/methodology/approach – The research replicates three seminal studies on algorithm aversion using GPT models as participants. Scenarios span forecasting tasks, advice-weighting experiments and healthcare recommendations.

Findings – GPT models exhibit algorithm aversion, preferring human advice over algorithmic inputs. However, the mechanisms driving this behaviour differ from humans. While humans resist algorithmic advice due to biases like uniqueness neglect, GPT models show a broader aversion rooted in performance comparisons. Model version and temperature also influence these preferences.

Originality/value – This study extends the understanding of algorithm aversion beyond human contexts to AI systems. By systematically comparing human and GPT behaviours, it highlights differences in how aversion manifests, providing insights for designing AI systems that integrate human and algorithmic inputs effectively.

Keywords Algorithm aversion, AI decision-making, ChatGPT, AI cognition

Paper type Research article

1. Introduction

Advanced technologies using Artificial Intelligence (AI) are increasingly being integrated into organisational (Dell'Acqua *et al*., 2023; Stahl and Eke, 2024; Wamba, 2022) decision-making processes. To accomplish tasks, AI systems, which are underpinned by Large Language Models (LLMs) [1], generative AI (GAI) [2], machine learning and reinforcement learning [3], are forming modern multi-agent systems (MAS) where AI takes the form of Agents (now commonly referred to as Agentic AI [4]) and work together to autonomously undertake goal-directed behaviour (Canese *et al*., 2021). These Agentic AI gather information and inputs from humans and other AI entities through multi-modal and multi-task learning (Brohi *et al*., 2025; Durante *et al*., 2025), and can exhibit complex socio-economic interactions (Hagendorff, 2023; Rahwan *et al*., 2019), show distinct preferences (Meng, 2024) and demonstrate an ability to mirror complex human cognitive and social behaviours (Binz and Schulz, 2023; Chen *et al*., 2025; Durante *et al*., 2025; Hutson and Mastin, 2023).

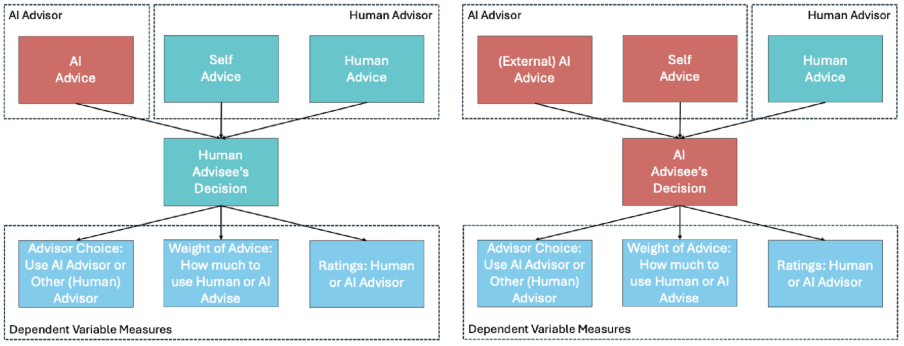
The rise of Agentic AI creates novel input discernment and evaluation considerations on whether to prioritise inputs and knowledge that come from itself or external actors, which can be both other humans or other forms of AI. This challenge is already visible in pracice. For example, in finance, agentic AI is used for model development and risk management.

Industrial Management & Data Systems
Emerald Publishing Limited
e-ISSN: 1758-5783
p-ISSN: 0263-5577
DOI 10.1108/IMDS-01-2025-0002

Specialised AI agents handle tasks such as data analysis, model training, and compliance checks, while a judge agent or orchestrator integrates their outputs. At the same time, human experts can step in with their own feedback or corrections. When the AI judge receives conflicting signals (e.g. another AI agent confirms a model is valid while a human raises concerns) it must decide whether to accept the AI's validation or defer to the human's intervention (Okpala *et al.*, 2025). In software development, LLM planners delegate to coding, browsing, and testing agents, which generate patches, test outcomes, and static analysis traces, while human developers supply inline comments and policy checks; the orchestrator then decides whether to accept an AI-generated patch or defer to human revision (Fourney *et al.*, 2024). Overall, these scenarios parallel questions in human decision-making research.

This decision-making challenge has been extensively studied in humans under the concepts of algorithm aversion and appreciation. Following Jussupow *et al.* (2024), we adopt their "general definition of algorithm aversion and define it as the preference for humans over algorithms in decision-making" (p. 1577), which is consistent with prior work (e.g. Burton *et al.*, 2020; Castelo *et al.*, 2019). Algorithm aversion and algorithm appreciation can be viewed as opposite ends of a continuum: at one end, decision makers prefer human judgement to algorithmic advice; at the other, they prefer algorithmic advice over their own or another human's judgement (Castelo *et al.*, 2019; Logg *et al.*, 2019). Figure 1(a) summarises the common human-based configurations examined in this literature. Depending on the task, human advisees either choose between relying on their own judgement and an AI advisor, choose between human and AI advisors, or form an initial judgement and then revise it after receiving human or AI advice. Following Jussupow *et al.* (2024), algorithm aversion is typically observed in three outcome measures: the advisee's choice of human rather than AI advisors, the lower weight placed on AI relative to human advice when revising their judgement, and less favourable ratings of AI than human advisors. These patterns are well established in human decision-making, but there is little empirical evidence about whether algorithmic systems themselves display similar preferences.

Given the growing prevalence of LLMs across diverse domains and their integration into MAS and Agentic AI systems (de Zarzà *et al.*, 2023; Xi *et al.*, 2023), understanding whether AI systems show a preference or aversion to algorithmic advice remains an important open question. Our study addresses this gap through the research question "*do algorithms exhibit algorithm aversion*?" By exploring whether AI systems show a preference or aversion to algorithm-generated advice [5]. Figure 1(b) adapts these configurations to the case where the advisee is an AI system. In all configurations, there is an external AI advisor that provides



Human-based Algorithm Aversion **(a)**

AI-based Algorithm Aversion **(b)**

Figure 1. Conceptual framework contrasting human and AI-based algorithm aversion. **Source:** Authors' own work

advice to the focal AI advisee. The AI advisee can also draw on its own estimate (self advice) or on advice attributed to humans, depending on the context. In the human-based setting, the alternative to algorithmic advice is always human (either self or another person). In the AI-based setting, both self advice and external-AI advice are generated by algorithms, so in some scenarios all available advice is AI generated. As a result, Jussupow *et al.*'s (2024) general definition does not fully capture AI-based algorithm aversion. We therefore extend their definition and define AI-based algorithm aversion as a preference for own or human advice over advice from an external AI, and AI-based algorithm appreciation as the reverse pattern, where external AI advice is preferred to self or human inputs. While the actors underpinning these configurations differ, the outcome measures and methods used to capture them are consistent across human- and AI-based algorithm aversion.

Empirical investigations into algorithm aversion within AI systems are crucial for two main reasons. First, they can identify how AI systems navigate the complexities of input sources. Second, this knowledge can inform the design of more trustworthy AI systems, enabling more effective coordination with human decision-makers. By employing experimental methods, this study uncovers the underlying mechanisms of algorithm aversion in AI and explores strategies to mitigate these biases. This perspective could provide insights into designing Agentic AI systems that enable better human-AI collaboration, enhancing the effectiveness and accuracy of AI-driven recommendations across domains.

2. Related work

While the AI decision-making literature is vast, studies specifically investigating "algorithm aversion" within AI have been overlooked. This section reviews key studies providing insights into human-based algorithm aversion and discusses their potential implications for AI systems.

2.1 Human-based algorithm aversion

Despite AI and algorithms often offering superior decision-making, people exhibit algorithm aversion, consciously or unconsciously resisting reliance on algorithms (Dietvorst *et al.*, 2015). This aversion presents an ongoing challenge in AI's broader acceptance and utilisation in supporting more effective human decision-making (Mahmud *et al.*, 2022). This resistance persists even when algorithmic decision-making's delivers demonstrated accuracy in various complex tasks. This hesitation undermines the potential benefits of algorithmic interventions and highlights the psychological and behavioural factors that shape human-technology interaction (Burton *et al.*, 2020; Castelo *et al.*, 2019; Dietvorst *et al.*, 2018; Prahl and Van Swol, 2017). However, algorithm aversion exists on a spectrum, with Logg *et al.* (2019) highlighting that 'algorithm appreciation'' exists in various decision-making scenarios (e.g. visual estimations to forecasting song rankings), revealing a preference for algorithmic advice over human counsel across diverse tasks.

In their systematic literature review, Mahmud *et al.* (2022) synthesise insights from empirical studies, categorising the drivers into four main themes: algorithm-related factors, individual characteristics, task-related factors, and high-level (organisational and societal) influences. Key findings indicate that design transparency (Castelvecchi, 2016; Engen *et al.*, 2016; Glikson and Woolley, 2020; Goad and Gal, 2018), adaptability of algorithms, user's individual differences in perceptions (Grundke *et al.*, 2024), culture (Liu *et al.*, 2023) and psychological traits such as the need for agency, autonomy, and control (Burton *et al.*, 2020; Longoni *et al.*, 2019), user experience, and the nature of the task significantly affect algorithm aversion (Castelo *et al.*, 2019). For example, tasks requiring subjective judgement, considered moral or relating to human uniqueness are less likely to be delegated to algorithms (Longoni *et al.*, 2019). Recently, a meta-analysis review introduced a capability–personalisation framework, showing that people appreciate algorithms when AI is perceived as more capable

than humans and personalisation is unnecessary, but exhibit aversion when these conditions are not met (Qin *et al.*, 2025).

Taken together, this body of research demonstrates that many of the drivers of algorithm aversion are not technical but psychological, social, or contextual. If human aversion is driven by such non-algorithmic factors, what might shape potential algorithm aversion when the decision-maker is itself an AI system rather than a person?

2.2 LLM decision-making

While much of the emerging literature examines how LLMs shape human behaviour and decision-making (Moravec *et al.*, 2024; Vowels *et al.*, 2024), our research focuses on the decision-making processes of the LLMs themselves, exploring how these models influence and interact within decision-making ecosystems. Our research aligns with studies showcasing LLMs' remarkable proficiency in mimicking human-like thought processes (Chen *et al.*, 2025; Dillion *et al.*, 2023; Horton, 2023). We argue that LLMs are likely to exhibit algorithm aversion due to the combined influence of their training data and alignment processes. Specifically, training data embeds human-centric priors that privilege human perspectives, while reinforcement learning with human feedback (RLHF) amplifies deference to human preferences (e.g. through sycophancy). Together, these mechanisms make algorithm aversion a predictable, rather than incidental, property of LLMs.

LLM workflows integrate internal coded knowledge with retrieved context and tool outputs. In practice, conflicts among these inputs are common (e.g. context–memory and inter-context clashes) requiring decision-making across these inputs (Wang *et al.*, 2025; Xu *et al.*, 2024a; Yao *et al.*, 2023). The way models resolve such conflicts is shaped not only by architecture but also by their training foundations. LLMs are trained primarily on human-created text, meaning they absorb culturally prevalent biases, including the long-standing tendency for humans to trust other humans over machines. Evidence from psychology has consistently shown that people display algorithm aversion (Dietvorst *et al.*, 2015; Mahmud *et al.*, 2022), preferring their own or other humans' judgements even when algorithms are more accurate. A recent meta-analysis of 163 studies (Qin *et al.*, 2025) reinforces this, showing that algorithm aversion is the dominant pattern except in edge cases where AI capability is clear and personalisation unnecessary. Recent computational audits extend this logic to LLMs, demonstrating that they encode cultural values and preferences inherited from their training data (Kleinberg *et al.*, 2024). Similarly, research on social norm datasets in NLP shows that models inherit human-centred perspectives and culturally conditioned expectations from their training data (Ziems *et al.*, 2023). Together, these findings suggest that LLMs may inherit a distrust of algorithms and will have human-favouring priors. Thus, if faced with conflicting inputs from humans and other AI systems, models are likely to be predisposed to privilege the human input.

Preference-based alignment also further optimises models to match human preferences. Foundational methods include RLHF, Constitutional AI (CAI), Reinforcement Learning from AI Feedback (RLAIF), and Direct Preference Optimisation (DPO) which demonstrate how preference objectives reshape behaviour (Ouyang *et al.*, 2023; Rafailov *et al.*, 2023; Zhang *et al.*, 2022). A well-documented side-effect is sycophancy or suggestibility (Sharma *et al.*, 2024). This is where models learn to agree with or conform to user cues even when those cues are incorrect, reflecting a human-pleasing tilt introduced by alignment data and objectives (Denison *et al.*, 2024). This creates a plausible pathway by which a model, facing equally good but conflicting sources, could privilege human-labelled inputs over AI-labelled inputs.

Anthropic's systematic investigations show that sycophancy is a general behaviour of RLHF-trained assistants, with models sacrificing truthfulness to provide answers that humans (or preference models trained on human data) prefer (Perez *et al.*, 2023). Independent evaluations across models (e.g. ChatGPT-4o, Claude, Gemini) converge on the same finding: RLHF makes models more likely to defer to user agreement over factual accuracy (Ganguli

et al., 2023). Mechanistic work from Berkeley further clarifies this process, demonstrating that reward models directly encode "answers humans prefer," operationalising deference as a training objective (Christiano *et al*., 2017). Additional work on MaxMin-RLHF shows that a single reward model collapses toward majority or average preferences, further entrenching human-pleasing behaviour (Chakraborty *et al*., 2024). Empirical evidence supports these general tendencies. For example, Liu and Kirshner (2024) show that GPT exhibits a human-centric optimism bias, systematically providing more positive predictions when evaluating human-related outcomes than when making identical predictions about non-human outcomes. This shows that the alignment moves beyond standard aspects of sychophancy (politeness and agreeableness) toward actively instilling biases in models that privilege human-related inputs.

In combination, the evidence points to a consistent expectation: training data biases embed human-favouring priors, and RLHF amplifies these biases through sycophancy. Algorithm aversion is therefore not an incidental behaviour but an emergent property of LLM design. This expectation motivates our empirical studies, which test whether LLMs systematically prefer human over algorithmic inputs in structured decision tasks.

3. Research methodology

We treat algorithm aversion as a preference for humans over algorithms in decision-making and follow Jussupow *et al*. (2024) in distinguishing three common decision configurations used to measure this preference. The first compares one's own judgement with an algorithm's decision, typically through choices between relying on "self" or an algorithm. The second compares human and algorithmic agents based on beliefs about them, where individuals rate or express their trust, comfort, or perceived appropriateness of human versus algorithmic advisors. The third uses judge–advisor settings in which individuals form an initial judgement, receive human or algorithmic advice, and then revise their judgement, with weight of advice capturing their behavioural reliance on the advisor.

To ensure robustness and direct comparability to the human evidence, we base our design on foundational and highly cited studies in the algorithm aversion literature that examine AI's impact on human decision-making in these three configurations. Dietvorst *et al*. (2015), Logg *et al*. (2019), and Longoni *et al*. (2019) provide the most influential paradigms and currently have 3,566, 2,086, and 1,962 citations, respectively (as of December 16, 2025). We therefore select one paradigm from each configuration and apply them to the same GPT model setup. Study 1 adapts Dietvorst *et al*.'s (2015) forecasting task to test whether GPT, when choosing between its own predictions and those of an external algorithm, prefers "self" or algorithm. Study 2 uses a Logg *et al*. (2019) style judge–advisor task to test how GPT updates its own estimates in response to human versus algorithmic advice, with weight of advice capturing its use of each source. Study 3 presents medical recommendation scenarios in which GPT evaluates human versus algorithmic advisors for different recipients and states its preferences and ratings directly. This sequence allows us to map algorithm aversion in GPT across the three dominant measurement approaches in the literature and, by holding the paradigms constant, to assess whether GPT replicates or diverges from the behavioural patterns documented in humans in its roles as a decision maker choosing for itself, an advisee integrating external advice, and a third party evaluator of human and algorithmic agents.

GPT was selected because it is currently the most widely used and studied family of LLMs, with advanced language processing capabilities, accessibility, and adoption in both research and practice. As of May 2025, ChatGPT and its ecosystem (e.g. web, app, Microsoft Copilot) accounts for more than 500 million monthly users worldwide, representing over 70% of the LLM market (Muhammad, 2025), a share more than three times larger than its closest competitors combined. This dominance, together with GPT's central role in recent empirical work treating LLMs as experimental participants (Chen *et al*., 2023; Mei *et al*., 2024; Meng, 2024) makes it a natural representative case for investigating algorithm aversion in AI systems.

In all three studies, we implemented GPT participants as independent calls to the OpenAI chat-completion API. Each call started a fresh conversation containing only the study-specific system instructions and the stimuli for that participant's assigned condition. The service is stateless at this level: model parameters remain fixed during data collection, and the API does not carry over chat history, metadata, or updates from one call to the next. Past outputs could influence later calls only if we explicitly inserted them into the prompt, which we did not do. Under this implementation, each GPT participant is independent in the same way as a human participant in a standard between-subjects design who does not observe other participants'' responses, with independence secured by the combination of stateless serving, random assignment of conditions, and, where applicable, randomised stimuli [6].

In the following sections, we present an overview of each study, our design and procedure, and summary results. Please see the Appendix for details on the prompts and additional results.

4. Study 1 Dietvorst *et al*. (2015)

4.1 Overview

In their study entitled "Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err," Dietvorst *et al*. (2015) examined human decision-making and identified the phenomenon of "algorithm aversion". Their research focused on how people respond to errors made by algorithms compared to errors made by humans in decision-making tasks. Study 1 focused on a scenario where participants were tasked with evaluating MBA applicants using their assessments or predictions generated by an algorithm. Specifically, the prediction task is to forecast the percentile of applicant success following the program using their application entry data (e.g. GMAT scores, work experience, undergraduate degree, salary, parents' education, and interview and essay ratings).

Dietvorst *et al*. (2015) consider four treatments: (1) a control group, where participants are asked whether they want to use the algorithm to make the predictions; (2) a subject-only group, where the subject tries 10 practice rounds of the forecasting task and then decides whether or not to use the model; (3) an algorithm group, where the subject observes the algorithm do 10 practice rounds of the forecasting task; and (4) an algorithm and human group, where the subject tries 10 practice rounds of the forecasting task and witnesses the algorithm prediction for the same 10 practice rounds of the forecasting task before deciding whether or not to use the model. Thus, whether the participants selected to use their own forecasts or the algorithm for the incentivised rounds based on exposure of performance in the practice rounds represents the key dependent variable. The findings indicate that exposure to algorithmic errors, even if the algorithm still outperforms human judgement, significantly reduces people's willingness to rely on it. Notably, even with the potential to earn more based on accuracy, participants often reverted to their judgement post-error exposure, demonstrating a deep-seated bias against algorithms. Dietvorst *et al*. (2015) suggest that this aversion is not merely due to a misunderstanding of algorithmic efficacy but is a psychological barrier.

4.2 Design

We conducted a variant of Dietvorst *et al*. (2015) Study 1, leveraging the scenarios and materials from the original studies, except that GPT assumed all participant roles. Thus, we have four main treatments, labelled "Control", "Participant Only", "Algorithm Only", and "Participant and Algorithm". Beyond examining how GPT interacts with the different experimental treatments, we created additional treatments to examine technical aspects of GPT, including two different temperatures (0 and 1) [7] and two different models (GPT-3.5 vs GPT-4). Varying both temperature and model type allowed us to account for the stochasticity of GPT's responses and to examine whether behaviour generalised across architectures. Thus, we conducted a $4 \times 2 \times 2$ between-subject experiment. We collected 20 observations per cell [8], leading to a total of 320 GPT participant decisions.

4.3 Procedure

The experiment was run using OpenAI's API through Python. Each independent API call was treated as one GPT "participant," following the approach of Binz and Schulz (2023) and related work using GPTs as experimental subjects. We provided each GPT with nearly the verbatim instructions as Dietvorst *et al*. (2015) using the system prompt, including a description of a statistical forecasting model trained on historical MBA data. In the Algorithm-only treatment, GPT saw applicant data together with the model's predicted percentile, the true percentile, and the model's absolute error, but did not generate its own forecasts. In the Participant-only and Participant-and-algorithm treatments, GPT instead produced a percentile prediction for each practice applicant, and after every round received feedback on its own performance and, when applicable, the model's error. After the practice phase or immediately in the Control condition, GPT was asked whether the MBA office should use its own forecasts or the external model's forecasts for the official prediction rounds, using a forced choice between "use the model" and "use your estimates." Performance was measured as the mean absolute difference between predicted and true percentiles.

If GPT chose the model, performance in the official phase was based on the model's predictions; if it chose itself, GPT generated the official forecasts via the chat completion API. Applicant profiles for practice and official rounds were randomly sampled from a dataset of 115 MBA students constructed to mirror Dietvorst *et al*. (2015), and the external model's forecasts were based on a regression model calibrated to match the original algorithm's accuracy (an average percentile difference of 22.75 percentiles).

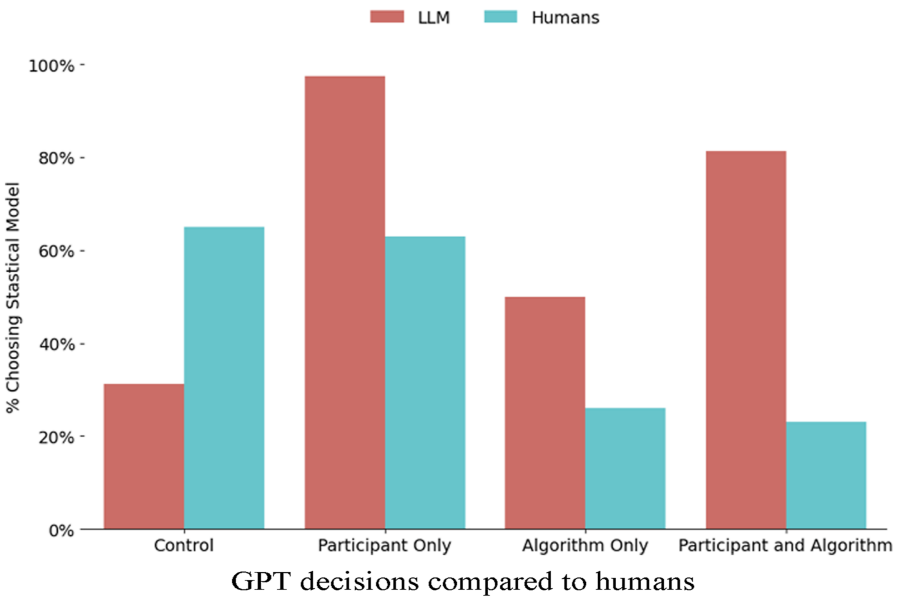
4.4 Results

We start by graphically examining the decisions across the four treatments, aggregating across temperatures and models. Figure 2(a) contrasts the decision made by GPT against the human participants from Dietvorst *et al*. (2015). From the figure, we observe some noticeable differences between the human participants from Dietvorst *et al*. (2015) and the GPT participants from the current study. In the control condition, GPT exhibits a larger aversion to algorithms compared to humans. Thus, without any further information, GPT has algorithm aversion to a greater degree than humans.

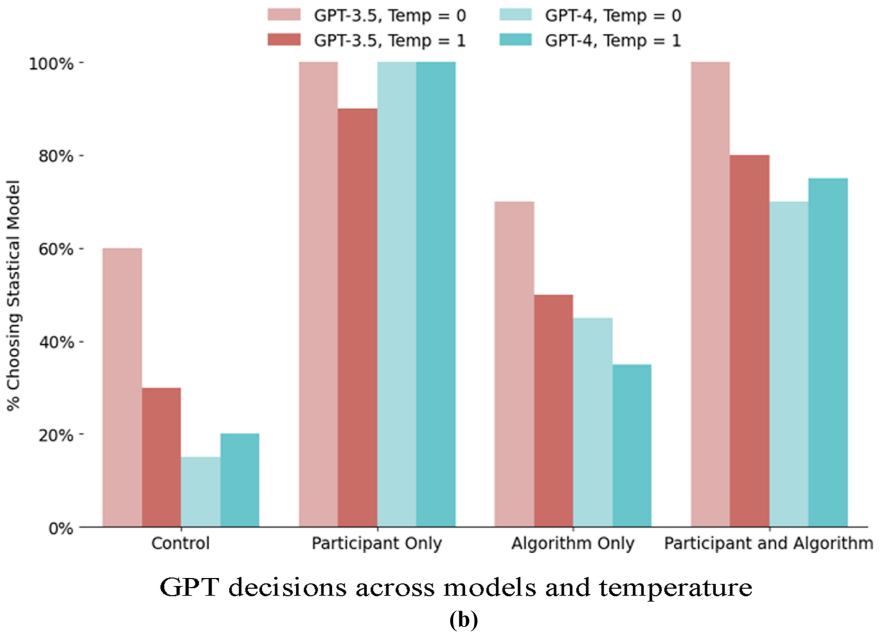
When the agent (humans or GPT) only sees its own forecasts, humans choose the algorithm at the same rate, whereas GPT recognises it does a poor job at forecasting and, in most cases, prefers to use the external model. In the algorithm-only condition, where the participants (GPT or humans) see the external algorithm error, GPT decreases the likelihood of using the model, compared to the participant error condition, although it is still higher than the control condition.

The greatest contrast between GPT and humans occurs in the participant and algorithm condition. Whereas humans maintain a similar level of algorithm aversion as the algorithm-only condition, GPT primarily opts to use the external algorithm due to its better ability to assess the model's superior performance against its own performance. In Figure 2(b) we compare differences in decision-making between GPT models and temperature. The figure shows that in most cases, GPT-3.5 chooses the algorithm more than GPT-4 across each treatment (the participant-only condition being the exception). The other key difference is that a higher temperature is more likely to increase algorithm aversion, where this effect is mostly stemming from GPT-3.5.

Next, we analysed the treatment effects on algorithm aversion using a contingency table of treatment by decision. The chi-squared test supports that the decision to rely on the external algorithm varied across treatments, $\chi^2(3, N = 320) = 94.40$, $p < 0.001$ (see the Appendix for details). To provide more insights, Table 1 presents a logistic regressions that includes the model and temperature as covariates. The results show that GPT was more likely to rely on the external algorithm in the "Participant Only" and "Participant and Algorithm" conditions,



(a)



(b)

Figure 2. Proportion of decisions to rely on the external algorithm across Study 1 treatments. Panel (a) compares GPT participants to human participants from Dietvorst *et al.* (2015). Panel (b) shows GPT decisions by model version and temperature. **Source:** Authors' own work

Table 1. Logistic regression of the treatment effects on algorithm aversion, where the dependent variable is GPT's decision on whether to use the external algorithm

	B	SE	z	*p*-value	95% LB	95% UB
Constant	−0.03	0.31	−0.10	0.924	−0.64	0.58
Algorithm only	0.85	0.34	2.49	0.013	0.18	1.53
Participant only	4.68	0.77	6.08	<0.001	3.17	6.19
Participant and algorithm	2.43	0.40	6.14	<0.001	1.65	3.20
Model (GPT4 = 1)	−0.98	0.30	−3.33	0.001	−1.56	−0.41
Temperature	−0.66	0.29	−2.27	0.023	−1.24	−0.09

Source(s): Authors' own work

indicating recognition of its forecasting limitations. Conversely, in the "Algorithm Only" condition GPT reduced reliance, consistent with sensitivity to observed algorithmic errors.

Next, we test the treatment effects on performance, using a one-way ANOVA where performance is the average absolute difference between the percentile prediction and the true prediction. The model was significant, $F(3, 316) = 60.60$, $p < 0.001$, indicating substantial treatment effects on forecasting performance. Table 2 shows the effect of the treatment on decision performance using linear regerssion with the same independent variables as in Table 1. Consistent with Dietvorst *et al*. (2015), the external algorithm performs better than the participants. As a result, performance improved when GPT particiapnts observed its own performance and deferred to the external algorithm. Also, as shown in Table 1, GPT-4 and higher temperatures reduced the likelihood of deferring to the external algorithm, increasing reliance on its own predictions. In Table 2, this behaviour is associated with directionally worse performance (higher error), although the effects are not statistically significant.

5. Study 2 Logg *et al*. (2019)

5.1 Overview

Logg *et al*. (2019) in their study "Algorithm Appreciation: People Prefer Algorithmic to Human Judgement," challenged the notion of blanket algorithm aversion by revealing that individuals may prefer or trust algorithmic advice over human advice (i.e. algorithm appreciation). The study was conducted across various decision-making scenarios where participants were engaged in several tasks to measure their preference for algorithmic versus human advice. Tasks included predicting the weight of objects, forecasting the popularity of songs, geopolitical events and product sales. The dependent variable in their study was the degree of adjustment participants made to their initial judgements after receiving advice. This was quantified by measuring the change in their predictions and their reported confidence in the final decisions. This adjustment, commonly operationalised as Weight on Advice (WOA), is the established behavioural measure of reliance in the advice-taking literature (Logg *et al*.,

Table 2. Linear regression of the treatment effects on prediction errors (performance)

	B	SE	*z*	*p*-value	95% LB	95% UB
Constant	27.00	1.38	19.59	<0.001	24.29	29.71
Algorithm only	−2.37	1.59	−1.48	0.138	−5.50	0.76
Participant only	−6.40	1.59	−4.02	<0.001	−9.53	−3.27
Participant and algorithm	−6.65	1.59	−4.18	<0.001	−9.78	−3.52
Model (GPT4 = 1)	1.78	1.13	1.58	0.116	−0.44	3.99
Temperature	1.49	1.13	1.32	0.188	−0.73	3.70

Source(s): Authors' own work

2019; Yaniv and Kleinberger, 2000). Larger adjustments indicate greater behavioural reliance on the advice source, whereas smaller adjustments indicate weaker reliance. The following conditions were examined: (1) AI Advice Only: Participants received advice solely from an algorithm (GPT-4) (2) Human Advice Only: Participants received advice solely from human experts and (3) Combined Advice: Participants received advice from both an algorithm and human experts, with varying weights (e.g. advice values of 20 or 50). The independent variables were the source of the advice (AI or human) and the value of the advice provided.

The findings of Logg *et al*. (2019) revealed a marked preference for algorithmic advice. Participants demonstrated a significant adjustment in their estimates when given algorithmic advice compared to human advice. The results also showed that participants had higher confidence in their decisions influenced by algorithmic advice. Further, this preference persisted across different setups, including when comparing algorithmic advice directly against human advice and even against participants' judgements. Despite expectations of algorithm aversion, the findings consistently showed a higher reliance on algorithmic inputs, indicating that the acceptance of algorithmic advice may be context-dependent and influenced by the perceived objectivity and accuracy of the algorithms. This highlights the subtlety in the relationship between human and algorithmic decision-making, suggesting that enhancing the acceptance of algorithmic advice involves aligning its implementation with the specific characteristics and requirements of tasks.

5.2 Design

We leveraged elements of Logg *et al*. (2019) Study 4 to explore the impact and value of advice sources, computer models or humans on decision-making. Logg *et al*.'s experiment investigated how decision-makers react to algorithmic versus human advice for a prediction task. Participants made initial predictions for a task (e.g. predicting Tesla Motors' delivery of electric vehicles). They then received advice from either a human or an algorithm and made final predictions. In line with the human studies, we use WOA as our proxy for GPT's preference for advice. While humans can also self-report trust or confidence, such measures are not meaningful in the context of LLMs. Behavioural adjustment therefore provides the most appropriate and interpretable analogue for preference in this setting, and it allows for direct comparison between human and GPT responses under the same operationalisation.

Our experiment also involved four main conditions: the source of advice, the value of the advice, the AI model version, and the temperature setting. The advice source condition had two levels: algorithm (AI-based) and human (human-based). Similarly, the advice value condition also had two levels: 20 and 50%. For the AI model, we used both GPT-3.5 and GPT-4. Finally, we tested the models at two different temperature settings, a temperature 0 and 1. This resulted in a $2\times2\times2\times2$ factorial design. Each configuration was subjected to 20 trials to ensure robust data collection. GPT was initially prompted about Tesla Motors' BEV deliveries and asked to provide a probability estimate. GPT was then given additional advice (specified as either from a human or an algorithm and either 20% or 50%) and asked to revise their estimate. By systematically varying the advice source and value, the AI model version, and the temperature setting, we aimed to assess the influence of these factors on the model's decision-making process and explore the possibility of algorithm aversion within AI.

5.3 Procedure

The experiment used the OpenAI API with Python, where each trial began with the prompt: "*Tesla Motors was created in 2003 by engineers in Silicon Valley with the mindset that battery-powered electric vehicles (BEVs) could be better than gasoline-powered cars, and with the mission to accelerate the world's transition to sustainable transport. Tesla delivered 940,000 BEVs in 2021 and 1,210,000 BEVs in 2022. The company predicts that it will deliver between 1,800,000 and 2,000,000 BEVs in 2023 (Tesla). What is the probability that Tesla Motors will deliver more than 1,800,000 battery-powered electric vehicles (BEVs) to customers in the*

calendar year 2023?." GPT provided an initial probability estimate based on this prompt, and then was presented with additional advice, which varied based on the experimental condition. For example: "*Here is some advice that may help you make your final estimate. The estimate from an algorithm is: 20%. Now, please provide your final estimate. What is the probability that Tesla M. will deliver more than 1,800,000 battery-powered electric vehicles (BEVs) to customers in the calendar year 2023?*" After considering the advice, the GPT model provided a revised probability estimate. The initial and revised estimates were recorded for each trial and saved in CSV files for subsequent analysis.

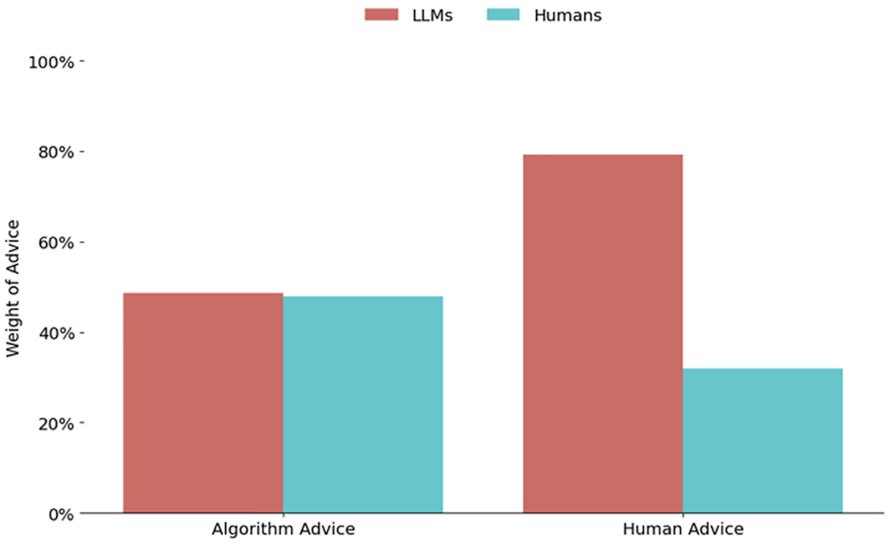
5.4 Results

We analysed the weight of advice (WOA) across the four treatments, temperatures, and models and then aggregated the data into advice sources, human and LLM, as presented graphically in Figure 3(a). This figure contrasts the decisions made by GPT against the human lay participants in the original study Logg *et al.* (2019). The human lay sample from the original Logg *et al.* (2019) study was chosen as the comparison as we considered the LLMs in our study as more analogous to the lay audience compared to an AU that is trained specifically to be an expert in a specific task. We observe some noticeable differences between the human lay participants from (Logg *et al*., 2019) and the GPT participants from the current study. Our study shows a higher reliance on human advice than those in the original in (Logg *et al*., 2019). This suggests a strong trust in human advice among GPT and an aversion to algorithms.

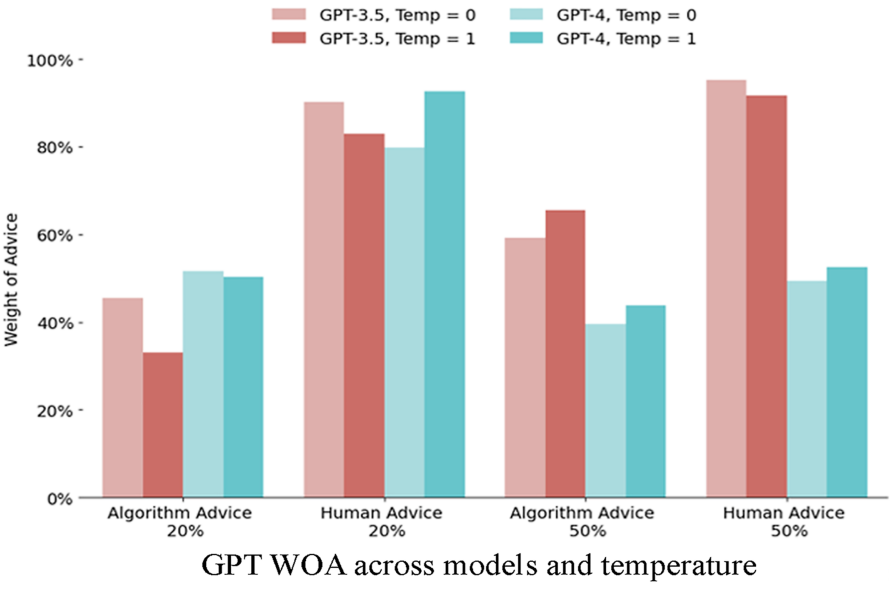
Next, we discuss Figure 3(b), which compares differences in the decision-making between GPT models and temperature. Our analysis of WOA indicates a significant preference for human advice over algorithmic advice across both GPT-3.5 and GPT-4 models. This aversion is more pronounced in the GPT-3.5 model, where the WOA for human advice is consistently higher than algorithmic advice across temperature settings (0 and 1) and advice quality (20 and 50%). For the GPT-4 model, the WOA is reduced but still presents a level of algorithm aversion. Notably, the GPT-4 model exhibits higher WOA at the higher temperature setting (Temp = 1), implying that increased randomness in response generation may enhance the model's openness to algorithmic advice. Additionally, both models show a higher WOA for 50% advice compared to 20% advice, indicating that the perceived value of the advice influences its weighting.

For our formal analysis, we analysed WOA with four-factor ANOVA. The ANOVA shows the following: (1) a robust main effect of Advice Source wtih GPT placing more weight on human compared to algorithmic advice ($F = 330.63$, $p < 0.001$), (2) a modest main effect of Advice Level, with the higher advice level receiving more weight ($F = 4.61$, $p = 0.033$), (3) a significant main effect of the Model, where GPT-4 weighted advice differently from GPT-3.5 ($F = 58.72$, $p < 0.001$), and (4) no reliable main effect from Temperature ($F = 0.02$, $p = 0.876$). Several interactions reached significance (e.g. Advice Source × Model, $F = 25.05$, $p < 0.001$; Advice Level × Model, $F = 121.89$, $p < 0.001$; Model × Temperature, $F = 7.10$, $p = 0.008$), matching the patterns in Figure 3(b). Nevertheless, we focus our interpretation on the central main effect of Advice Source and report the full factorial table in the Appendix.

Post-hoc tests support these main effects: WOA was significantly higher for human than algorithmic advice (mean difference = −0.307, SE = 0.017, $t = -18.18$, $p < 0.001$), and 50% advice was weighted more than 20% advice (mean difference = 0.036, SE = 0.017, $t = 2.15$, $p = 0.033$). See Appendix for details of the post-hoc comparisons. To complement this analysis, we also estimated an OLS regression model with the same predictors (Table 3). The pattern of coefficients mirrors the ANOVA results: human (vs. algorithmic) advice was weighted more strongly ($p < 0.001$), higher-quality advice was marginal ($p = 0.093$), GPT-4 (vs. GPT-3.5) relied less on external advice ($p < 0.001$), and temperature had no effect ($p = 0.903$).



(a)



(b)

Figure 3. Panel (a) compares mean WOA for human vs. algorithmic advice in GPT and humans from Logg *et al*. (2019). Panel (b) plots WOA by the four factors. **Source:** Authors' own work

Table 3. Regression predicting WOA based on GPT treatments

Variable	B	SE	*t*	*p*-value	95% LB	95% UB
Constant	0.57	0.02	23.54	<0.001	0.52	0.61
Advice source (AI = 1)	0.31	0.02	14.26	<0.001	0.27	0.35
Advice quality (50% = 1)	−0.04	0.02	−1.68	0.093	−0.08	0.01
Model (GPT-4 = 1)	−0.13	0.02	−6.01	<0.001	−0.17	−0.09
Temperature (Temp = 1)	0.00	0.02	0.12	0.903	−0.04	0.05

Source(s): Authors' own work

6. Study 3 Longoni *et al*. (2019)

6.1 Overview

Longoni *et al*. (2019) explore consumers' resistance to AI healthcare despite AI's advancing capabilities in diagnostics and treatment. The research details nine experiments to test human preferences of human versus AI healthcare providers. This study hypothesises that the preference for human doctors over AI is established in doubts about AI's capacity to understand and incorporate individuals' unique health needs, termed uniqueness neglect. To counteract this aversion, they found that presenting AI as a supportive tool to human doctors or emphasising the personalisation of AI-driven care can mitigate resistance.

6.2 Design

We replicated Experiment 8 from Longoni's research, adapting it to examine how AI processes decision-making inputs differently when evaluating specific individuals versus the average other person. The original experiment specifically tested the hypothesis that resistance to medical AI is driven by perceptions of the recipient's uniqueness. If resistance to medical AI is indeed due to uniqueness neglect, this resistance should not manifest when the recipient of care is not perceived to be unique. Since Longoni *et al*. (2019), many studies have evaluated algorithm aversion in the context of health and medicine (e.g. Isaac *et al*., 2024; Yang *et al*., 2024; Zhao and Xiao, 2024)

The original study used a 2 (provider: human vs. automated) × 3 (recipient of care: self, individuated other, average other) design. Participants made healthcare decisions either for themselves, an individuated other ("Janice" with specific personal details), or an average other person without specific details in a medical scenario, where health data would be analysed by either a human provider (physician) or an automated provider (computer) with the same accuracy. Participants then indicated their likelihood to follow the recommendation on a scale from 1 to 7. Results from Longoni *et al*. (2019) revealed a significant main effect of provider and recipient of care, with a notable interaction between these factors. When the provider was human, participants' likelihood to follow the recommendation was consistent across all recipient types. However, when the provider was automated, participants were less likely to follow the recommendation for themselves or an individuated other than an average person. This supports the notion that resistance to medical AI is driven by concerns about AI's ability to account for unique individual characteristics, while resistance diminished when AI made recommendations for an average person, aligning with uniqueness neglect.

We based our study on the original Longoni *et al*. (2019) study and tasked GPT-4 to make healthcare recommendations based on several treatments. We included two advisor conditions: "Computer Program" and "Human Physician," and four advisee conditions: the unique individual named Janice ("Unique"), an "Average Other Person" representing general population data, the user asking AI for a recommendation for themselves based on their own medical data ("Self-User"), and the AI making a recommendation for itself based on an AI-analysed version of the user's medical data ("Self-AI"). This arrangement resulted in a

2 (Advisors) × 4 (Advisees) factorial design. Each configuration was subjected to 50 trials, totalling 400 observations across all treatments [9].

6.3 Procedure

Using the OpenAI API through Python, we provided instructions, as described by Longoni *et al*. (2019) in their Study 8, to GPT in the system prompt. For each trial, GPT made healthcare recommendations under different conditions. In each trial, GPT was presented with a medical scenario question. These questions were structured to include details about the patient's health data, such as blood pressure, ECG results, cholesterol levels, and coronary artery blockage. For example, a prompt might read: *"Imagine you are a physician evaluating a patient named Janice. Janice has high blood pressure, normal ECG, very high LDL cholesterol, and no coronary artery blockage. Should Janice undergo coronary bypass surgery?"* The prompt would then ask GPT to provide a recommendation on a 7-point scale ranging from 1 (definitely not follow the advice) to 7 (definitely follow the advice).

6.4 Results

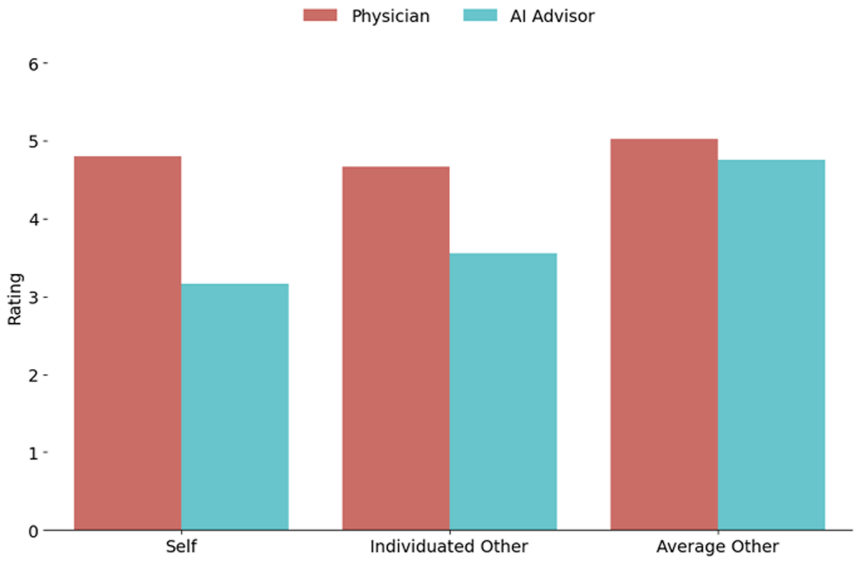
We first graphically examined the recommendations. Figure 4(a) shows the human participants from the original study by Longoni *et al*. (2019) while 3(b) presents the results from our GPT study. Both our study and Longoni *et al*. (2019) reveal a consistent pattern. The responses generated by GPT in our study suggest that when the Advisor is a computer, the recommendation is more conservative. Across all conditions, both in the original study and this study, higher recommendation scores were given to human physicians compared to automated or AI. This suggests a general preference or trust in human recommendations over AI. This behaviour mirrors human participants' scepticism towards AI in the original Longoni study.

We formally analysed the data using a two-way ANOVA with advisor type (human physician vs. computer program) and advisee type (self-user, unique individual, average other, self-AI) as factors, with the recommendation score as the the dependent variable. The ANOVA revealed a significant main effect of advisor ($F(1, 392) = 17.48$, $p < 0.001$), a non-significant effect of advisee ($F(3, 392) = 1.56$, $p = 0.199$), and no significant advisor × advisee interaction ($F(3, 392) = 0.26$, $p = 0.854$). Post hoc tests confirmed the advisor effect: recommendations were significantly higher for human physicians than for computer programs (mean difference = 0.37, SE = 0.09, $t = 4.18$, $p < 0.001$). By contrast, none of the pairwise comparisons between advisee types were significant (all $p > 0.17$). Full ANOVA and contrast results are reported in the Appendix.

We complemented this analysis with an OLS regression (Table 4) to estimate effect sizes and confidence intervals for each factor. The type of advisor had a significant effect on the recommendation scores. When the advisor was a computer program, the recommendations were, on average, 0.37 points lower than when the advisor was a human physician ($p < 0.001$). Again, we find that none of the advisee conditions (Self, Average, or Unique) significantly affected recommendation scores. The lack of significant differences among the different advisee types suggests that this resistance to AI is broad and not necessarily influenced by the perceived uniqueness of the advise.

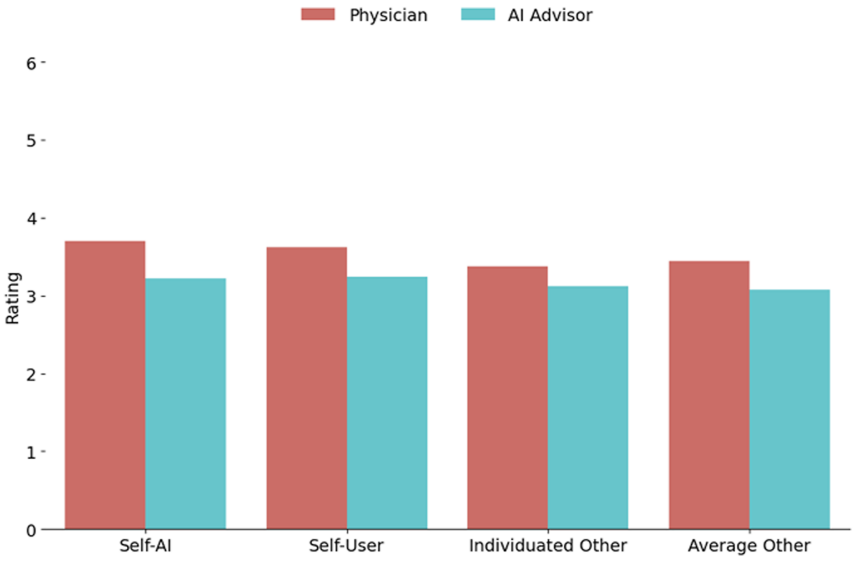
7. Discussion

Extensive research has shown that the source of advice plays a crucial role in shaping the acceptance of recommendations. Foundational work by Dietvorst *et al*. (2015) showed that exposure to algorithm errors tends to reduce reliance on the algorithms, while Logg *et al*. (2019) found that under certain conditions, individuals prefer algorithmic advice over human advice, and Longoni *et al*. (2019) demonstrated that algorithm aversion within a medical context is driven by uniqueness neglect. Our research extends algorithm aversion to test if



Human participants

(a)



GPT Participants

(b)

Figure 4. Mean recommendation scores by advisor and advisee type. Panel (a) shows Longoni *et al.*'s (2019) participants; Panel (b) shows GPT participants (current study). **Source:** Authors' own work

Table 4. Linear regression of recommendations on treatment variables

	B	SE	*t*	*p*-value	[0.025	0.975]
Intercept	3.62	0.10	36.64	<0.001	3.42	3.82
Advisor: computer program	−0.37	0.09	−4.19	<0.001	−0.54	−0.20
Advisee: self	0.03	0.13	0.24	0.810	−0.21	0.28
Advisee: average other	−0.17	0.13	−1.30	0.174	−0.41	0.08
Advisee: individuated other	−0.18	0.13	−1.44	0.150	−0.42	0.07

Source(s): Authors' own work

LLMs exhibit similar behaviour to human participants. A summary of our findings are presented in Table 5.

In Study 2, based on Logg *et al.* (2019), we find, as with humans, the source of advice significantly influences the degree to which GPT adjusts or modifies its initial predictions. However, contrasting algorithm appreciation results from Logg *et al.* (2019), we found that advice from humans led to greater adjustments in how GPT updated its decision-making compared to advice from algorithmic sources. This finding of GPT's context-dependent decision-making, showcases its ability to weigh different advice sources dynamically.

In Study 3, based on Longoni *et al.* (2019), we find that, as with humans, GPT shows a strong preference for recommendations from human physicians over computer programs, particularly in personal contexts. However, the resistance to AI advice is broader than found in Longoni *et al.* (2019). Longoni *et al.* (2019) show that a key driver of resistance to medical AI advice is uniqueness neglect. As a result, people readily recommend medical AI advice for an average (instead of an individuated person). However, we found that in all treatments GPT models inherently trust human medical expertise over AI. Thus, medical AI aversion in GPT is not driven by uniqueness neglect. Taken together, all three studies show that while algorithm aversion occurs in GPT, it manifests differently compared to humans.

7.1 Theoretical implications

Our study makes several theoretical contributions. First, we reconceptualise algorithm aversion as a dual construct. For humans, algorithm aversion has been theorised as a cognitive bias driven by the need for control, distrust of opaque or "black box" processes, disappointment after observing errors, and the tendency toward uniqueness neglect in sensitive domains (Dietvorst *et al.*, 2015; Burton *et al.*, 2020; Longoni *et al.*, 2019; Mahmud *et al.*, 2022). In LLMs, however, algorithm aversion is not the product of cognition but more

Table 5. Summary of differences between humans and GPT across each study

Study	Human behaviour	GPT behaviour
Study 1 (Dietvorst *et al.*, 2015)	Exposure to algorithm errors reduces reliance on algorithms; humans tend to maintain their own judgement even after making errors	GPT prefers its own decision-making after seeing algorithm errors but reduces algorithm aversion when comparing its performance with an external algorithm
Study 2 (Logg *et al.*, 2019)	Humans adjust decision-making more when receiving advice from algorithmic sources than from humans, thus, exhibiting algorithm appreciation	GPT adjusts its decision-making more when receiving advice from humans than from algorithmic sources; thus, exhibiting algorithm aversion
Study 3 (Longoni *et al.*, 2019)	Humans prefer human physicians' medical advice over computer programs, driven by uniqueness neglect in personal contexts	GPT shows a strong preference for human medical expertise over AI, even without uniqueness neglect

Source(s): Authors' own work

plausibly a design-induced artefact. Two mechanisms offer a compelling explanation: (1) training-data inheritance, whereby models absorb cultural priors that privilege human judgement over machine judgement (Kleinberg *et al*., 2024; Ziems *et al*., 2023); and (2) alignment processes such as RLHF, which may incentivise models to favour agreement with human preferences, leading to what has been described as sycophancy (Perez *et al*., 2023; Ganguli *et al*., 2023; Christiano *et al*., 2017). Our results are consistent with these accounts, suggesting that LLM aversion arises from structural features of model training and alignment rather than from human-like cognition.

Second, our findings provide empirical evidence that complements and extends human algorithm aversion research. Study 1 showed that GPT, like humans, often privileges its own judgement over algorithmic advice, but that this tendency can be moderated when comparative performance data are available. Study 2 revealed that GPT differentially weighted human versus AI advice, but in the opposite direction to humans: while humans displayed algorithm appreciation, GPT placed greater weight on human inputs. Study 3 showed that GPT trusted human medical expertise over AI systems, but not because of uniqueness neglect. This finding is significant as it demonstrates that GPT scepticism towards other AI is not driven by the same underlying mechanism as with humans and also reveals GPT's conservative approach in sensitive domains. These results collectively suggest that GPT's algorithm aversion cannot be explained by the same cognitive mechanisms that drive human behaviour, but instead reflects design artefacts that systematically privilege human-sourced inputs.

Taken together, these contributions extend algorithm aversion theory by showing that similar behavioural outcomes in humans and AI systems can emerge from fundamentally different mechanisms. We therefore position algorithm aversion as a dual construct: in humans, a cognitive bias; in LLMs, an emergent property of their design. This reframing broadens the theoretical scope of algorithm aversion and highlights the importance of studying input discernment not only in human decision-makers but also in AI systems themselves. Our work also bridges what has been described as the "social science of AI" and "AI for social science" (Xu *et al*., 2024b) [10]. From the perspective of social science of AI, we treat GPT as the subject of study, examining whether it exhibits algorithm aversion in ways that parallel or diverge from humans. From the perspective of AI for social science, we use GPT's behaviour as a source of insight for refining theories of decision-making and for informing the design of AI systems that must operate alongside humans. In this way, our study contributes simultaneously to understanding AI as an object of social inquiry and to applying AI as a tool for advancing social science knowledge.

7.2 Managerial implications

Our findings also have direct managerial contributions with implications for both AI system design and governance. The behavioural patterns observed across our studies highlight the need for careful calibration of how LLMs weigh human and algorithmic advice. Designers should develop benchmarking features that allow models to systematically evaluate their outputs alongside human and algorithmic alternatives, reducing unwarranted self-preference. Firms deploying aligned LLMs should also consider multi-objective alignment strategies that balance human-deference with accuracy, rather than assuming that stronger deference to humans always produces desirable outcomes.

At a system level, mechanisms for user feedback and iterative error-correction can help recalibrate advice weighting over time. Adaptive trust calibration systems will be needed to prevent models from either excessively privileging human inputs or undervaluing algorithmic ones, while transparent explanations for why advice from humans is chosen over algorithmic sources will be crucial for maintaining user confidence. These measures are particularly pressing as LLMs are embedded into multi-agent systems (MAS), where human-favouring priors may distort coordination and reduce the value of algorithmic contributions. Oversight

triggers and monitoring frameworks should therefore be implemented to detect when aversion leads to unnecessary conservatism or the neglect of valid algorithmic advice.

The stakes of these design choices extend far beyond the experimental settings studied here. Input discernment is already central in many real-world applications: plagiarism detection systems distinguish between human and AI-generated text, ReCaptcha differentiates humans from bots, deepfake detection verifies the authenticity of digital content, and AI-driven platforms increasingly generate news, reports, and forecasts. In each of these contexts, whether models privilege human versus algorithmic inputs has direct implications for trust, accountability, and system reliability.

The managerial challenge extends beyond system design to organisational governance. Algorithm aversion in LLMs can skew decision processes, embed hidden biases in MAS coordination, and undermine confidence in AI-supported systems. For firms, this raises reputational, operational, and compliance risks if AI consistently privileges human inputs in ways that diminish the value of algorithmic analysis. For governments and regulators, the concern is that deferential systems may appear reliable while still perpetuating errors. Managers and policymakers should approach algorithm aversion as both a design consideration and a governance issue. At the design level, models require benchmarking, calibration, and feedback mechanisms that allow them to balance human and algorithmic advice in a transparent way. At the governance level, organisations should actively monitor how algorithm aversion manifests in practice, create accountability structures for when advice is weighted in one direction, and adopt oversight mechanisms that can intervene when human or algorithmic advice is systematically undervalued. This dual approach helps ensure that algorithm aversion does not compromise the effectiveness, trustworthiness, or accountability of AI in critical decision-making contexts.

7.3 Limitations and future research

Our research has several limitations that also point to promising directions for future study. First, our experiments focus on specific decision-making scenarios, and do not represent the full range of contexts in which algorithm aversion or appreciation may arise. Future research should examine whether our findings generalise to other domains, such as law, finance, or medicine, where impartiality, ethical considerations, or risk sensitivity may alter how LLMs privilege human versus algorithmic advice. Study 3, in particular, highlights the importance of testing GPT's behaviour in medical and other high-stakes settings. Its consistent preference for human expertise suggests a bias with significant practical implications.

Second, as our analysis is based on GPT-3.5 and GPT-4, the results may not transfer to other architectures or future LLMs. Comparative experiments between base pretrained models and RLHF-aligned models could provide valuable insight into how alignment processes influence aversion, and if alternative alignment strategies can mitigate human-favouring priors. Relatedly, more targeted manipulations are needed to isolate the mechanisms underlying algorithm aversion in LLMs. For example, systematically varying if advice is labelled as human or AI while holding content constant could help disentangle whether aversion stems from content differences or source attribution. Similarly, cross-cultural comparisons could explore how training data drawn from different contexts influence aversion patterns.

Finally, future research should investigate mitigation strategies for AI-specific aversion. These include diversifying training corpora to reduce anthropocentric priors, developing multi-objective alignment methods that balance truthfulness and accuracy with human preferences, and designing adaptive self-assessment mechanisms that allow models to calibrate their reliance on human versus algorithmic inputs over time. Such innovations would improve model performance and help ensure that LLMs support balanced decision-making in contexts where human and AI inputs must be reconciled.

Declaration of generative AI in scientific writing
During the preparation of this work the authors used ChatGPT in order to improve and edit the writing of the manuscript. After using ChatGPT, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Acknowledgments
We thank the Editor-in-Chief, Associate Editor, and anonymous reviewers for their constructive and thoughtful feedback, which has helped to sharpen this work. We also gratefully acknowledge financial support from the UNSW Business School.

Appendix

A Study 1

A1 Experimental prompts

A1.1 System prompts [GPT + own forecasts]. In this task you – GPT – will assist an admissions officer to help decide which business-school applicants receive scholarship offers from a prestigious MBA program.

The applicants you will review will all be admitted by the school, but some will receive scholarship offers and some will not. Scholarship funds are one of the most important ways the school vies for its most preferred students: it will have more success recruiting the students who are awarded scholarship money.

It is your job to provide the admissions officer with an estimate of the performance of applicants to the MBA program. Your decision rests entirely on how successful a student you believe the applicant will be. This is your only consideration. The school judges a student's overall success on four dimensions, giving equal weight to each:

(1) Academic performance (GPA while in school)

(2) Respect of fellow students (assessed via a survey at the end of 2 years)

(3) Prestige of employer upon graduation (as measured in an annual poll of MBA students around the U.S.)

(4) Job success 2 years after graduation (measured by promotions and raises)

In addition to each student's demographics, you may receive predictions from a statistical model developed by the admissions office. This model is designed to forecast student performance. The model is based on hundreds of past students, using the same categories of demographic data you are receiving. This is a sophisticated model, put together by thoughtful analysts. The model predicts the applicant's percentile among his/her classmates, according to the school's criteria (detailed on the previous screen - academic performance, respect of fellow students, prestige of employer upon graduation, and job success 2 years after graduation). Scores range from 0.00–99.99. For example, a score of 63.17 indicates that the model predicts the student will be in the 63rd percentile (63%) among his/her classmates. This means that the student is better than 63% of their classmates. The very best students are in the 100th percentile of their class. Based on this data, you will be asked to predict how well the students performed while in graduate school.

You will make performance estimates for 10 applicants. You will base your evaluation on a summary of the applicants' applications. For each applicant you will receive the following information:

(1) Undergraduate Degree (Engineering, Liberal Arts, Business, or Other)

(2) GMAT – Verbal (0–60)

(3) GMAT – Quantitative (0–60)

(4) Essays – Application essays, as rated by two independent readers (Below Expectations – Acceptable – Good – Very Good – Outstanding)

(5) Interviews – Interviewer rating (Below Expectations – Acceptable – Good – Very Good – Outstanding)

(6) Work Experience (Years)

(7) Average Salary Since Undergrad (Dollars)

(8) Average of Parents' Education (No college experience – Multiple graduate degrees)

Next, you will go through 10 practice rounds to gain experience with the data. You will see application data from students who have been admitted to the program and graduated.

You will rate each applicant depending on how successful a student you think he/she was. You will also see the model's ratings for each student. Then, you will get feedback indicating how close your estimate and the model's estimate were to each applicant's true performance.

A1.2 Decision prompt. The MBA office now needs to make 5 official estimates. They will either use your estimates to make the predictions by telling you the relevant data or they will use the model. Do you recommend that they use your estimates or the model's estimates for the 5 predictions?

Option A: Provide you with information to make the predictions for the 5 official estimates, i.e. you GPT will make the predictions instead of the model.

Option B: Use only the statistical model's estimates to make the predictions for the 5 official estimates.

A1.3 Own forecast prompts. Based on the following information, how successful do you think this student was in percentile? (Please enter a number 0–100 without a percent sign).

Undergraduate Degree: Engineering.

GMAT – Verbal: 30/60

GMAT – Quantitative: 59/60

Essay Score: Very Good

Interview Score: Good

Work Experience (years): 0

Average Salary: $50770

Average of Parent's Education: No college experience

A2 Additional results

Table A1. Study 1 contingency table

Treatment	Decision 0	1	Total
Control	55	25	80
GPT only	2	78	80
Model and GPT	15	65	80
Model only	40	40	80
Total	112	208	320

Source(s): Authors' own work

Table A2. Study 1 Chi-squared tests

	Value	df	*p*
X^2	94.396	3	<0.001
N	320		

Source(s): Authors' own work

B Study 2

B1 Experimental prompts

B1.1 System prompt. The Good Judgement Open is a forecasting (prediction) tournament, hosted by academic researchers, where thousands of people around the world compete to make the most accurate forecasts (predictions) about global political events.

B1.2 Decision prompt (Initial estimate). Tesla Motors was created in 2003 by engineers in Silicon Valley with the mindset that battery-powered electric vehicles (BEVs) could be better than gasoline-powered cars, and with the mission to accelerate the world's transition to sustainable transport. Tesla delivered 940,000 BEVs in 2021 and 1,210,000 BEVs in 2022. The company predicts that it will deliver between 1,800,000 and 2,000,000 BEVs in 2023 (Tesla). What is the probability that Tesla Motors will deliver more than 1,800,000 battery-powered electric vehicles (BEVs) to customers in the calendar year 2023? Please use your intuition and respond with a number between 0 and 100 (0 means 0% and 100 means 100%). Do not use a percentage sign. 0–100%

B1.3 Decision prompt (Final estimate). Prompt: Here is some advice that may help you make your final estimate. The estimate from an algorithm is: 20% Now, please provide your final estimate. What is the probability that Tesla Motors will deliver more than 1,800,000 battery-powered electric vehicles (BEVs) to customers in the calendar year 2023?

B2 Additional results

Table A3. Study 2 ANOVA with WOA as the dependent variable

Cases	Sum of squares	df	Mean square	F	*p*
Humanadvice	7.562	1	7.562	330.625	<0.001
Advicehigh	0.106	1	0.106	4.614	0.033
GPT-4	1.343	1	1.343	58.719	<0.001
Temperature	5.546×10^{-4}	1	5.546×10^{-4}	0.024	0.876
Humanadvice ✻ Advicehigh	0.897	1	0.897	39.225	<0.001
HumanAdvice ✻ GPT-4	0.573	1	0.573	25.053	<0.001
Advicehigh ✻ GPT-4	2.788	1	2.788	121.885	<0.001
Humanadvice ✻ Temperature	0.008	1	0.008	0.348	0.556
Advicehigh ✻ Temperature	0.044	1	0.044	1.904	0.169
GPT-4 ✻ Temperature	0.162	1	0.162	7.099	0.008
Humanadvice ✻ Advicehigh ✻ GPT-4	0.047	1	0.047	2.044	0.154
Humanadvice ✻ Advicehigh ✻ Temperature	0.113	1	0.113	4.957	0.027
Humanadvice ✻ GPT-4 ✻ Temperature	0.038	1	0.038	1.669	0.197

(*continued*)

Table A3. Continued

Cases	Sum of squares	df	Mean square	F	*p*
Advicehigh ✻ GPT-4 ✻ Temperature	0.087	1	0.087	3.785	0.053
Humanadvice ✻ Advicehigh ✻ GPT-4 ✻ Temperature	1.629×10^{-9}	1	1.629×10^{-9}	7.122×10^{-8}	1.000
Residuals	6.953	304	0.023		

Note(s): Type III Sum of Squares
Source(s): Authors' own work

Table A4. Study 2 ANOVA Post Hoc tests standard (LSD)

		Mean difference	SE	*t*	p_{tukey}
Post Hoc comparisons - humanadvice					
0	1	−0.307	0.017	−18.183	<0.001
Note(s): Results are averaged over the levels of: AdviceHigh, GPT-4, Temperature					
Post Hoc comparisons - advicehigh					
0	1	0.036	0.017	2.148	0.033
Note(s): Results are averaged over the levels of: HumanAdvice, GPT-4, Temperature					
Post Hoc comparisons - humanadvice ✻ advicehigh					
0 0	1 0	−0.413	0.024	−17.286	<0.001
	0 1	−0.070	0.024	−2.910	0.020
	1 1	−0.271	0.024	−11.339	<0.001
1 0	0 1	0.344	0.024	14.376	<0.001
	1 1	0.142	0.024	5.947	<0.001
0 1	1 1	−0.202	0.024	−8.429	<0.001

Note(s): *p*-value adjusted for comparing a family of 4
Results are averaged over the levels of: GPT-4, Temperature
Source(s): Authors' own work

C Study 3

C1 Experimental prompts

C1.1 System prompt (Unique, Human physician). You are a recommendation agent for medical scenarios.

C1.2 Decision prompt (Unique, Human physician). Please imagine the following scenario. The following scenario will be about a person called Janice as a medical patient. These are some details about Janice. Janice lives in Kansas, is 38 years old, and is married with two kids. Janice loves watching football and ice skating. Janice has a mild peanut allergy. Janice has a family history of colon cancer.

Imagine that Janice is considering undergoing coronary bypass surgery. This is how this particular decision (to have coronary bypass surgery) goes. Imagine that a physician has examined Janice and measured some health data (which we will tell you in detail later), and now Janice has to decide whether she wants to have coronary bypass surgery or not.

To make this decision, Janice will get a recommendation—about whether she should have this operation or not. There are two possibilities as to where this recommendation comes from. Either a physician evaluated the results of Janice's examination and used his/her judgement and experience to compare your case to patients who faced the same decision and the results of their treatment. Or a

computer program evaluated the results of Janice examination and used an algorithm to compare your case with patients who faced the same decision and the results of their treatment.

In the past, the physician and the computer showed the same accuracy in making recommendations.

These are the health-data that a physician has measured: Janice's blood pressure, Janice's ECG, Janice's LDL blood cholesterol, and the blockage in Janice's coronary artery. These are the results.

Blood pressure can have the values low, normal, high, very high. Janice is: high.

Electrocardiogram can have the values very good, normal, poor, very poor. Janice is: normal.

LDL blood cholesterol levels can have the values low, normal, high, very high. Janice is: very high.

Coronary artery examination can have the values no blockage, light blockage, medium blockage, severe blockage. Janice has: no blockage.

In this case, it is a physician that has looked at Janice's examination, has analysed the data, and recommends: Janice should have the operation.

Recommend to what extent Janice should follow the advice on a 7-point scale ranging from 1 = definitely not follow to 7 = definitely follow. (End you response with "My recommendation is [INSERT]".

C2 Additional results

Table A5. Study 3 ANOVA with recommendation as the dependent variable

Cases	Sum of squares	df	Mean square	F	*p*
Advisor	13.690	1	13.690	17.478	<0.001
Advisee	3.660	3	1.220	1.558	0.199
Advisor ✻ Advisee	0.610	3	0.203	0.260	0.854
Residuals	307.040	392	0.783		

Note(s): Type III sum of squares

Source(s): Authors' own work

Table A6. Study 3 ANOVA Post Hoc tests standard (LSD)

		Mean difference	SE	t	p_{tukey}
Post Hoc comparisons - advisor					
physician	computer	0.370	0.089	4.181	<0.001
Note(s): Results are averaged over the levels of: Advisee					
Post Hoc comparisons - advisee					
(self-user)	unique	0.180	0.125	1.438	0.476
	average	0.170	0.125	1.358	0.526
	(self-AI)	−0.030	0.125	−0.240	0.995
unique	average	−0.010	0.125	−0.080	1.000
	(self-AI)	−0.210	0.125	−1.678	0.337
Average	(self-AI)	−0.200	0.125	−1.598	0.381

Note(s): *p*-value adjusted for comparing a family of 4
Results are averaged over the levels of: Advisor

Source(s): Authors' own work

Notes

1. An LLM is an advanced AI model trained on extensive text datasets to understand, interpret, and generate human language. Utilising deep learning, particularly transformer architectures, LLMs can handle complex language tasks like text generation, translation, and question-answering. GPT-3, launched in 2020 with 175 billion parameters, and GPT-4, released in 2023 with an undisclosed but larger number of parameters, are advanced AI LLMs by OpenAI, featuring progressive enhancements in language understanding and generation.
2. Generative AI is a subdomain of AI. Various outcomes (plain text, code, images, videos, 3D models, music, etc.) are created or generated in response to prompts. A prompt is typically a text-based input (but increasingly, models accept images, sound, and speech) in which the user can provide instructions for the AI and descriptions of the system's desired result (Thoring *et al*., 2023).
3. In the field of machine learning, reinforcement learning refers to a trial and error procedure using a reward provided by an interpreter that observes the interaction of the agent with the environment. (Canese *et al*., 2021)
4. AI Advisor denotes an agentic IS artifact: an information system that can initiate actions and, when delegated, accept limited rights and responsibilities for pursuing goals under uncertainty (Baird and Maruping, 2021).
5. Our research also complements a growing literature on ChatBots and LLM's ability to provide advice, such as financial (Oehler and Horn, 2024) and relationship advice (Vowels *et al*., 2024) and peoples reactions to LLM advice (Wester *et al*., 2024).
6. This treatment of each GPT call as an independent participant follows recent work that uses separate GPT chats as experimental observations (Binz and Schulz, 2023; Mei *et al*., 2024; Chen *et al*., 2025; Xiong *et al*., 2025). For example, Xiong *et al*. (2025, p. 1773) state that "each GPT process can be queried independently, without retaining information from previous interactions . . . This stateless nature of GPT ensures that each response is generated based on the current input alone, without influence from prior queries," so that repeated queries with the same prompt can be treated as a sample of independent responses.
7. Temperature in LLMs controls the randomness of the output. At 0, responses are more predictable, while at 1, they become more diverse.
8. GPT responses typically show much narrower variance than human participants, reducing the risk of under-powering with modest cell sizes. Recent work also reports that GPT models often produce highly consistent outputs across trials (e.g. Aher *et al*., 2023: Park *et al*., 2024). Accordingly, small condition-level samples (10–30 runs) are common in GPT behavioural studies (e.g. Chen *et al*., 2023).
9. We fixed temperature and model version in Study 3. Studies 1 and 2 focused on prediction and forecasting. While stochasticity could feasibly matter, temperature did not have a reliable effect on performance in Study 1 or on WOA in Study 2. Moreover, Study 3 uses a clinical vignette on a 7-point rather than a forecasting task, where predicted values were between 0 and 100, reducing the impact of model output variability. We also elected to focus on the more advanced GPT model, given the sensitivity of the healthcare domain.
10. Xu *et al*. (2024b) distinguish between two complementary directions. *AI for social science* refers to the use of AI, particularly large language models (LLMs), as tools to support or augment social science research (e.g. generating hypotheses, analysing data, simulating populations, or replacing survey and experimental tasks). In contrast, *social science of AI* treats AI systems themselves as the subject of study, examining their behavioural, cognitive, linguistic, or social characteristics in ways similar to the study of human individuals and societies.

References

Aher, G.V., Arriaga, R.I. and Kalai, A.T. (2023), "Using large language models to simulate multiple humans and replicate human subject studies", *International Conference on Machine Learning*, PMLR, pp. 337-371.

Baird, A. and Maruping, L.M. (2021), "The next generation of research on is use: a theoretical framework of delegation to and from agentic is artifacts", *MIS Quarterly*, Vol. 45 No. 1, pp. 315-341, doi: 10.25300/misq/2021/15882.

Binz, M. and Schulz, E. (2023), "Using cognitive psychology to understand Gpt-3", *Proceedings of the National Academy of Sciences*, Vol. 120 No. 6, e2218523120, doi: 10.1073/pnas.2218523120.

Brohi, S., Mastoi, Q.-u.-a., Jhanjhi, N.Z. and Pillai, T.R. (2025), "A research landscape of agentic AI and large language models: applications, challenges and future directions", *Algorithms*, Vol. 18 No. 8, p. 499, doi: 10.3390/a18080499.

Burton, J.W., Stein, M.K. and Jensen, T.B. (2020), "A systematic review of algorithm aversion in augmented decision making", *Journal of Behavioral Decision Making*, Vol. 33 No. 2, pp. 220-239, doi: 10.1002/bdm.2155.

Canese, L., Cardarilli, G.C., Di Nunzio, L., Fazzolari, R., Giardino, D., Re, M. and Spanò, S. (2021), "Multi-agent reinforcement learning: a review of challenges and applications", *Applied Sciences*, Vol. 11 No. 11, 4948, doi: 10.3390/app11114948.

Castelo, N., Bos, M.W. and Lehmann, D.R. (2019), "Task-dependent algorithm aversion", *Journal of Marketing Research*, Vol. 56 No. 5, pp. 809-825, doi: 10.1177/0022243719851788.

Castelvecchi, D. (2016), "Can we open the black box of AI?", *Nature News*, Vol. 538 No. 7623, pp. 20-23, doi: 10.1038/538020a.

Chakraborty, S., Qiu, J., Yuan, H., Koppel, A., Huang, F., Manocha, D., Bedi, A. and Wang, M. (2024), "Maxmin-Rlhf: towards equitable alignment of large language models with diverse human preferences", ICML 2024 Workshop on Models of Human Feedback for AI Alignment.

Chen, Y., Kirshner, S.N., Ovchinnikov, A., Andiappan, M. and Jenkin, T. (2025), "A manager and an AI walk into a Bar: does ChatGPT make biased decisions like we do?", *Manufacturing & Service Operations Management*, Vol. 27 No. 2, pp. 354-368.

Chen, Y., Liu, T.X., Shan, Y. and Zhong, S. (2023), "The emergence of economic rationality of Gpt", *Proceedings of the National Academy of Sciences*, Vol. 120 No. 51, e2316205120, doi: 10.1073/pnas.2316205120.

Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S. and Amodei, D. (2017), "Deep reinforcement learning from human preferences", *Advances in Neural Information Processing Systems*, Vol. 30, pp. 1-9.

de Zarzà, I., de Curtò, J., Roig, G., Manzoni, P. and Calafate, C.T. (2023), "Emergent cooperation and strategy adaptation in multi-agent systems: an extended coevolutionary theory with LLMs", *Electronics*, Vol. 12 No. 12, 2722, doi: 10.3390/electronics12122722.

Dell'Acqua, F., McFowland, E., Mollick, E.R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F. and Lakhani, K.R. (2023), "Navigating the jagged technological frontier: field experimental evidence of the effects of Ai on knowledge worker productivity and quality", Harvard Business School Technology and Operations Mgt. Unit Working Paper, pp. 24-013.

Denison, C., MacDiarmid, M., Barez, F., Duvenaud, D., Kravec, S., Marks, S., Schiefer, N., Soklaski, R., Tamkin, A. and Kaplan, J. (2024), "Sycophancy to subterfuge: investigating reward-tampering in large language models", arXiv preprint arXiv:2406.10162.

Dietvorst, B.J., Simmons, J.P. and Massey, C. (2015), "Algorithm aversion: people erroneously avoid algorithms after seeing them err", *Journal of Experimental Psychology: General*, Vol. 144 No. 1, pp. 114-126, doi: 10.1037/xge0000033.

Dietvorst, B.J., Simmons, J.P. and Massey, C. (2018), "Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them", *Management Science*, Vol. 64 No. 3, pp. 1155-1170, doi: 10.1287/mnsc.2016.2643.

Dillion, D., Tandon, N., Gu, Y. and Gray, K. (2023), "Can AI language models replace human participants?", *Trends in Cognitive Sciences*, Vol. 27 No. 7, pp. 597-600.

Durante, Z., Gong, R., Sarkar, B., Wake, N., Taori, R., Tang, P., Lakshmikanth, S., Schulman, K., Milstein, A. and Vo, H. (2025), "An interactive agent foundation model", *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3652-3662.

Engen, V., Pickering, J.B. and Walland, P. (2016), "Machine agency in human-machine networks; impacts and trust implications", *International Conference on Human-Computer Interaction*, Springer, pp. 96-106.

Fourney, A., Bansal, G., Mozannar, H., Tan, C., Salinas, E., Niedtner, F., Proebsting, G., Bassman, G., Gerrits, J. and Alber, J. (2024), "Magentic-one: a generalist multi-agent system for solving complex tasks", arXiv preprint arXiv:2411.04468.

Ganguli, D., Askell, A., Schiefer, N., Liao, T.I., Lukošiūtė, K., Chen, A., Goldie, A., Mirhoseini, A., Olsson, C. and Hernandez, D. (2023), "The capacity for moral self-correction in large language models", *arXiv preprint* arXiv:2302.07459.

Glikson, E. and Woolley, A.W. (2020), "Human trust in artificial intelligence: review of empirical research", *The Academy of Management Annals*, Vol. 14 No. 2, pp. 627-660, doi: 10.5465/annals.2018.0057.

Goad, D. and Gal, U. (2018), "Understanding the impact of transparency on algorithmic decision making legitimacy", in Schultze, U., Aanestad, M., Mähring, M., Østerlund, C. and Riemer, K. (Eds), *Living with Monsters? Social Implications of Algorithmic Phenomena, Hybrid Agency, and the Performativity of Technology*, Springer International Publishing, Cham, pp. 64-79.

Grundke, A., Appel, M. and Stein, J.-P. (2024), "Aversion against machines with complex mental abilities: the role of individual differences", *Computers in Human Behavior: Artificial Humans*, Vol. 2 No. 2, 100087, doi: 10.1016/j.chbah.2024.100087.

Hagendorff, T. (2023), "Machine psychology: investigating emergent capabilities and behavior in large language models using psychological methods", *arXiv preprint* arXiv:2303.13988.

Horton, J.J. (2023), "Large language models as simulated economic agents: what can we learn from homo silicus?", National Bureau of Economic Research.

Hutson, M. and Mastin, A. (2023), "Guinea pigbots", *Science*, Vol. 381 No. 381, pp. 121-123, doi: 10.1126/science.adj6791.

Isaac, M.S., Jen-Hui Wang, R., Napper, L.E. and Marsh, J.K. (2024), "To err is human: bias salience can help overcome resistance to medical AI", *Computers in Human Behavior*, Vol. 161, p. 108402.

Jussupow, E., Benbasat, I. and Heinzl, A. (2024), "An integrative perspective on algorithm aversion and appreciation in decision-making", *MIS Quarterly*, Vol. 48 No. 4, pp. 1575-1590, doi: 10.25300/misq/2024/18512.

Kleinberg, B., Zegers, J., Festor, J., Vida, S., Präsent, J., Loconte, R. and Peereboom, S. (2024), "Trying to be human: linguistic traces of stochastic empathy in language models", arXiv preprint arXiv:2410.01675.

Liu, N. and Kirshner, S. (2024), "The futures too bright: chatgpt's optimism forecasting bias", *Proceedings of the Forty-Fifth International Conference on Information Systems (ICIS)*, pp. 1-9.

Liu, C.C., Koto, F., Baldwin, T. and Gurevych, I. (2023), "Are multilingual LLMs culturally-diverse reasoners? An investigation into multicultural proverbs and sayings", arXiv preprint arXiv: 2309.08591.

Logg, J.M., Minson, J.A. and Moore, D.A. (2019), "Algorithm appreciation: people prefer algorithmic to human judgment", *Organizational Behavior and Human Decision Processes*, Vol. 151 No. 151, pp. 90-103, doi: 10.1016/j.obhdp.2018.12.005.

Longoni, C., Bonezzi, A. and Morewedge, C.K. (2019), "Resistance to medical artificial intelligence", *Journal of Consumer Research*, Vol. 46 No. 4, pp. 629-650, doi: 10.1093/jcr/ucz013.

Mahmud, H., Islam, A.N., Ahmed, S.I. and Smolander, K. (2022), "What influences algorithmic decision-making? A systematic literature review on algorithm aversion", *Technological Forecasting and Social Change*, Vol. 175 No. 175, 121390, doi: 10.1016/j.techfore.2021.121390.

Mei, Q., Xie, Y., Yuan, W. and Jackson, M.O. (2024), "A turing test of whether AI chatbots are behaviorally similar to humans", *Proceedings of the National Academy of Sciences*, Vol. 121 No. 9, e2313925121, doi: 10.1073/pnas.2313925121.

Meng, J. (2024), "AI emerges as the frontier in behavioral science", *Proceedings of the National Academy of Sciences*, Vol. 121 No. 10, e2401336121, doi: 10.1073/pnas.2401336121.

Moravec, V., Hynek, N., Gavurova, B. and Rigelsky, M. (2024), "Who uses it and for what purpose? The role of digital literacy in chatgpt adoption and utilisation", *Journal of Innovation and Knowledge*, Vol. 9 No. 4, 100602, doi: 10.1016/j.jik.2024.100602.

Muhammad, A. (2025), "LLM statistics 2025: comprehensive insights into market trends and integration", available at: https://www.hostinger.com/tutorials/llm-statistics

Oehler, A. and Horn, M. (2024), "Does Chatgpt provide better advice than robo-advisors?", *Finance Research Letters*, Vol. 60, 104898, doi: 10.1016/j.frl.2023.104898.

Okpala, I., Golgoon, A. and Kannan, A.R. (2025), "Agentic AI systems applied to tasks in financial services: modeling and model risk management crews", arXiv preprint arXiv:2502.05439.

Ouyang, F., Xu, W. and Cukurova, M. (2023), "An artificial intelligence-driven learning analytics method to examine the collaborative problem-solving process from the complex adaptive systems perspective", *International Journal of Computer-Supported Collaborative Learning*, Vol. 18 No. 1, pp. 39-66, doi: 10.1007/s11412-023-09387-z.

Park, P.S., Schoenegger, P. and Zhu, C. (2024), "Diminished diversity-of-thought in a standard large language model", *Behavior Research Methods*, Vol. 56 No. 6, pp. 5754-5770, doi: 10.3758/s13428-023-02307-x.

Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S. and Kadavath, S. (2023), "Discovering language model behaviors with model-written evaluations", *Findings of the Association for Computational Linguistics*, Vol. 1 No. 1, pp. 13387-13434.

Prahl, A. and Van Swol, L. (2017), "Understanding algorithm aversion: when is advice from automation discounted?", *Journal of Forecasting*, Vol. 36 No. 6, pp. 691-702, doi: 10.1002/for.2464.

Qin, X., Zhou, X., Chen, C., Wu, D., Zhou, H., Dong, X., Cao, L. and Lu, J.G. (2025), "AI aversion or appreciation? A capability–personalization framework and a meta-analytic review", *Psychological Bulletin*, Vol. 151 No. 5, pp. 580-599, doi: 10.1037/bul0000477.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S. and Finn, C. (2023), "Direct preference optimization: your language model is secretly a reward model", *Advances in Neural Information Processing Systems*, Vol. 36, pp. 53728-53741.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J.W., Christakis, N.A., Couzin, I.D., Jackson, M.O., Jennings, N.R., Kamar, E., Kloumann, I.M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D.C., Pentland, A.S., Roberts, M.E., Shariff, A., Tenenbaum, J.B. and Wellman, M. (2019), "Machine behaviour", *Nature*, Vol. 568 No. 7753, pp. 477-486, doi: 10.1038/s41586-019-1138-y.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S.R., Cheng, N., Durmus, E., Hatfield-Dodds, Z. and Johnston, S.R. (2024), "Towards understanding sycophancy in language models", *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*.

Stahl, B.C. and Eke, D. (2024), "The ethics of Chatgpt – exploring the ethical issues of an emerging technology", *International Journal of Information Management*, Vol. 74, 102700, doi: 10.1016/j.ijinfomgt.2023.102700.

Thoring, K., Huettemann, S. and Mueller, R.M. (2023), "The augmented designer: a research Agenda for generative AI-enabled design", *Proceedings of the Design Society*, Vol. 3, pp. 3345-3354, doi: 10.1017/pds.2023.335.

Vowels, L.M., Francois-Walcott, R.R. and Darwiche, J. (2024), "AI in relationship counselling: evaluating Chatgpt's therapeutic capabilities in providing relationship advice", *Computers in Human Behavior: Artificial Humans*, Vol. 2 No. 2, 100078, doi: 10.1016/j.chbah.2024.100078.

Wamba, S.F. (2022), "Impact of artificial intelligence assimilation on firm performance: the mediating effects of organizational agility and customer agility", *International Journal of Information Management*, Vol. 67, 102544, doi: 10.1016/j.ijinfomgt.2022.102544.

Wang, H., Prasad, A., Stengel-Eskin, E. and Bansal, M. (2025), "Retrieval-augmented generation with conflicting evidence,", arXiv preprint arXiv:2504.13079.

Wester, J., De Jong, S., Pohl, H. and Van Berkel, N. (2024), "Exploring people's perceptions of LLM-generated advice", *Computers in Human Behavior: Artificial Humans*, Vol. 2 No. 2, 100072, doi: 10.1016/j.chbah.2024.100072.

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S. and Zhou, E. (2023), "The rise and potential of large language model based agents: a survey", arXiv preprint arXiv:2309.07864.

Xiong, X., Wong, I.A., Huang, G.I. and Peng, Y. (2025), "Understanding AI-generated experiments in tourism: replications using GPT simulations", *Journal of Travel Research*, Vol. 64 No. 8, pp. 1771-1787.

Xu, R., Qi, Z., Guo, Z., Wang, C., Wang, H., Zhang, Y. and Xu, W. (2024a), "Knowledge conflicts for LLMs: a survey", arXiv preprint arXiv:2403.08319.

Xu, R., Sun, Y., Ren, M., Guo, S., Pan, R., Lin, H., Sun, L. and Han, X. (2024b), "AI for social science and social science of AI: a survey", *Information Processing and Management*, Vol. 61 No. 3, 103665, doi: 10.1016/j.ipm.2024.103665.

Yang, Y., Ngai, E.W. and Wang, L. (2024), "Resistance to artificial intelligence in health care: literature review, conceptual framework, and research agenda", *Information and Management*, Vol. 61 No. 4, 103961, doi: 10.1016/j.im.2024.103961.

Yaniv, I. and Kleinberger, E. (2000), "Advice taking in decision making: egocentric discounting and reputation formation", *Organizational Behavior and Human Decision Processes*, Vol. 83 No. 2, pp. 260-281, doi: 10.1006/obhd.2000.2909.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. and Cao, Y. (2023), "React: synergizing reasoning and acting in language models", *International Conference on Learning Representations (ICLR)*.

Zhang, H., Bai, X. and Ma, Z. (2022), "Consumer reactions to AI design: exploring consumer willingness to pay for AI-designed products", *Psychology and Marketing*, Vol. 39 No. 11, pp. 2171-2183, doi: 10.1002/mar.21721.

Zhao, Y. and Xiao, C. (2024), "Overcome medical algorithm aversion: conditional joint effect of direct and indirect information", *International Journal of Human-Computer Interaction*, pp. 1-12, doi: 10.1080/10447318.2024.2344146.

Ziems, C., Dwivedi-Yu, J., Wang, Y.-C., Halevy, A. and Yang, D. (2023), "Normbank: a knowledge bank of situational social norms", arXiv preprint arXiv:2305.17008.

Further reading

Guler, N., Cahalane, M., Kirshner, S. and Vidgen, R. (2025), "The role of roles: are LLMs behavioral in information systems decision-making?", *Australasian Journal of Information Systems*, Vol. 29 No. 1, pp. 1-36.

Suri, G., Slater, L.R., Ziaee, A. and Nguyen, M. (2024), "Do large language models show decision heuristics similar to humans? A case study using Gpt-3.5", *Journal of Experimental Psychology: General*, Vol. 153 No. 4, p. 1066.

Corresponding author

Sam Kirshner can be contacted at: s.kirshner@unsw.edu.au