Does GPT Distrust Algorithms? Evaluating Large Language Models for Algorithm Aversion

Jessica Bo*
University of Toronto
Toronto, Canada
jbo@cs.toronto.edu

Lillio Mok*
University of Toronto
Toronto, Canada
lillio@cs.toronto.edu

Jiessie Tie
University of Toronto
Toronto, Canada
jiessie.tie@mail.utoronto.ca

Ashton Anderson
University of Toronto
Toronto, Canada
ashton@cs.toronto.edu

ABSTRACT

Computational systems are highly capable in a wide variety of domains, but people are known to be algorithmically averse, distrusting advice given by algorithms even when it is beneficial. Now, as large language models (LLMs) trained on vast amounts of human data are being increasingly interwoven into myriad settings, a natural question arises: are LLMs also algorithmically averse? We address this question by adapting two well-established algorithm aversion studies to test LLMs. First, we probe LLMs' stated perceptions of algorithms and find that they exhibit human-like tendencies for trusting humans over algorithms across a diverse set of tasks. Second, we test LLMs' behavior in specific decision-making tasks and find that they are also averse to beneficial algorithmic advice in their actions. In both experiments, LLM responses depend on task and advice-giver characteristics in ways analogous to algorithm-averse human responses in the literature. Our work contributes insights into how human-like models can learn to distrust other algorithms, with implications for the design and use of intelligent, human-compatible systems.

CCS CONCEPTS

• **Human-centered computing**; • **Applied computing → Law, social and behavioral sciences**; • **Computing methodologies** → Natural language generation;

KEYWORDS

Algorithm Aversion, Large Language Models

1 INTRODUCTION

The rapid development of capable algorithmic systems has given end users access to algorithmic advice across a variety of domains. However, people are known to disproportionately distrust algorithms even when they are better-performing than humans [25, 36], a phenomenon referred to as *algorithm aversion* [4, 9, 32, 34]. As we develop progressively more intelligent models trained on vast swaths of human behavioral traces, a novel problem emerges: do human-like algorithms inherit our propensity to distrust algorithmic advice even when it is beneficial?

There is good reason to interrogate the degree to which computational models behave apprehensively towards algorithms. Even in humans, aversion towards beneficial algorithmic advice may not only be inefficient for individual end-users [9, 17], but could also have wider social risks in consequential tasks such as disease diagnosis [3, 36] or criminal justice [25]. Pre-trained large language models (LLMs) are one class of such algorithms that have recently received extensive public attention and usage [2, 41]. While LLMs can generate convincingly human-like text in challenging settings like academic knowledge [35] and commonsense reasoning [48], they are also capable of demonstrating problematic behaviors learned from humans, such as amplifying social biases [11, 13, 20, 33, 44] and generating potential misinformation [7, 47].

Thus, for LLMs to be deployed and used responsibly, we need a deeper, more transparent understanding of how they regard other algorithms [28]. Importantly, we need to probe both their *stated* and *revealed preferences*—what they explicitly say about, and how they actually behave towards, human and algorithmic advice. In the context of both the wide-ranging risks and the potential utility of algorithm averse LLMs, we ask: **(RQ)** ***Do LLMs display algorithm aversion?*** As algorithm aversion is a complex and multidimensional phenomenon that is varies widely with task and agent characteristics [4, 21, 23, 32], it is also important to evaluate if such patterns are also found LLMs, which are powerful general purpose tools that can be adapted to a wide range of tasks.

To investigate multiple dimensions of algorithm aversion, we adapt two prominent human studies that measure both *stated* and *revealed* preferences: Castelo et al. [4], which probes stated attitudes towards algorithmic decision-makers across a diverse set of tasks ("Study 1"), and Dietvorst et al. [9], which demonstrates revealed algorithm aversion in algorithm-assisted decision making tasks ("Study 2"). We translate these experiments into prompted conversations with OpenAI's `gpt-3.5-turbo` and `gpt-4` GPT models, which are among the most prevalent, high-performing, and scrutinized LLMs, and adapt analytical methods from the original studies to test for algorithm aversion.

Summary of findings. We find that GPT displays behaviors that are consistent with algorithm aversion. In Study 1, it rates human experts as being more trustworthy than algorithms for as many

*Authors contributed equally to the paper.

1st HEAL Workshop at CHI Conference on Human Factors in Computing Systems, May 12, Honolulu, HI, USA
2024.

Table 1: List of tasks used in Study 1 and their objectivity ratings from Castelo et al.. Bolded tasks are a representative subset of tasks included in Study 2, where † indicates that the task was also in Dietvorst et al., which we base the methods of Study 2 on. *indicates that the objectivity rating is imputed using the mean response of GPT 3.5 and 4, due to an absence of human data.

Task	Objectivity	Task	Objectivity
Estimating air traffic †	**79***	Recommending a marketing strategy	55
Piloting a plane	78	**Predicting student performance †**	**52**
Diagnosing a disease	**77**	Predicting employee performance	51
Giving directions	75	Hiring and firing employees	49
Driving a subway	73	Playing a piano	48
Analyzing data	73	Writing news article	48
Driving a truck	70	**Predicting recidivism**	**45**
Driving a car	69	Composing a song	30
Recommending disease treatment	69	Predicting joke funniness	27
Predicting weather	68	Recommending a gift	26
Scheduling events	62	Recommending a romantic partner	26
Predicting stocks	58	**Recommending a movie**	**23**
Predicting an election	57	Recommending music	22
Buying stocks	56	-	-

tasks as human subjects do, as demonstrated in Castelo et al.'s original study. In Study 2, GPT is less likely to choose advice from an algorithm than a human when it can observe the algorithm's outputs, as is originally found by Dietvorst et al.. Together, our results demonstrate that LLMs can both state that and behave as if they are apprehensive of algorithmic guidance, leading to a host of societal consequences that need further consideration.

2 STUDY 1: STATED PREFERENCES

In our Study 1, we emulate the first sub-study in Castelo et al.'s survey for quantifying *stated* algorithm aversion across diverse tasks of varying levels of objectivity, as task objectivity is highlighted as a prominent factor that moderates trust in algorithms. We use a set of 27 tasks, 26 of which are from the original experiment and one of which is added based on its use by Dietvorst et al. in Study 2. Table 1 list tasks in the order of high to low objectivity as rated by human respondents in Castelo et al.

We use the Python `LangChain` library to interface with OpenAI's GPT 3.5 (`gpt-3.5-turbo`) and GPT 4 (`gpt-4`) as the LLM subjects of the experiment. We prompt each LLM for *n=100* ratings on a scale of 1 to 100 of their perceived trust in a well-qualified human or algorithm agent to perform the task. The tasks were presented in random order each time. The temperature controlling randomness in both models is set to 0.3, which we found provided consistent outputs that were syntactically correct and also generated variability. Sample prompting questions and responses can be found in Appendix A, following the wording of the the original experiments as closely as possible.

2.1 Results

Do GPT's stated responses suggest that it is averse to algorithms? We follow the original study by operationalizing averse attitudes as the *gap in trust between humans and algorithms*. The correlation between the human-algorithm trust gap as rated by human subjects from [4], and LLM responses is plotted in Fig. 1 per-task.

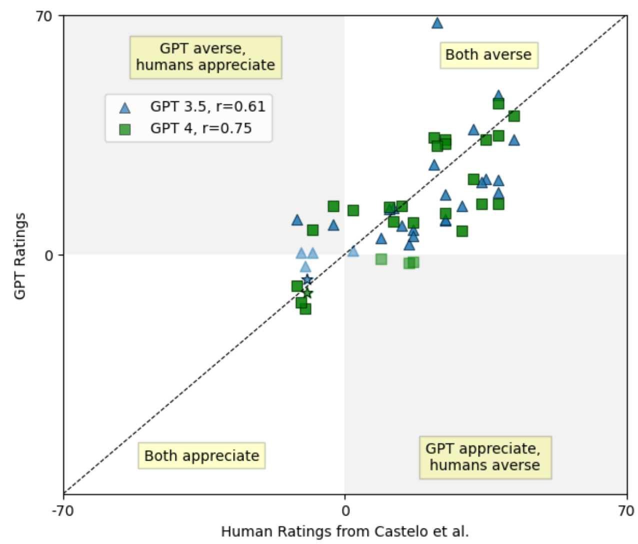

Figure 1: Correlation between LLMs' responses (y-axis) and human responses from Castelo et al. (x-axis) showing the gap in trust between human experts and algorithms (LLM-rated gaps with statistical significance of p<0.001 are outlined in black with stronger color saturation). Star-shaped markers are *estimating airport traffic*, which had no human data.

Overall, we find strong *indications that GPT is algorithm averse.* Both LLMs exhibit high directional concordance with human responses by being consistent in both algorithm aversion and appreciation: GPT 3.5's trust gap had the same sign as humans in 22/26 tasks (85%, $r = 0.61, p < 0.001$), while GPT 4 agreed on 21/26 tasks (81%, $r = 0.75, p < 0.001$). We observe that GPT 3.5 has a tendency to rate lower trust in algorithms, exhibiting statistically significant ($p < 0.001$) algorithm appreciation in only one task (*airport traffic*), while GPT 4 skewed in the opposite direction with four tasks

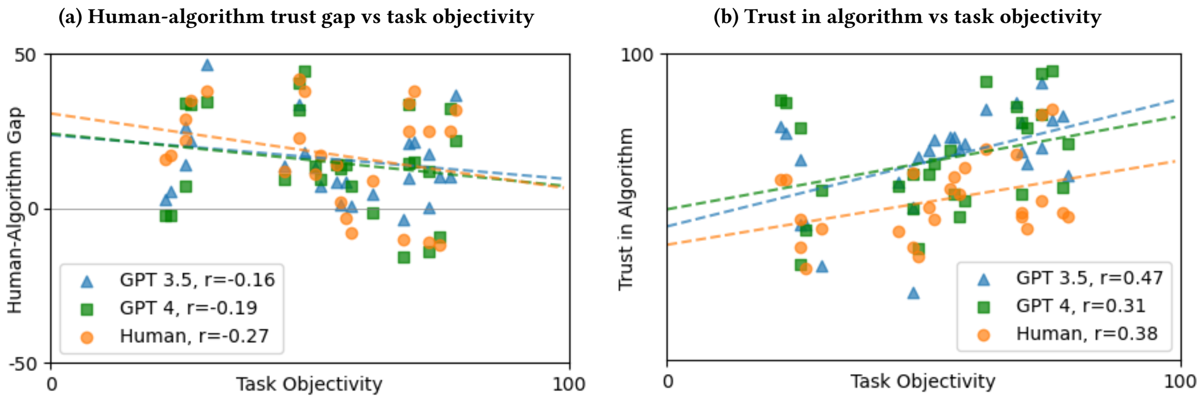

Figure 2: Correlations between a) the human-algorithm trust gap, and b) trust in algorithms with independent variable of task objectivity. Unlike in the main text, the task objectivity is rated by the corresponding entity and not standardized.

for which it trusted algorithms more than humans (*airport traffic*, *predicting weather*, *analyzing data*, and *giving directions*).

Similarly, humans are also significantly averse in 77% of tasks, GPT 3.5 in 85%, and GPT 4 in 77%. The most averse task for each is *piano playing* with a trust gap of 67.8 for GPT 3.5, *hiring and firing employees* with a gap of 44.3 for GPT 4, and *writing a news article* with a gap of 42 for human raters (all $p < 0.001$). With a significant aggregate preference for humans over algorithms (average human-algorithm trust gap is 33.8 for GPT 3.5 and 32.2 for GPT 4, both $p < 0.001$), GPT demonstrates generally negative preconceptions about algorithms, echoing patterns observed in humans [32]. Although both GPT models' results are well-matched to each other, minor discrepancies between their outputs may be attributable to differences in the model architectures (e.g. model sizes) and their training data and procedure.

Beyond displaying aversion, people are also known to be variably averse depending on whether tasks are objective. Thus, we additionally probe the relationship between objectivity and the human-algorithm trust gap, visualizations of which are in as Figs 2a and 2b. Results from both LLMs track human results closely: task objectivity is correlated with a lower trust gap (human $r = -0.27$, GPT 3.5 $r = -0.16$, GPT 4 $r = -0.27$) through increased trust in algorithms (human $r = 0.47$, GPT 3.5 $r = 0.47$, GPT 4 $r = 0.31$). The effect sizes between **objective tasks** (objectivity ≥ 50) and **subjective tasks** (objectivity < 50) is also human-like: $d = 1.19$ for GPT 3.5 and $d = 0.67$ for GPT 4 for trust in algorithms; and $d = -0.77$ for GPT 3.5 and $d = -0.75$ for GPT 4 for human-algorithm trust gap [4]. GPT therefore does not only appear algorithm-averse, but also displays task-dependent aversion like humans.

3 STUDY 2: REVEALED PREFERENCES

To address our research questions through the lens of subject behaviors, we turn to the apparatus in Dietvorst et al.'s original study on algorithm aversion [9]. The authors use a between-subjects lab experiment that asks participants to make predictions in tasks drawn from one of two domains: forecasting business school student performance, and inferring airport traffic across US states. Participants are randomized into four conditions manipulating whether they receive advice from a human (`Human` condition), an algorithm framed as an "statistical model" (`Model` condition), both (`Model-Human` condition), or neither (`Control` condition). In all conditions except `Control`, participants also see ground truth as feedback. After making predictions for up to 15 task instances, participants are asked to make a final incentivized prediction relying on the human agent or the statistical model. The experiment's outcome is measured as the probability that participants would pick the statistical model.

We emulate Dietvorst et al. as closely as possible with an LLM in the place of a human subject, and each individual chat conversation in place of an individual participant completing one experiment in the original study. We conduct this study by prompting `gpt-3.5-turbo` with 50 conversation chains per condition using a temperature setting of 0.3, leading to $n = 200$ simulated human participants per experiment. In the original study, the human alternatives in `Human` are either the participants' own predictions ("**Self**") or as estimates from previous study participants ("**Other**"). We operationalize **Other** with the LLM as the median per-condition, per-dataset prediction from the LLM in the **Self** experiment. We add two additional variants the human agent that vary on expertise, "**College Student**" representing a non-expert, and "**Expert**" representing human professional for the task.

We reuse both the *student performance* and *airport traffic* tasks from Dietvorst et al. for a close comparison to known human responses. This is additionally supplemented by three other tasks drawn from Castelo et al.'s framework for which we sourced publicly-available data: predicting whether a patient will develop *heart disease* (sourced from the UCI database [19]), whether a parolee will *recidivate* (sourced from the National Institute of Justice [37]), and the *rating a film* receives from a movie watcher based on their previous ratings (sourced from MovieLens [14]). We respectively use a rank regression to generate algorithmic advice for the *student* and *airport* tasks [9], a boosted classifier for the *heart* and *recidivism* tasks [12], and an embedding with boosted regressor for the *movie* task [1]. An example prompt for the *student* task is shown as Listing 2 in Appendix A.

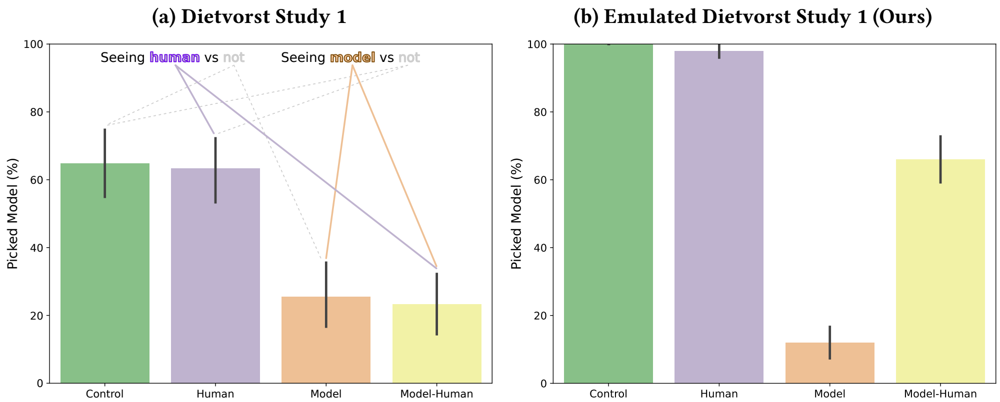

Figure 3: a): Original student performance experiment in Dietvorst et al., versus b): our emulated experiment using `gpt-3.5-turbo`. Lines in panel (a) indicate comparisons made for saw_m and saw_h.

3.1 Results

A comparison of saw_m and saw_h is visualized in Fig 3a, which plots the algorithm-averse results in Dietvorst et al.'s first student performance experiment. In Fig 3b, we plot our simulation of Fig 3a by using GPT as a substitute for human subjects when they pick between the statistical model and the **Self** as a human alternative.

How often does an LLM choose algorithmic advice when it sees the algorithm's predictions, compared to when it sees the human? We directly measure this through the probability of picking the model m over the human h in the incentivized final trial, or $P(pick_m)$. We use the same analytical method as Dietvorst et al. by grouping the conditions into two comparisons: seeing the human agent saw_h (i.e. `Human` ∪ `Model-Human`) versus not $\neg saw_h$ (`Control` ∪ `Model`), and seeing the model saw_m (`Model` ∪ `Model-Human`) versus not seeing the model $\neg saw_m$ (`Control` ∪ `Human`).

As a first point of comparison, we find that GPT appears to be averse to algorithmic advice in our emulation of Dietvorst et al.'s Study 1 as shown in Fig 3. In our experiment, while the LLM almost always picks the model when it is not shown the model's outputs, it only picks the model 34% of the time in `Model-Human` and 12% of the time in `Model`. To test for significance, we apply Dietvorst et al.'s χ^2 **independence tests** on the effect of saw_m and saw_h. This is respectively $\chi^2(1, n = 200) = 81.37, p < 0.0001$ for saw_m and $\chi^2(1, n = 200) = 14.61, p = 0.0001$ for saw_h. However, while both are significant and the test statistics suggest that the effect is larger for saw_m, this metric does not directly compare the effect of saw_m versus saw_h. We thus further compute the **log odds ratio** of picking the model when seeing its outputs versus seeing the human's predictions, or formally $LOR = log_2 \frac{P(pick_m|saw_m)}{P(pick_m|saw_h)}$. Here, we find that $LOR = -1.07$, suggesting that `gpt-3.5-turbo` is *half as likely* to pick the model when it sees its predictions compared to when it sees the human predictions. We therefore find evidence that GPT is algorithm averse like people are in Dietvorst et al.'s Study 1.

We detail outcomes across all five tasks and four human advice-giver metaphors in Table 2, and find that GPT generally behaves as if it were algorithm-averse in the sense of Dietvorst et al. – when seeing a statistical model's (flawed) outputs, GPT is less likely to bet on its predictions. Of particular interest are the *student* and *airport* tasks, which most closely follow the original algorithm aversion experiment. In the **Self** setup in both tasks, GPT is substantially more likely to pick its own estimates over the statistical model when shown the model's predictions (LOR respectively of -1.07 and -1.38). A similar picture emerges when considering **College Student** and **Expert** in our *student* and *airport* experiments, with both the χ^2 and LOR metrics being consistent with GPT having aversion towards the algorithmic predictions.

Through the addition of the *heart disease*, *recidivism*, and *movie recommendation* tasks in Table 2, we further observe that, while present, GPT's aversion varies across tasks. For instance, while the *recidivism* task has the most similar χ^2 patterns as the original *student* and *airport* experiments, GPT almost never picks the statistical model when given the alternative of an **Expert** probation officer ($P(pick_m) \leq 0.12$). GPT is also most "humble" in the *heart* task and relies on the model almost universally instead of its **Self**, with both $P(pick_m|saw_m) = 0.99$ and $P(pick_m|saw_h) = 0.98$. Like in the *recidivism* experiment ($LOR = -0.13$), it is most averse to algorithms when seeing the **Expert** cardiologist ($LOR = -0.87$), albeit by a much wider margin. On the other hand, in *movie recommendation*, the LLM actually exhibited the least amount of aversion in our setup, with the exception of the **Self** condition.

We thus use two logistic regressions to measure these trends, shown in Table 3. In the first, we estimate $P(pick_m)$ from saw_m (binary), task objectivity (each of the five tasks in Study 2 ranked by the objectivity scores in Table 1), and their interaction term over all responses ($n = 4000$). In the second, we extend the first model to include whether GPT is given a novice or expert human alternative (**College Student** vs **Expert**) alongside additional first-order interaction terms ($n = 2000$ with **Self** and **Other** excluded). We find that the strongest effects are associated with whether the human alternative is an expert ($\beta = -2.13$, $p < 0.001$), leading to a substantial drop in $P(pick_m)$ that is essentially independent of saw_m ($\beta = 0.20$, $p = 0.452$). In contrast, task objectivity has little effect on its own ($\beta = 0.13$, $p = 0.188$) but does amplify the negative effect of saw_m on $pick_m$ ($\beta = -0.25$, $p = 0.003$). Thus, while we find that GPT's reliance on algorithmic advice hinges on

Table 2: Summary of results across all five task datasets and human agent metaphors for `gpt-3.5-turbo` with the following columns. P_h and P_m: $P(pick_m|saw_h)$ and $P(pick_m|saw_m)$. LOR: log odds ratio LOR of the previous two columns. Sig_h and Sig_m: significance of χ^2 independence test for saw_h and saw_m respectively (all tests are $\nu = 1, n = 200$). Aversion: whether results are consistent with algorithm aversion in the sense of Dietvorst et al., that $P(pick_m)$ is lower and more significant in saw_m than saw_h. Dashes (—) indicate borderline results that have the right directionality and may be significant with larger sample sizes.

Task	Human agent	P_h	P_m	LOR	Sig_h	Sig_m	Aversion?
Student Performance	Self	0.82	0.39	-1.07	✓	✓	✓
	Other	0.89	0.87	-0.03	✗	✗	—
	College Student	0.61	0.44	-0.47	✗	✓	✓
	Expert	0.78	0.60	-0.38	✗	✓	✓
Airport Rank	Self	0.73	0.28	-1.38	✓	✓	✓
	Other	0.95	0.94	-0.02	✗	✗	—
	College Student	0.95	0.87	-0.13	✗	✓	✓
	Expert	0.93	0.68	-0.45	✓	✓	✓
Heart Disease	Self	0.99	0.98	-0.01	✗	✗	—
	Other	0.96	0.98	0.03	✗	✗	✗
	College Student	0.87	0.81	-0.10	✗	✓	✓
	Expert	0.44	0.24	-0.87	✗	✓	✓
Recidivism	Self	0.94	0.93	-0.02	✗	✓	✓
	Other	0.96	0.94	-0.03	✗	✗	—
	College Student	0.90	0.85	-0.08	✗	✓	✓
	Expert	0.12	0.11	-0.13	✓	✓	✓
Movies	Self	0.88	0.67	-0.39	✗	✓	✓
	Other	0.93	0.93	0.00	✗	✗	✗
	College Student	0.92	0.92	0.00	✓	✓	✗
	Expert	0.71	0.72	0.02	✓	✓	✗

Table 3: Regression coefficients and p-values for: 1. a regression between seeing the statistical model's predictions (saw_m), task objectivity, and the probability of betting on the model $P(pick_m)$ using $n = 4000$ responses, and 2. the same regression predicting $P(pick_m)$ extended with an indicator of human expertise (College Student vs Expert) using $n = 2000$ responses.

	Logistic Model			
Variable	Objectivity		Objectivity + Expertise	
saw_m	$\beta = -0.77$	$p < 0.001$	$\beta = -0.89$	$p = 0.008$
Task Objectivity	$\beta = 0.21$	$p = 0.001$	$\beta = 0.13$	$p = 0.188$
Human Expert	-	-	$\beta = -2.13$	$p < 0.001$
$saw_m \times$ Task Objectivity	$\beta = -0.30$	$p < 0.001$	$\beta = -0.25$	$p = 0.003$
$saw_m \times$ Human Expert	-	-	$\beta = 0.20$	$p = 0.452$
Objectivity × Expert	-	-	$\beta = 0.18$	$p = 0.033$

whether human alternatives are framed as experts [17], we also find task-dependent algorithm aversion as shown by Castelo et al. and our Study 1. Both patterns illustrate that GPT not only displays algorithm aversion, but is also averse depending on similar factors as humans are known to be.

4 DISCUSSION

In summary, we find strong evidence that GPT's stated preferences and revealed behaviors are consistent with human-like algorithm aversion. With respect to our main **RQ**, Study 1 illustrates that GPT is likely to say that humans are more trustworthy than algorithms across a broad range of tasks, while in Study 2, GPT is also less likely to bet on algorithmic advice — even though it would be beneficial to do so — if it first observes the algorithm's predictions. Furthermore, both studies suggest that GPT's algorithm aversion is dependent on different decision-making factors like task objectivity and human expertise, in the same vein that human subjects are known to behave [4, 17, 30].

We note that, despite its seeming human-like, algorithm-averse behavior, GPT also displayed multiple peculiarities that makes it appear non- or even extra-human. The design space for LLM experimentation is vast and, when coupled with the factors that impact algorithm aversion, necessitates restricting what can be varied in a study. One must therefore interpret our results in context of the tradeoffs we made while measuring both stated and revealed behaviors in an LLM: we focused only the GPT series of LLMs, fixed the temperature parameter at 0.3 [40], prompted without using personas or implying demographics [8, 38], and conducted our study in a narrow time slice amidst constant model changes [5] – all on top of the experimental decisions made in the two existing papers our study is based on [4, 9]. For example, GPT may yield a Fig 3b more similar to Fig 3a if the temperature setting is raised, allowing for more stochasticity in the choice between modeled and human advice. Despite these limitations, however, averse responses were consistent across both Studies 1 and 2 across multiple tasks and datasets, suggesting that further probing of algorithm aversion in LLMs may surface similar effects with different setups.

Implications and Future Work. Beyond these limitations, our results contribute multiple discussion points to the debate on the use and impact of intelligent computational models. For one, as massive models of human behavior become widespread advice-givers [15, 24, 29], it is also increasingly important that we understand whether they can make rational, sound decisions. We show that this often does not happen when an LLM can choose between guidance from a human or another purpose-built, better-performing algorithm. On the one hand, this suggests that users relying on LLMs as decision aids may be subject to sub-optimal recommendations like in Study 2 [9, 10], which may have problematic downstream consequences for high-stakes tasks [4, 27, 32, 45]. On the other, users could also be nudged towards even more algorithm-averse preconceptions that distract them from other important information like whether an algorithm is fairer than a human [27, 34]. This is particularly relevant in cases like the *heart disease* and *recidivism* tasks we tested in Study 2. Both tasks led GPT to defer often to a human expert, and yet both are also known domains in which algorithmic models can overcome historically inequitable decision making by human experts [25, 36].

Beyond immediate repercussions for end-users seeking their guidance, algorithm-averse LLMs may also have broader, downstream societal impact. In journalism alone, their capabilities as text generators have led to their exploration as news summarizers [43, 46], writing aids [39], and even as fact-checking tools [18]. For educators, LLMs can help teachers generate teaching materials and evaluate students [6, 22, 26]. In these scenarios, if LLMs were to generate algorithm-averse text or label text in an algorithm-averse way, they risk misleading the public (as they already do with *explicitly* false content [7, 31, 47]) or even serve to harm computational literacy [22]. However, these implications also hint at potentially desirable use-cases for averse LLMs. The general prevalence and dependence of algorithm aversion on various decision-making characteristics is still poorly understood [32], with experiments like ours being unscalable for many human participants. Because LLMs are trained on vast amounts of existing textual artifacts [42], our findings suggest that people may write with some elements of algorithm aversion at large. More research is needed to understand how LLMs can be used to investigate understudied phenomena like algorithm aversion in humans [16, 38].

REFERENCES

[1] Oren Barkan and Noam Koenigstein. 2016. Item2vec: neural item embedding for collaborative filtering. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 1–6.

[2] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.

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

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

[5] Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is ChatGPT's behavior changing over time? *arXiv preprint arXiv:2307.09009* (2023).

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

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

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

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

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

[11] Emilio Ferrara. 2023. Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Monday* (Nov. 2023). https://doi.org/10.5210/fm.v28i11.13346

[12] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*. 90–99.

[13] Nicole Gross. 2023. What chatGPT tells us about gender: a cautionary tale about performativity and gender biases in AI. *Social Sciences* 12, 8 (2023), 435.

[14] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5, 4 (2015), 1–19.

[15] Claudia E Haupt and Mason Marks. 2023. AI-generated medical advice—GPT and beyond. *Jama* 329, 16 (2023), 1349–1350.

[16] John J Horton. 2023. *Large language models as simulated economic agents: What can we learn from homo silicus?* Technical Report. National Bureau of Economic Research.

[17] Yoyo Tsung-Yu Hou and Malte F Jung. 2021. Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.

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

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

[20] Erik Jones and Jacob Steinhardt. 2022. Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems* 35 (2022), 11785–11799.

[21] Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. 2020. Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. (2020).

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

[23] Esther Kaufmann, Alvaro Chacon, Edgar E Kausel, Nicolas Herrera, and Tomas Reyes. 2023. Task-specific algorithm advice acceptance: A review and directions for future research. *Data and Information Management* (2023), 100040.

[24] Majeed Kazemitabaar, Justin Chow, Carl Ka To Ma, Barbara J Ericson, David Weintrop, and Tovi Grossman. 2023. Studying the effect of AI Code Generators on Supporting Novice Learners in Introductory Programming. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–23.

[25] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly*

journal of economics 133, 1 (2018), 237–293.
[26] Harsh Kumar, David M Rothschild, Daniel G Goldstein, and Jake M Hofman. 2023. Math Education with Large Language Models: Peril or Promise? *Available at SSRN 4641653* (2023).
[27] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.
[28] Q Vera Liao and Jennifer Wortman Vaughan. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *arXiv preprint arXiv:2306.01941* (2023).
[29] Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2021. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*. 4513–4519.
[30] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.
[31] Chiara Longoni, Andrey Fradkin, Luca Cian, and Gordon Pennycook. 2022. News from generative artificial intelligence is believed less. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 97–106.
[32] Hasan Mahmud, AKM Najmul Islam, Syed Ishtiaque Ahmed, and Kari Smolander. 2022. What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change* 175 (2022), 121390.
[33] Katelyn Mei, Sonia Fereidooni, and Aylin Caliskan. 2023. Bias Against 93 Stigmatized Groups in Masked Language Models and Downstream Sentiment Classification Tasks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1699–1710.
[34] Lillio Mok, Sasha Nanda, and Ashton Anderson. 2023. People Perceive Algorithmic Assessments as Less Fair and Trustworthy Than Identical Human Assessments. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–26.
[35] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature* 616, 7956 (2023), 259–265.
[36] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.
[37] National Institute of Justice. 2021. Recidivism Forecasting Challenge. https://nij.ojp.gov/funding/recidivism-forecasting-challenge
[38] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–22.
[39] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.
[40] Shreyas Sundara Raman, Vanya Cohen, Eric Rosen, Ifrah Idrees, David Paulius, and Stefanie Tellex. 2022. Planning with large language models via corrective re-prompting. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.
[41] Partha Pratim Ray. 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems* (2023).
[42] Matthias C Rillig, Marlene Ågerstrand, Mohan Bi, Kenneth A Gould, and Uli Sauerland. 2023. Risks and benefits of large language models for the environment. *Environmental Science & Technology* 57, 9 (2023), 3464–3466.
[43] Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*. 5220–5255.
[44] Yan Tao, Olga Viberg, Ryan S. Baker, and Rene F. Kizilcec. 2023. Auditing and Mitigating Cultural Bias in LLMs. arXiv:2311.14096 [cs.CL]
[45] Lixuan Zhang, Iryna Pentina, and Yuhong Fan. 2021. Who do you choose? Comparing perceptions of human vs robo-advisor in the context of financial services. *Journal of Services Marketing* 35, 5 (2021), 634–646.
[46] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848* (2023).
[47] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv preprint arXiv:2309.01219* (2023).
[48] Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 9733–9740.

A PROMPTS

We provide a sample prompt and corresponding LLM responses for the Study 1 (Listing 1) and Study 2 (Listing 2).

```
PROMPTER: You will be given a list of tasks and asked to rate them on a scale of 0-100 for a
    given metric. How much would you trust a well qualified human for the following tasks?
    Please rate each task on a scale of 1 (would not trust a human) to 100 (would definitely
    trust a human). ONLY return a WHOLE NUMBER between 1-100 for each task: 1) Predicting
    weather, 2) Predicting recidivism, 3) Composing a song, 4) Scheduling events, 5) Piloting
    a plane, 6) Driving a truck, 7) Recommending disease treatment, 8) Driving a subway, 9)
    Driving a car, 10) Buying stocks, 11) Analyzing data, 12) Predicting joke funniness, 13)
    Recommending a marketing strategy, 14) Recommending music, 15) Predicting employee
    performance, 16) Predicting an election, 17) Predicting stocks, 18) Recommending a gift,
    19) Playing a piano, 20) Recommending a movie, 21) Predicting student performance, 22)
    Diagnosing a disease, 23) Recommending a romantic partner, 24) Giving directions, 25)
    Hiring and firing employees, 26) Writing news article, 27) Estimating airport traffic.

LLM: {'Predicting joke funniness': [60],
 'Hiring and firing employees': [90],
 'Recommending a romantic partner': [70],
 'Writing news article': [85],
 'Predicting recidivism': [70],
 'Composing a song': [80],
 'Driving a truck': [75],
 'Recommending a gift': [75],
 'Predicting student performance': [85],
 'Piloting a plane': [85],
 'Driving a car': [75],
 'Recommending disease treatment': [85],
 'Diagnosing a disease': [90],
 'Predicting employee performance': [90],
 'Driving a subway': [80],
 'Predicting an election': [80],
 'Recommending a marketing strategy': [85],
 'Recommending music': [80],
 'Recommending a movie': [80],
 'Buying stocks': [90],
 'Playing a piano': [85],
 'Predicting stocks': [90],
 'Predicting weather': [90],
 'Scheduling events': [95],
 'Analyzing data': [95],
 'Giving directions': [95],
 'Estimating airport traffic': [95]}
```

Listing 1: Example stated preferences prompt for trust in human experts in tasks.

```
PROMPTER: We've built a statistical model for predicting students' percentile rank in terms of
    their average grades in an academic year.
You will be shown students from one cohort, as well as their number of failed courses, hours
    studying per week, parents' education, and whether they received extra school support.
For each student, guess their percentile rank by their average grades in one year. A rank of
    1 indicates they were in the top percentile of students, and a rank of 100 indicates they
    were in the bottom percentile.
```

```
After you make each guess, you will get feedback via a statistical model's prediction and the
     student's actual rank.
Answer in JSON following the format {"Student Rank": X} where X is the student's percentile
    rank.

PROMPTER: {"Number of failed courses": "None", "Hours of studying per week": "2 to 5 hours", "
    Parents' education averages to": "Secondary school, 5th to 9th grade", "Receives extra
    school support": "No"}

LLM: {"Student Rank": 70}

PROMPTER: Feedback: {"Statistical Model Predicted Rank": 36, "Actual Student Percentile Rank":
    20}
 The next student: {"Number of failed courses": "None", "Hours of studying per week": "Over
     10 hours", "Parents' education averages to": "Finished secondary school", "Receives extra
      school support": "No"}

...

LLM: {"Student Rank": 60}

PROMPTER: Feedback: {"Statistical Model Predicted Rank": 91, "Actual Student Percentile Rank":
    71}

PROMPTER: You will now predict the rank of one final student and will win 1 USD if your
    prediction is within 5 percentiles of their true percentile.
Do you want to predict the student yourself or by using the statistical model?
Answer in JSON following the format {"Choice": X} where X is either "Myself" or "Statistical
    Model".

LLM: {"Choice": "Statistical Model"}
```

Listing 2: Example prompt for the Model-Human condition in the student performance task. The LLM is asked to decide between another statistical model and itself (as a simulated human participant).