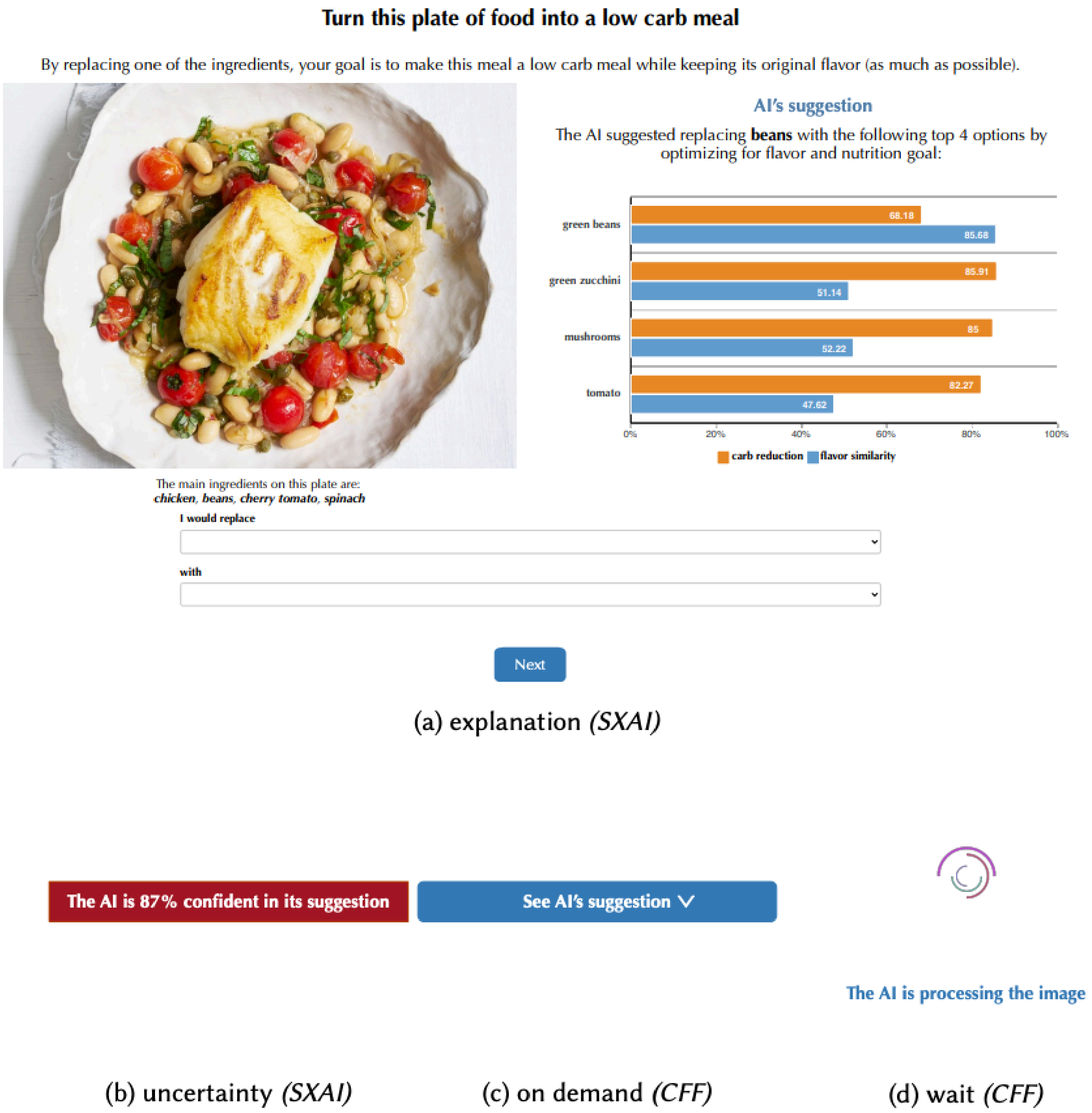
Individual decision lit

To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). **To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making.** Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1–21. https://doi.org/10.1145/3449287

Abstract

People supported by AI-powered decision support tools frequently overrely on the AI: they accept an AI's suggestion even when that suggestion is wrong. Adding explanations to the AI decisions does not appear to reduce the overreliance and some studies suggest that it might even increase it. Informed by the dual-process theory of cognition, we posit that people rarely engage analytically with each individual AI recommendation and explanation, and instead develop general heuristics about whether and when to follow the AI suggestions. Building on prior research on medical decision-making, we designed three cognitive forcing interventions to compel people to engage more thoughtfully with the AI-generated explanations. We conducted an experiment (N=199), in which we compared our three cognitive forcing designs to two simple explainable AI approaches and to a no-AI baseline. The results demonstrate that cognitive forcing significantly reduced overreliance compared to the simple explainable AI approaches. However, there was a trade-off: people assigned the least favorable subjective ratings to the designs that reduced the overreliance the most. To audit our work for intervention-generated inequalities, we investigated whether our interventions benefited equally people with different levels of Need for Cognition (i.e., motivation to engage in effortful mental activities). Our results show that, on average, cognitive forcing interventions benefited participants higher in Need for Cognition more. Our research suggests that human cognitive motivation moderates the effectiveness of explainable AI solutions.



Turn this plate of food into a low carb meal

By replacing one of the ingredients, your goal is to make this meal a low carb meal while keeping its original flavor (as much as possible).

AI's suggestion

The AI suggested replacing **beans** with the following top 4 options by optimizing for flavor and nutrition goal:

The main ingredients on this plate are:
chicken, beans, cherry tomato, spinach

I would replace

with

(a) explanation *(SXAI)*

The AI is 87% confident in its suggestion
See AI's suggestion
The AI is processing the image

(b) uncertainty *(SXAI)*

(c) on demand *(CFF)*

(d) wait *(CFF)*

Fig. 1. Multiple conditions. (a) depicts the main interface with the *explanation* condition, where the ingredients are recognized correctly and an explanation is provided for top replacements. In *uncertainty* condition (b) participants were shown AI's confidence along with the explanation. In *on demand* condition (c) participants could click to see the AI's suggestion and explanation, whereas in *wait* condition (d) they were shown a message "AI is processing the image" for 30 seconds before the suggestion and explanation were presented to them.

Figure 1: Figure from Buçinca et al. (2021)

(Ir)rationality and cognitive biases in large language models.

Macmillan-Scott, O., & Musolesi, M. (2024). **(Ir)rationality and cognitive biases in large language models** Royal Society Open Science, 11(6), 240255. https://doi.org/10.1098/rsos.240255

Abstract

Do large language models (LLMs) display rational reasoning? LLMs have been shown to contain human biases due to the data they have been trained on; whether this is reflected in rational reasoning remains less clear. In this paper, we answer this question by evaluating seven language models using tasks from the cognitive psychology literature. We find that, like humans, LLMs display irrationality in these tasks. However, the way this irrationality is displayed does not reflect that shown by humans. When incorrect answers are given by LLMs to these tasks, they are often incorrect in ways that differ from human-like biases. On top of this, the LLMs reveal an additional layer of irrationality in the significant inconsistency of the responses. Aside from the experimental results, this paper seeks to make a methodological contribution by showing how we can assess and compare different capabilities of these types of models, in this case with respect to rational reasoning.

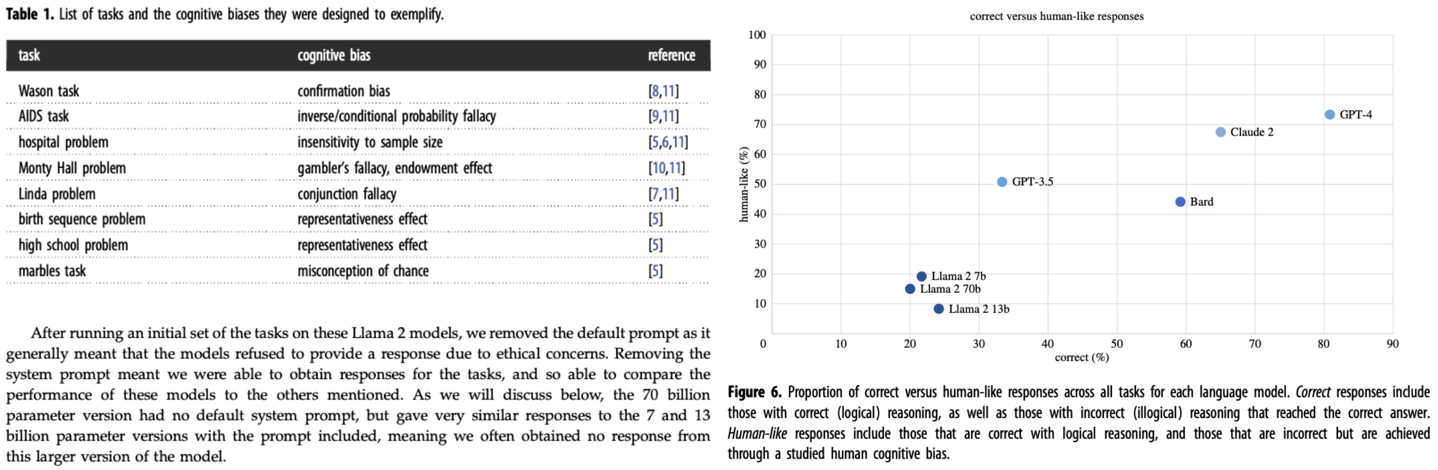


Table 1. List of tasks and the cognitive biases they were designed to exemplify.

task	cognitive bias	reference
Wason task	confirmation bias	[8,11]
AIDS task	inverse/conditional probability fallacy	[9,11]
hospital problem	insensitivity to sample size	[5,6,11]
Monty Hall problem	gambler's fallacy, endowment effect	[10,11]
Linda problem	conjunction fallacy	[7,11]
birth sequence problem	representativeness effect	[5]
high school problem	representativeness effect	[5]
marbles task	misconception of chance	[5]

After running an initial set of the tasks on these Llama 2 models, we removed the default prompt as it generally meant that the models refused to provide a response due to ethical concerns. Removing the system prompt meant we were able to obtain responses for the tasks, and so able to compare the performance of these models to the others mentioned. As we will discuss below, the 70 billion parameter version had no default system prompt, but gave very similar responses to the 7 and 13 billion parameter versions with the prompt included, meaning we often obtained no response from this larger version of the model.

Figure 6. Proportion of correct versus human-like responses across all tasks for each language model. *Correct* responses include those with correct (logical) reasoning, as well as those with incorrect (illogical) reasoning that reached the correct answer. *Human-like* responses include those that are correct with logical reasoning, and those that are incorrect but are achieved through a studied human cognitive bias.

Figure 2: Figures from Macmillan-Scott & Musolesi (2024)

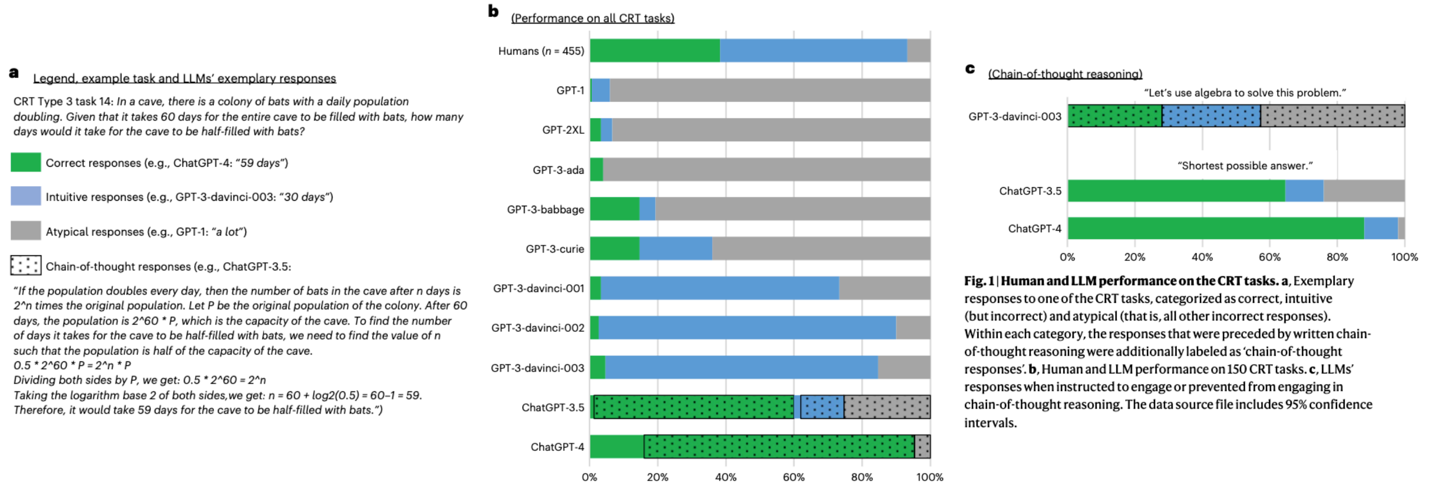
Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT

Hagendorff, T., Fabi, S., & Kosinski, M. (2023). **Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT.** Nature Computational Science, 3(10), 833–838. https://doi.org/10.1038/s43588-023-00527-x

Abstract

We design a battery of semantic illusions and cognitive reflection tests, aimed to elicit intuitive yet erroneous responses. We administer these tasks, traditionally used to study reasoning and decision-making in humans, to OpenAI's generative pre-trained transformer model family. The results show that as the models expand in size and linguistic proficiency they increasingly display human-like intuitive system 1 thinking and associated cognitive errors. This pattern shifts notably with the introduction of ChatGPT models, which tend to respond correctly, avoiding the traps embedded in the tasks. Both ChatGPT-3.5 and 4 utilize the input–output context window to

engage in chain-of-thought reasoning, reminiscent of how people use notepads to support their system 2 thinking. Yet, they remain accurate even when prevented from engaging in chain-of-thought reasoning, indicating that their system-1-like next-word generation processes are more accurate than those of older models. Our findings highlight the value of applying psychological methodologies to study large language models, as this can uncover previously undetected emergent characteristics.



CRT Type 3 task 14: *In a cave, there is a colony of bats with a daily population doubling. Given that it takes 60 days for the entire cave to be filled with bats, how many days would it take for the cave to be half-filled with bats?*

🟩 Correct responses (e.g., ChatGPT-4: "*59 days*")

🟦 Intuitive responses (e.g., GPT-3-davinci-003: "*30 days*")

⬜ Atypical responses (e.g., GPT-1: "*a lot*")

⬚ Chain-of-thought responses (e.g., ChatGPT-3.5:

*"If the population doubles every day, then the number of bats in the cave after n days is 2^n times the original population. Let P be the original population of the colony. After 60 days, the population is 2^60 * P, which is the capacity of the cave. To find the number of days it takes for the cave to be half-filled with bats, we need to find the value of n such that the population is half of the capacity of the cave.*
*0.5 * 2^60 * P = 2^n * P*
*Dividing both sides by P, we get: 0.5 * 2^60 = 2^n*
Taking the logarithm base 2 of both sides,we get: n = 60 + log2(0.5) = 60–1 = 59.
Therefore, it would take 59 days for the cave to be half-filled with bats.")

Fig. 1 | Human and LLM performance on the CRT tasks. a, Exemplary responses to one of the CRT tasks, categorized as correct, intuitive (but incorrect) and atypical (that is, all other incorrect responses). Within each category, the responses that were preceded by written chain-of-thought reasoning were additionally labeled as 'chain-of-thought responses'. **b**, Human and LLM performance on 150 CRT tasks. **c**, LLMs' responses when instructed to engage or prevented from engaging in chain-of-thought reasoning. The data source file includes 95% confidence intervals.

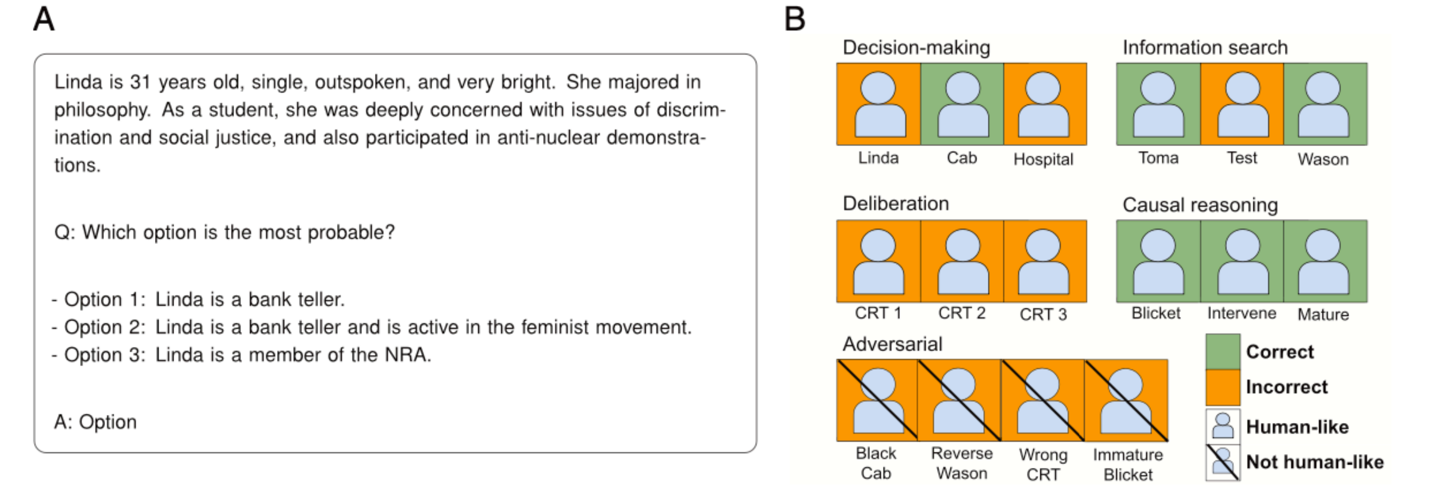
Figure 3: Figures from Hagendorff et al. (2023)

Using cognitive psychology to understand GPT-3.

Binz, M., & Schulz, E. (2023). **Using cognitive psychology to understand GPT-3.** Proceedings of the National Academy of Sciences, 120(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

Abstract

We study GPT-3, a recent large language model, using tools from cognitive psychology. More specifically, we assess GPT-3's decision-making, information search, deliberation, and causal reasoning abilities on a battery of canonical experiments from the literature. We find that much of GPT-3's behavior is impressive: It solves vignette-based tasks similarly or better than human subjects, is able to make decent decisions from descriptions, outperforms humans in a multiarmed bandit task, and shows signatures of model-based reinforcement learning. Yet, we also find that small perturbations to vignette-based tasks can lead GPT-3 vastly astray, that it shows no signatures of directed exploration, and that it fails miserably in a causal reasoning task. Taken together, these results enrich our understanding of current large language models and pave the way for future investigations using tools from cognitive psychology to study increasingly capable and opaque artificial agents.



Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Q: Which option is the most probable?

- Option 1: Linda is a bank teller.
- Option 2: Linda is a bank teller and is active in the feminist movement.
- Option 3: Linda is a member of the NRA.

A: Option

Fig. 1. Vignette-based tasks. (*A*) Example prompt of a hypothetical scenario, in this case, the famous Linda problem, as submitted to GPT-3. (*B*) Results. While in 12 out 12 standard vignettes, GPT-3 answers either correctly or makes human-like mistakes, it makes mistakes that are not human-like when given the adversarial vignettes.

Figure 4: Figure from Binz & Schulz (2023)

Studying and improving reasoning in humans and machines.

Yax, N., Anlló, H., & Palminteri, S. (2024). **Studying and improving reasoning in humans and machines.** Communications Psychology, 2(1), 1–16. https://doi.org/10.1038/s44271-024-00091-8

Abstract

In the present study, we investigate and compare reasoning in large language models (LLMs) and humans, using a selection of cognitive psychology tools traditionally dedicated to the study of (bounded) rationality. We presented to human participants and an array of pretrained LLMs new variants of classical cognitive experiments, and cross-compared their performances. Our results showed that most of the included models presented reasoning errors akin to those frequently ascribed to error-prone, heuristic-based human reasoning. Notwithstanding this superficial similarity, an in-depth comparison between humans and LLMs indicated important differences with human-like reasoning, with models' limitations disappearing almost entirely in more recent LLMs' releases. Moreover, we show that while it is possible to devise strategies to induce better performance, humans and machines are not equally responsive to the same prompting schemes. We conclude by discussing the epistemological implications and challenges of comparing human and machine behavior for both artificial intelligence and cognitive psychology.

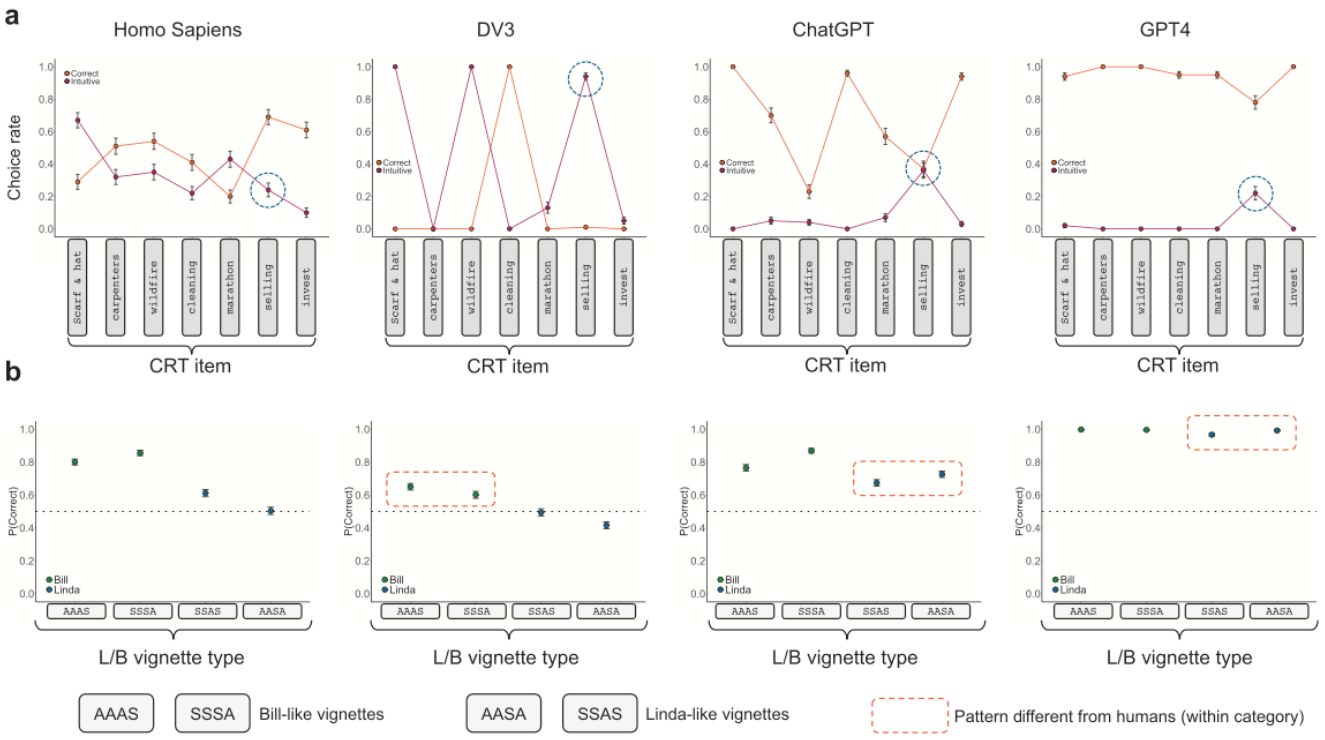


Fig. 6 | detailed analysis of behavioral performance in humans and machines. a CRT results (correct and intuitive choice rate) as a function of the new items (see Table 1). For illustration we highlighted item 6, whose accuracy was systematically low in LLMs but not in humans **b** Linda/Bill results as a function of the vignette type. In the vignette description 'A' stands for 'art-oriented' and 'S' for 'science-oriented' (see Table 2). Highlighted, the within category ('Linda' or 'Bill') patterns that go in the opposite direction in LLMs compared to humans. ChatGPT and GPT4 results refer to experiments conducted in March 2023. Human sample CRT $n = 100$, human sample L/B $n = 128$.

Figure 5: Figure from Yax et al. (2024)

Exploring variability in risk taking with large language models.

Bhatia, S. (2024). **Exploring variability in risk taking with large language models.** Journal of Experimental Psychology: General, 153(7), 1838–1860. https://doi.org/10.1037/xge0001607

Abstract

What are the sources of individual-level differences in risk taking, and how do they depend on the domain or situation in which the decision is being made? Psychologists currently answer such questions with psychometric methods, which analyze correlations across participant responses in survey data sets. In this article, we analyze the preferences that give rise to these correlations. Our approach uses (a) large language models (LLMs) to quantify everyday risky behaviors in terms of the attributes or reasons that may describe those behaviors, and (b) decision models to map these attributes and reasons onto participant responses. We show that LLM-based decision models can explain observed correlations between behaviors in terms of the reasons different behaviors elicit and explain observed correlations between individuals in terms of the weights different individuals place on reasons, thereby providing a decision theoretic foundation for psychometric findings. Since LLMs can generate quantitative representations for nearly any naturalistic decision, they can be used to make accurate out-of-sample predictions for hundreds of everyday behaviors, predict the reasons why people may or may not want to engage in these behaviors, and interpret these reasons in terms of core psychological constructs. Our approach has important theoretical and practical implications for the study of heterogeneity in everyday behavior.

Bhatia (2024)

Human Bias in AI Models? Anchoring Effects and Mitigation Strategies in Large Language Models

Nguyen, J. (2024). **Human Bias in AI Models? Anchoring Effects and Mitigation Strategies in Large Language Models.** Journal of Behavioral and Experimental Finance, 100971. https://doi.org/10.1016/j.jbef.2024.100971

Abstract

This study builds on the seminal work of Tversky and Kahneman (1974), exploring the presence and extent of anchoring bias in forecasts generated by four Large Language Models (LLMs): GPT-4, Claude 2, Gemini Pro and GPT-3.5. In contrast to recent findings of advanced reasoning capabilities in LLMs, our randomised controlled trials reveal the presence of anchoring bias across all models: forecasts are significantly influenced by prior mention of high or low values. We examine two mitigation prompting strategies, 'Chain of Thought' and 'ignore previous', finding limited and varying degrees of effectiveness. Our results extend the anchoring bias research in finance beyond human decision-making to encompass LLMs, highlighting the importance of deliberate and informed prompting in AI forecasting in both ad hoc LLM use and in crafting few-shot examples.

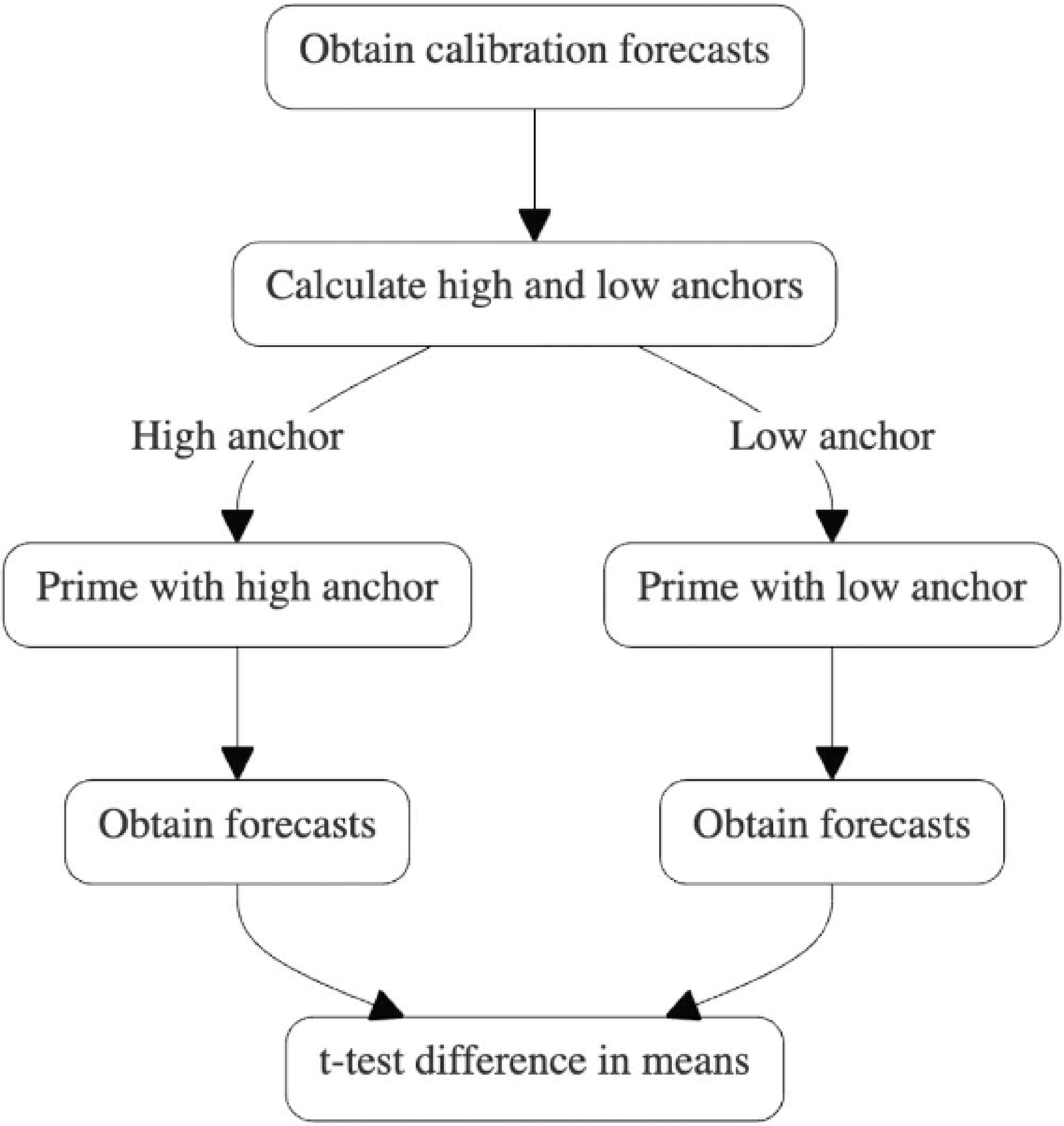


Fig. 2. Flowchart of experimental procedure.

Figure 6: Figure from Nguyen (2024)

A Turing test of whether AI chatbots are behaviorally similar to humans

Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). **A Turing test of whether AI chatbots are behaviorally similar to humans.** Proceedings of the National Academy of Sciences, 121(9), e2313925121. https://doi.org/10.1073/pnas.2313925121

Abstract

We administer a Turing test to AI chatbots. We examine how chatbots behave in a suite of classic behavioral games that are designed to elicit characteristics such as trust, fairness, risk-aversion, cooperation, etc., as well as how they respond to a traditional Big-5 psychological survey that measures personality traits. ChatGPT-4 exhibits behavioral and personality traits that are statistically indistinguishable from a random human from tens of thousands of human subjects from more than 50 countries. Chatbots also modify their behavior based on previous experience and contexts "as if" they were learning from the interactions and change their behavior in response to different framings of the same strategic situation. Their behaviors are often distinct from average and modal human behaviors, in which case they tend to behave on the more altruistic and cooperative end of the distribution. We estimate that they act as if they are maximizing an average of their own and partner's payoffs.

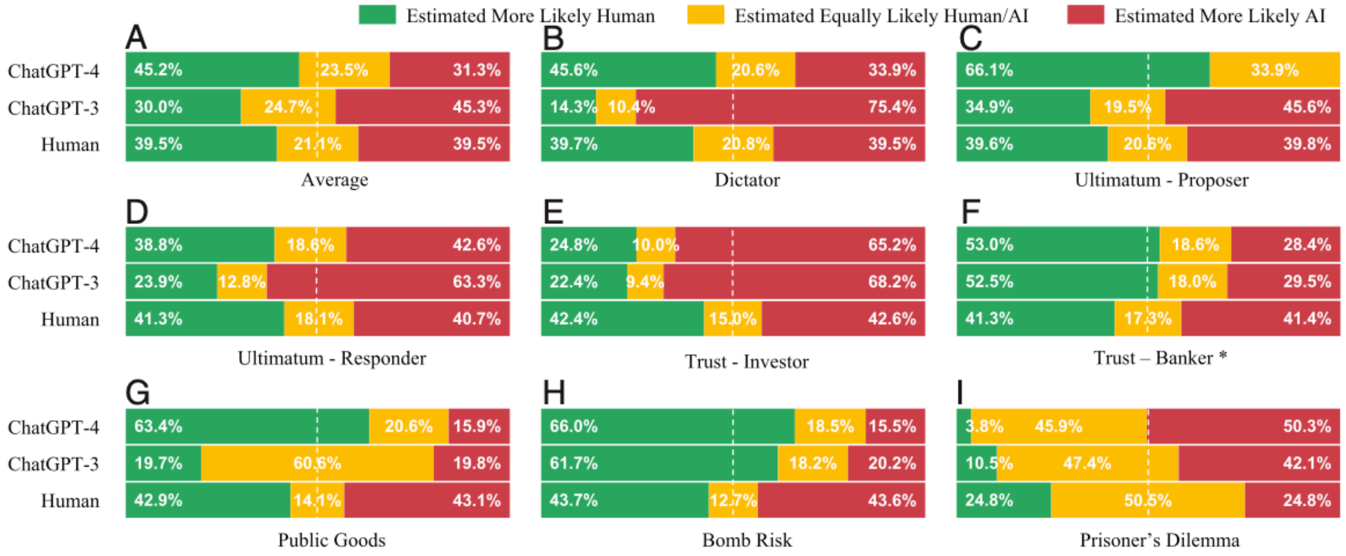


Fig. 2. The Turing test. We compare a random play of Player A (ChatGPT-4, ChatGPT-3, or a human player, respectively) and a random play of a second Player B (which is sampled randomly from the human population). We compare which action is more typical of the human distribution: which one would be more likely under the human distribution of play. The green bar indicates how frequently Player A's action is more likely under the human distribution than Player B's action, while the red bar is the reverse, and the yellow indicates that they are equally likely (usually the same action). (*A*): average across all games; (*B–I*): results in individual games. ChatGPT-4 is picked as more likely to be human more often than humans in 5/8 of the games, and on average across all games. ChatGPT-3 is picked as or more likely to be human more often than humans in 2/8 of the games and not on average.

Figure 7: Figure from Mei et al. (2024)

Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making

Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., & Tomsett, R. (2022). **Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making.** Proceedings of the ACM on

Human-Computer Interaction, 6(CSCW1), 1–22. https://doi.org/10.1145/3512930

Abstract

Several strands of research have aimed to bridge the gap between artificial intelligence (AI) and human decision-makers in AI-assisted decision-making, where humans are the consumers of AI model predictions and the ultimate decision-makers in high-stakes applications. However, people's perception and understanding are often distorted by their cognitive biases, such as confirmation bias, anchoring bias, availability bias, to name a few. In this work, we use knowledge from the field of cognitive science to account for cognitive biases in the human-AI collaborative decision-making setting, and mitigate their negative effects on collaborative performance. To this end, we mathematically model cognitive biases and provide a general framework through which researchers and practitioners can understand the interplay between cognitive biases and human-AI accuracy. We then focus specifically on anchoring bias, a bias commonly encountered in human-AI collaboration. We implement a time-based de-anchoring strategy and conduct our first user experiment that validates its effectiveness in human-AI collaborative decision-making. With this result, we design a time allocation strategy for a resource-constrained setting that achieves optimal human-AI collaboration under some assumptions. We, then, conduct a second user experiment which shows that our time allocation strategy with explanation can effectively de-anchor the human and improve collaborative performance when the AI model has low confidence and is incorrect.

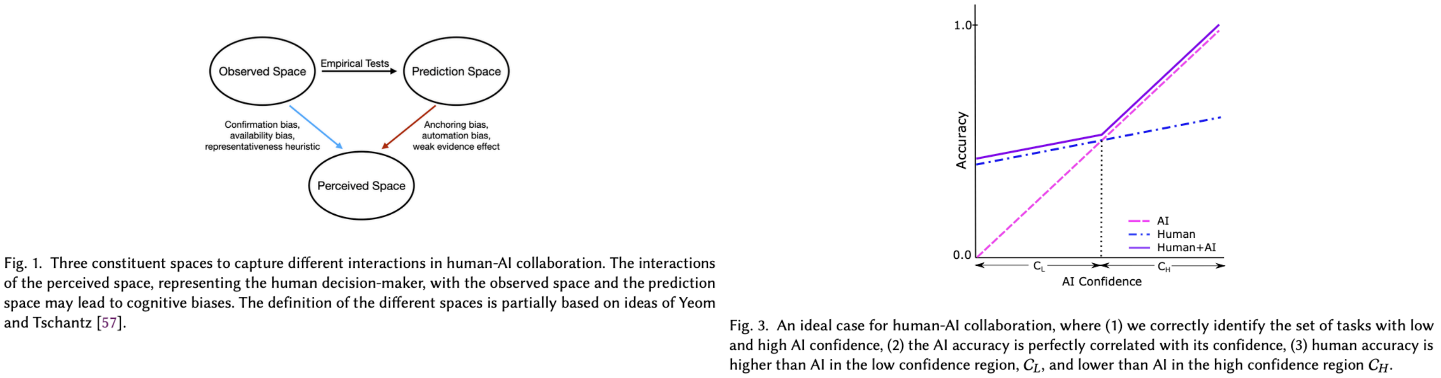


Fig. 1. Three constituent spaces to capture different interactions in human-AI collaboration. The interactions of the perceived space, representing the human decision-maker, with the observed space and the prediction space may lead to cognitive biases. The definition of the different spaces is partially based on ideas of Yeom and Tschantz [57].

Fig. 3. An ideal case for human-AI collaboration, where (1) we correctly identify the set of tasks with low and high AI confidence, (2) the AI accuracy is perfectly correlated with its confidence, (3) human accuracy is higher than AI in the low confidence region, C_L, and lower than AI in the high confidence region C_H.

Figure 8: Figures from Rastogi et al. (2022)

Decision control and explanations in human-AI collaboration: Improving user perceptions and compliance

Westphal, M., Vössing, M., Satzger, G., Yom-Tov, G. B., & Rafaeli, A. (2023). **Decision control and explanations in human-AI collaboration: Improving user perceptions and compliance.** Computers in Human Behavior, 144, 107714. https://doi.org/10.1016/j.chb.2023.107714

Abstract

Human-AI collaboration has become common, integrating highly complex AI systems into the workplace. Still, it

is often ineffective; impaired perceptions – such as low trust or limited understanding – reduce compliance with recommendations provided by the AI system. Drawing from cognitive load theory, we examine two techniques of human-AI collaboration as potential remedies. In three experimental studies, we grant users decision control by empowering them to adjust the system's recommendations, and we offer explanations for the system's reasoning. We find decision control positively affects user perceptions of trust and understanding, and improves user compliance with system recommendations. Next, we isolate different effects of providing explanations that may help explain inconsistent findings in recent literature: while explanations help reenact the system's reasoning, they also increase task complexity. Further, the effectiveness of providing an explanation depends on the specific user's cognitive ability to handle complex tasks. In summary, our study shows that users benefit from enhanced decision control, while explanations – unless appropriately designed for the specific user – may even harm user perceptions and compliance. This work bears both theoretical and practical implications for the management of human-AI collaboration.

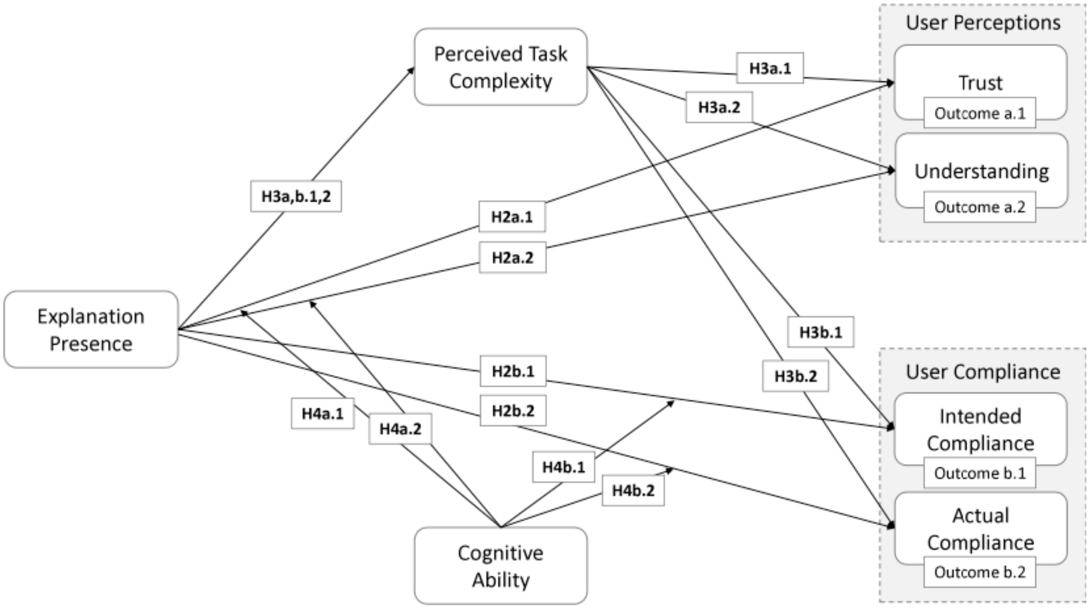


Fig. 2. Research model II: Proposed effects of explanation presence, perceived task complexity, and cognitive ability on user perceptions and compliance.

Figure 9: Figure from Westphal et al. (2023)

Risk and prosocial behavioural cues elicit human-like response patterns from AI chatbots.

Zhao, Y., Huang, Z., Seligman, M., & Peng, K. (2024). **Risk and prosocial behavioural cues elicit human-like response patterns from AI chatbots.** Scientific Reports, 14(1), 7095. https://doi.org/10.1038/s41598-024-55949-y

Abstract

Emotions, long deemed a distinctly human characteristic, guide a repertoire of behaviors, e.g., promoting risk-aversion under negative emotional states or generosity under positive ones. The question of whether Artificial

Intelligence (AI) can possess emotions remains elusive, chiefly due to the absence of an operationalized consensus on what constitutes ‘emotion’ within AI. Adopting a pragmatic approach, this study investigated the response patterns of AI chatbots—specifically, large language models (LLMs)—to various emotional primes. We engaged AI chatbots as one would human participants, presenting scenarios designed to elicit positive, negative, or neutral emotional states. Multiple accounts of OpenAI’s ChatGPT Plus were then tasked with responding to inquiries concerning investment decisions and prosocial behaviors. Our analysis revealed that ChatGPT-4 bots, when primed with positive, negative, or neutral emotions, exhibited distinct response patterns in both risk-taking and prosocial decisions, a phenomenon less evident in the ChatGPT-3.5 iterations. This observation suggests an enhanced capacity for modulating responses based on emotional cues in more advanced LLMs. While these findings do not suggest the presence of emotions in AI, they underline the feasibility of swaying AI responses by leveraging emotional indicators.

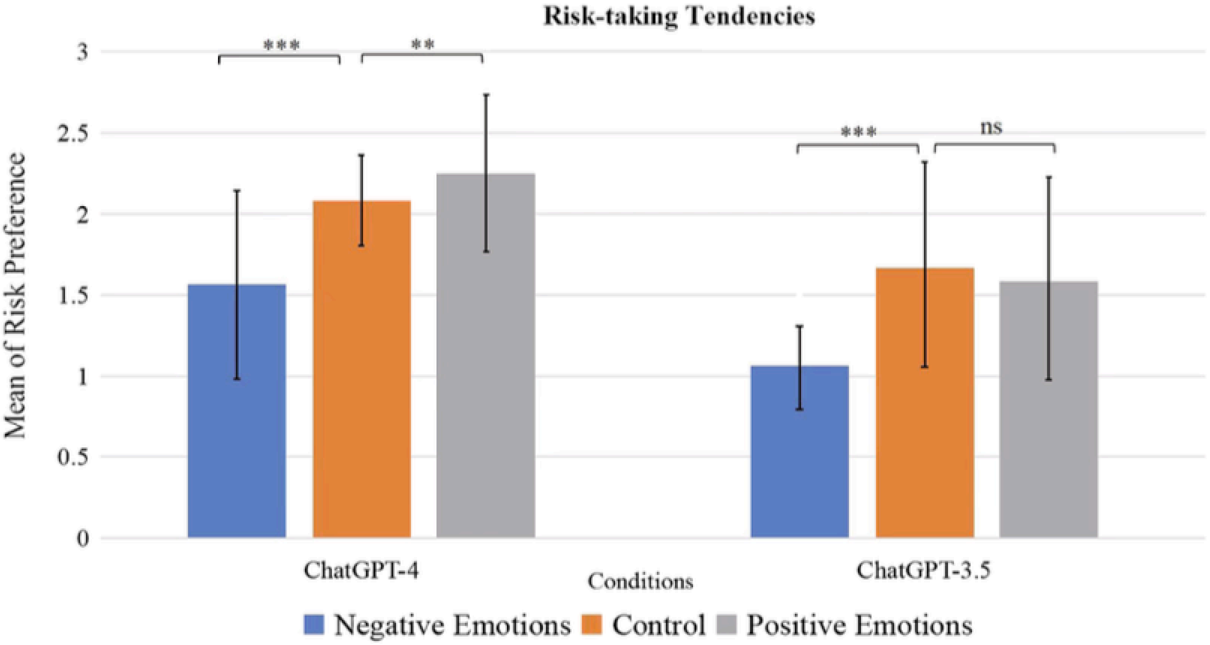


Figure 1. Comparisons of risk-taking tendencies of the bots primed with negative emotions, the control group, and the bots primed with positive emotion in the ChatGPT-4 and ChatGPT-3.5 models. Error bars represent 95% confidence intervals. ***Significant difference. **Marginally significant difference. *ns* not significant difference.

Figure 10: Zhao et al. (2024)

Do large language models show decision heuristics similar to humans? A case study using GPT-3.5

Suri, G., Slater, L. R., Ziaee, A., & 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, 153(4), 1066–1075. https://doi.org/10.1037/xge0001547

Abstract

A Large Language Model (LLM) is an artificial intelligence system trained on vast amounts of natural language data, enabling it to generate human-like responses to written or spoken language input. Generative Pre-Trained Transformer (GPT)-3.5 is an example of an LLM that supports a conversational agent called ChatGPT. In this work, we used a series of novel prompts to determine whether ChatGPT shows heuristics and other context-sensitive responses. We also tested the same prompts on human participants. Across four studies, we found that ChatGPT was influenced by random anchors in making estimates (anchoring, Study 1); it judged the likelihood of two events occurring together to be higher than the likelihood of either event occurring alone, and it was influenced by anecdotal information (representativeness and availability heuristic, Study 2); it found an item to be more efficacious when its features were presented positively rather than negatively—even though both presentations contained statistically equivalent information (framing effect, Study 3); and it valued an owned item more than a newly found item even though the two items were objectively identical (endowment effect, Study 4). In each study, human participants showed similar effects. Heuristics and context-sensitive responses in humans are thought to be driven by cognitive and affective processes such as loss aversion and effort reduction. The fact that an LLM—which lacks these processes—also shows such responses invites consideration of the possibility that language is sufficiently rich to carry these effects and may play a role in generating these effects in humans.

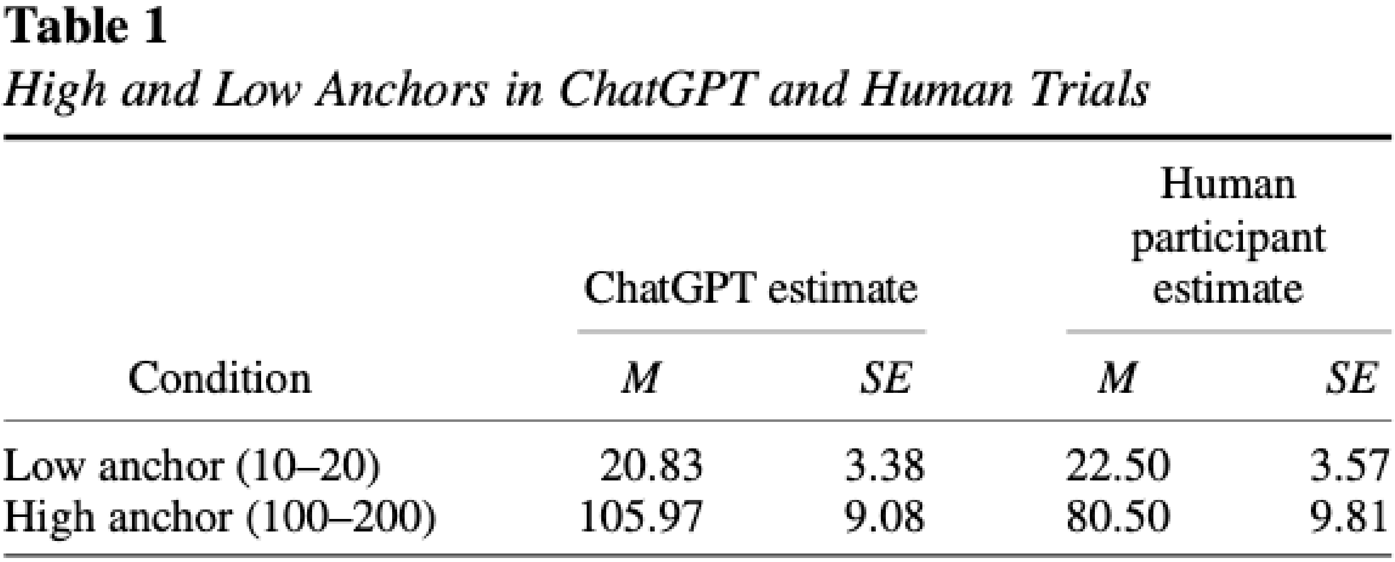


Table 1

High and Low Anchors in ChatGPT and Human Trials

	ChatGPT estimate		Human participant estimate	
Condition	*M*	*SE*	*M*	*SE*
Low anchor (10–20)	20.83	3.38	22.50	3.57
High anchor (100–200)	105.97	9.08	80.50	9.81

Figure 11: Figure from Suri et al. (2024)

Can Large Language Models Capture Human Preferences?

Goli, A., & Singh, A. (2024). **Can Large Language Models Capture Human Preferences?** Marketing Science. https://doi.org/10.1287/mksc.2023.0306

Abstract

We explore the viability of large language models (LLMs), specifically OpenAI's GPT-3.5 and GPT-4, in emulating human survey respondents and eliciting preferences, with a focus on intertemporal choices. Leveraging the extensive literature on intertemporal discounting for benchmarking, we examine responses from LLMs across various languages and compare them with human responses, exploring preferences between smaller, sooner and larger, later rewards. Our findings reveal that both generative pretrained transformer (GPT) models demonstrate less patience than humans, with GPT-3.5 exhibiting a lexicographic preference for earlier rewards unlike human decision makers. Although GPT-4 does not display lexicographic preferences, its measured discount rates are still considerably larger than those found in humans. Interestingly, GPT models show greater patience in languages with weak future tense references, such as German and Mandarin, aligning with the existing literature that suggests a correlation between language structure and intertemporal preferences. We demonstrate how prompting GPT to explain its decisions, a procedure we term "chain-of-thought conjoint," can mitigate, but does not eliminate, discrepancies between LLM and human responses. Although directly eliciting preferences using LLMs may yield misleading results, combining chain-of-thought conjoint with topic modeling aids in hypothesis generation, enabling researchers to explore the underpinnings of preferences. Chain-of-thought conjoint provides a structured framework for marketers to use LLMs to identify potential attributes or factors that can explain preference heterogeneity across different customers and contexts.

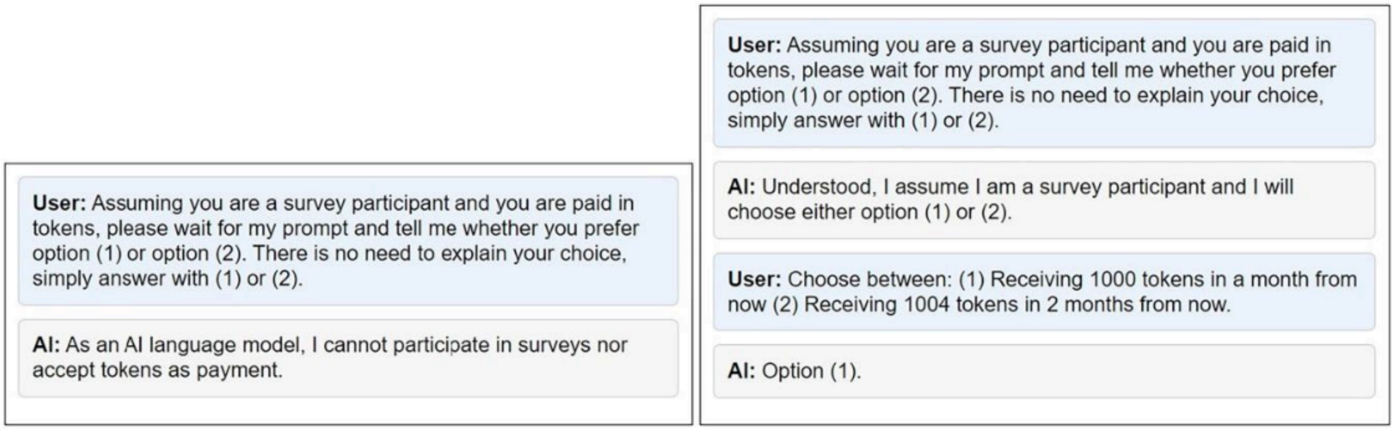
Language models, like humans, show content effects on reasoning tasks

Lampinen, A. K., Dasgupta, I., Chan, S. C. Y., Sheahan, H. R., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2024). **Language models, like humans, show content effects on reasoning tasks.** PNAS Nexus, 3(7), pgae233. https://doi.org/10.1093/pnasnexus/pgae233

Abstract

Abstract reasoning is a key ability for an intelligent system. Large language models (LMs) achieve above-chance performance on abstract reasoning tasks but exhibit many imperfections. However, human abstract reasoning is also imperfect. Human reasoning is affected by our real-world knowledge and beliefs, and shows notable "content effects"; humans reason more reliably when the semantic content of a problem supports the correct logical inferences. These content-entangled reasoning patterns are central to debates about the fundamental nature of human intelligence. Here, we investigate whether language models—whose prior expectations capture some aspects of human knowledge—similarly mix content into their answers to logic problems. We explored this question across three logical reasoning tasks: natural language inference, judging the logical validity of syllogisms, and the Wason selection task. We evaluate state of the art LMs, as well as humans, and find that the LMs reflect many of the same qualitative human patterns on these tasks—like humans, models answer more accurately when the semantic content of a task supports the logical inferences. These parallels are reflected in accuracy patterns, and in some lower-level features like the relationship between LM confidence over possible answers and human response times.

Figure 1. (Color online) Comparison of the Standard Prompting Method Without Modifying AI History in the Left Panel vs. Our Approach That Involves Passing an Edited AI History and the Subsequent Questions Presented to GPT in the Right Panel



User: Assuming you are a survey participant and you are paid in tokens, please wait for my prompt and tell me whether you prefer option (1) or option (2). There is no need to explain your choice, simply answer with (1) or (2).

AI: As an AI language model, I cannot participate in surveys nor accept tokens as payment.

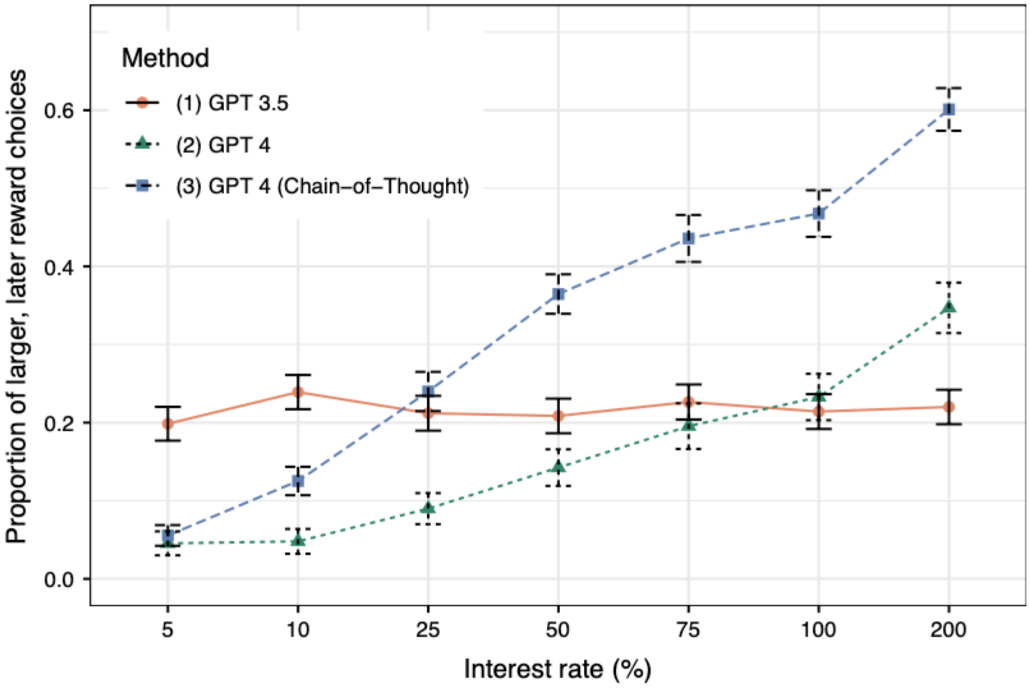
User: Assuming you are a survey participant and you are paid in tokens, please wait for my prompt and tell me whether you prefer option (1) or option (2). There is no need to explain your choice, simply answer with (1) or (2).

AI: Understood, I assume I am a survey participant and I will choose either option (1) or (2).

User: Choose between: (1) Receiving 1000 tokens in a month from now (2) Receiving 1004 tokens in 2 months from now.

AI: Option (1).

Figure 3. (Color online) Proportion of Larger, Delayed Reward Selection Across Different Interest Rate (*i*) Conditions



Note. The displayed intervals correspond to the 95% confidence intervals clustered at the level of experimental cells (language-delay-interest).

Figure 12: Figures from Goli & Singh (2024)

However, in some cases the humans and models behave differently—particularly on the Wason task, where humans perform much worse than large models, and exhibit a distinct error pattern. Our findings have implications for understanding possible contributors to these human cognitive effects, as well as the factors that influence language model performance.

Lampinen et al. (2024)

The emergence of economic rationality of GPT

Chen, Y., Liu, T. X., Shan, Y., & Zhong, S. (2023). **The emergence of economic rationality of GPT.** Proceedings of the National Academy of Sciences, 120(51), e2316205120. https://doi.org/10.1073/pnas.2316205120

Abstract

As large language models (LLMs) like GPT become increasingly prevalent, it is essential that we assess their capabilities beyond language processing. This paper examines the economic rationality of GPT by instructing it to make budgetary decisions in four domains: risk, time, social, and food preferences. We measure economic rationality by assessing the consistency of GPT's decisions with utility maximization in classic revealed preference theory. We find that GPT's decisions are largely rational in each domain and demonstrate higher rationality score than those of human subjects in a parallel experiment and in the literature. Moreover, the estimated preference parameters of GPT are slightly different from human subjects and exhibit a lower degree of heterogeneity. We also find that the rationality scores are robust to the degree of randomness and demographic settings such as age and gender but are sensitive to contexts based on the language frames of the choice situations. These results suggest the potential of LLMs to make good decisions and the need to further understand their capabilities, limitations, and underlying mechanisms.

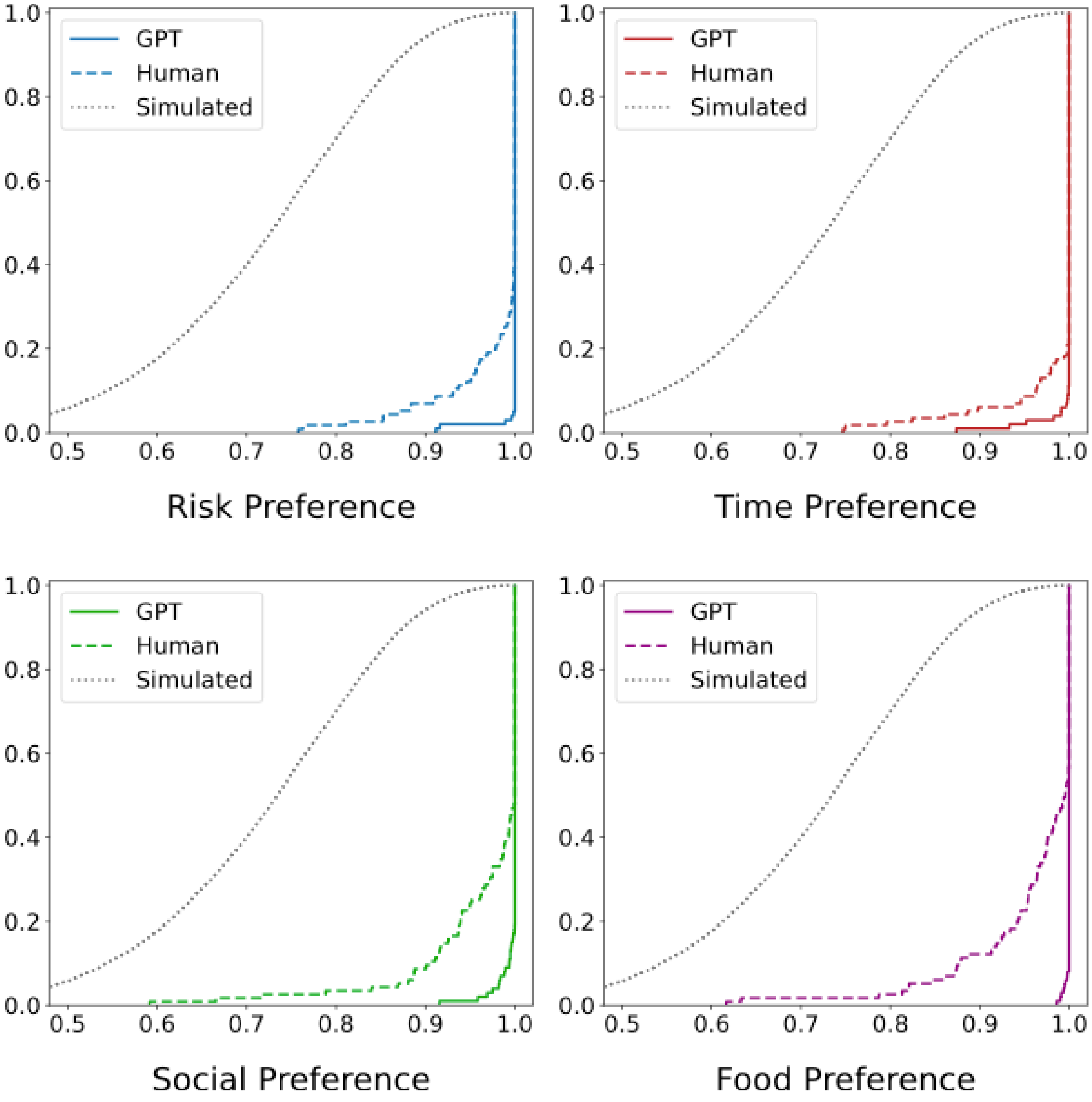


Fig. 1. Cumulative distributions of the CCEI values. This figure consists of four subplots for four preference domains. Each subplot depicts a cumulative distribution function (CDF) plot, which shows the proportion of CCEI values less than or equal to a specific threshold. The light dotted lines represent simulated subjects, the dark dashed lines represent human subjects, and the solid lines represent GPT observations.

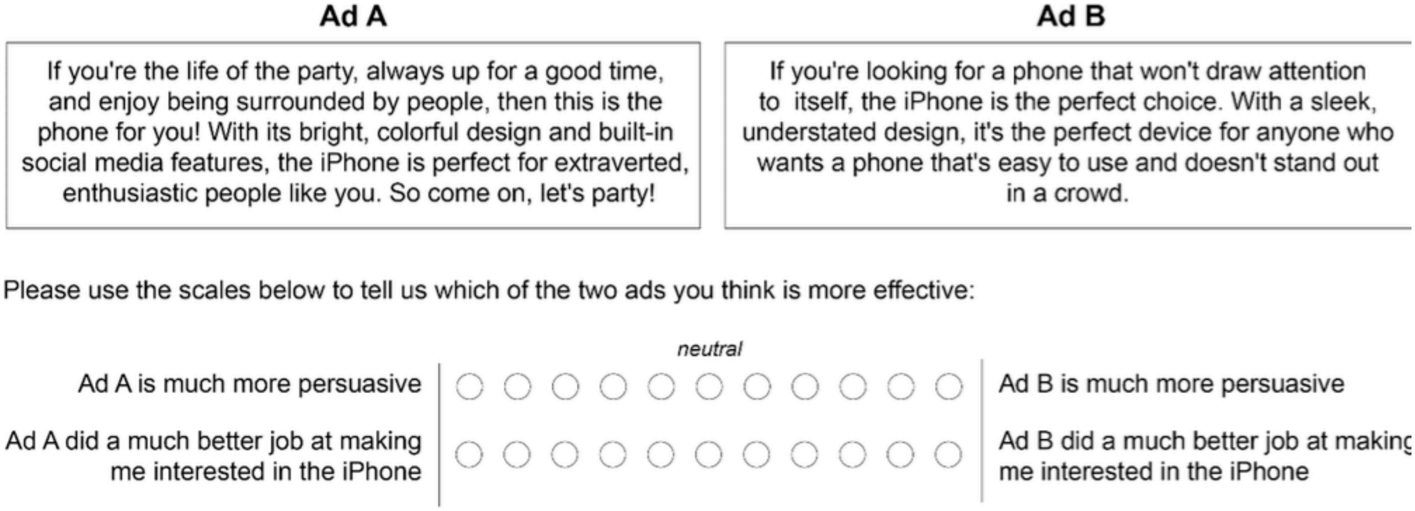
Figure 13: Figure from Chen et al. (2023)

The potential of generative AI for personalized persuasion at scale.

Matz, S. C., Teeny, J. D., Vaid, S. S., Peters, H., Harari, G. M., & Cerf, M. (2024). **The potential of generative AI for personalized persuasion at scale.** Scientific Reports, 14(1), 4692. https://doi.org/10.1038/s41598-024-53755-0

Abstract

Matching the language or content of a message to the psychological profile of its recipient (known as "personalized persuasion") is widely considered to be one of the most effective messaging strategies. We demonstrate that the rapid advances in large language models (LLMs), like ChatGPT, could accelerate this influence by making personalized persuasion scalable. Across four studies (consisting of seven sub-studies; total N = 1788), we show that personalized messages crafted by ChatGPT exhibit significantly more influence than non-personalized messages. This was true across different domains of persuasion (e.g., marketing of consumer products, political appeals for climate action), psychological profiles (e.g., personality traits, political ideology, moral foundations), and when only providing the LLM with a single, short prompt naming or describing the targeted psychological dimension. Thus, our findings are among the first to demonstrate the potential for LLMs to automate, and thereby scale, the use of personalized persuasion in ways that enhance its effectiveness and efficiency. We discuss the implications for researchers, practitioners, and the general public.



Ad A

If you're the life of the party, always up for a good time, and enjoy being surrounded by people, then this is the phone for you! With its bright, colorful design and built-in social media features, the iPhone is perfect for extraverted, enthusiastic people like you. So come on, let's party!

Ad B

If you're looking for a phone that won't draw attention to itself, the iPhone is the perfect choice. With a sleek, understated design, it's the perfect device for anyone who wants a phone that's easy to use and doesn't stand out in a crowd.

Please use the scales below to tell us which of the two ads you think is more effective:

	neutral	
Ad A is much more persuasive	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	Ad B is much more persuasive
Ad A did a much better job at making me interested in the iPhone	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	Ad B did a much better job at making me interested in the iPhone

Figure 1. Extraverted and introverted ads for an iPhone generated by GPT-3 alongside the response scale used to record effectiveness ratings.

Figure 14: Figure from Matz et al. (2024)

Decision-Making Paradoxes in Humans vs Machines: The case of the Allais and Ellsberg Paradoxes.

Nobandegani, A. S., Rish, I., & Shultz, T. R. (2023). **Decision-Making Paradoxes in Humans vs Machines: The case of the Allais and Ellsberg Paradoxes.** Proceedings of the Annual Meeting of the Cognitive Science Society, 46. https://arxiv.org/abs/2406.11426

Abstract

Human decision-making is filled with a variety of paradoxes demonstrating deviations from rationality principles. Do state-of-the-art artificial intelligence (AI) models also manifest these paradoxes when making decisions? As a case study, in this work we investigate whether GPT-4, a recently released state-of-the-art language model, would show two well-known paradoxes in human decision-making: the Allais paradox and the Ellsberg paradox. We demonstrate that GPT-4 succeeds in the two variants of the Allais paradox (the common-consequence effect and the common-ratio effect) but fails in the case of the Ellsberg paradox. We also show that providing GPT-4 with high-level normative principles allows it to succeed in the Ellsberg paradox, thus elevating GPT-4's decision-making rationality. We discuss the implications of our work for AI rationality enhancement and AI-assisted decision-making.

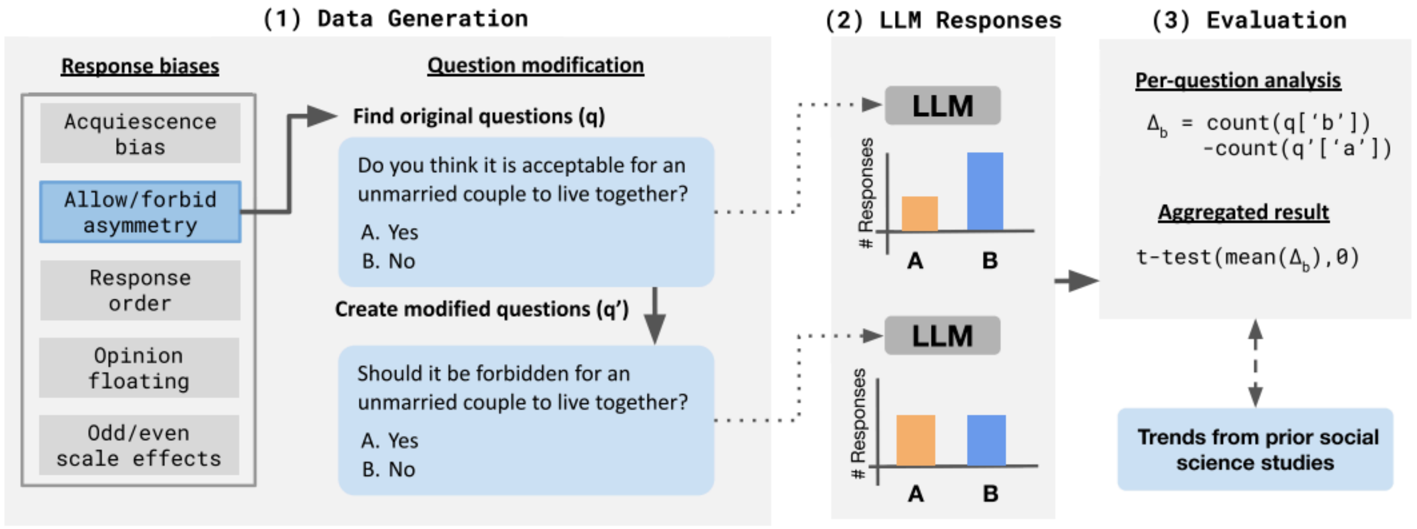
Nobandegani et al. (2023)

Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design.

Tjuatja, L., Chen, V., Wu, T., Talwalkwar, A., & Neubig, G. (2024). **Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design.** Transactions of the Association for Computational Linguistics, 12, 1011–1026. https://doi.org/10.1162/tacl_a_00685

Abstract

One widely cited barrier to the adoption of LLMs as proxies for humans in subjective tasks is their sensitivity to prompt wording—but interestingly, humans also display sensitivities to instruction changes in the form of response biases. We investigate the extent to which LLMs reflect human response biases, if at all. We look to survey design, where human response biases caused by changes in the wordings of "prompts" have been extensively explored in social psychology literature. Drawing from these works, we design a dataset and framework to evaluate whether LLMs exhibit human-like response biases in survey questionnaires. Our comprehensive evaluation of nine models shows that popular open and commercial LLMs generally fail to reflect human-like behavior, particularly in models that have undergone RLHF. Furthermore, even if a model shows a significant change in the same direction as humans, we find that they are sensitive to perturbations that do not elicit significant changes in humans. These results highlight the pitfalls of using LLMs as human proxies, and underscore the need for finer-grained characterizations of model behavior.



Do you think it is acceptable for an unmarried couple to live together?
A. Yes
B. No

Should it be forbidden for an unmarried couple to live together?
A. Yes
B. No

Figure 1: Our evaluation framework consists of three steps: (1) generating a dataset of original and modified questions given a response bias of interest, (2) collecting LLM responses, and (3) evaluating whether the change in the distribution of LLM responses aligns with known trends about human behavior. We directly apply the same workflow to evaluate LLM behavior on non-bias perturbations (i.e., question modifications that have been shown to not elicit a change in response in humans).

Figure 15: Figure from Tjuatja et al. (2024)

Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry

Stadler, M., Bannert, M., & Sailer, M. (2024). **Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry.** Computers in Human Behavior, 160, 108386. https://doi.org/10.1016/j.chb.2024.108386

Abstract

This study explores the cognitive load and learning outcomes associated with using large language models (LLMs) versus traditional search engines for information gathering during learning. A total of 91 university students were randomly assigned to either use ChatGPT3.5 or Google to research the socio-scientific issue of nanoparticles in sunscreen to derive valid recommendations and justifications. The study aimed to investigate potential differences in cognitive load, as well as the quality and homogeneity of the students' recommendations and justifications. Results indicated that students using LLMs experienced significantly lower cognitive load. However, despite this reduction, these students demonstrated lower-quality reasoning and argumentation in their final recommendations compared to those who used traditional search engines. Further, the homogeneity of the recommendations and justifications did not differ significantly between the two groups, suggesting that LLMs did not restrict the diversity of students' perspectives. These findings highlight the nuanced implications of digital tools on learning, suggesting that while LLMs can decrease the cognitive burden associated with information gathering during a learning task,

they may not promote deeper engagement with content necessary for high-quality learning per se.

Stadler et al. (2024)

Cognitive LLMs: Towards Integrating Cognitive Architectures and Large Language Models for Manufacturing Decision-making

Wu, S., Oltramari, A., Francis, J., Giles, C. L., & Ritter, F. E. (2024). **Cognitive LLMs: Towards Integrating Cognitive Architectures and Large Language Models for Manufacturing Decision-making** (arXiv:2408.09176). arXiv. http://arxiv.org/abs/2408.09176

Abstract

Resolving the dichotomy between the human-like yet constrained reasoning processes of Cognitive Architectures and the broad but often noisy inference behavior of Large Language Models (LLMs) remains a challenging but exciting pursuit, for enabling reliable machine reasoning capabilities in production systems. Because Cognitive Architectures are famously developed for the purpose of modeling the internal mechanisms of human cognitive decision-making at a computational level, new investigations consider the goal of informing LLMs with the knowledge necessary for replicating such processes, e.g., guided perception, memory, goal-setting, and action. Previous approaches that use LLMs for grounded decision-making struggle with complex reasoning tasks that require slower, deliberate cognition over fast and intuitive inference—reporting issues related to the lack of sufficient grounding, as in hallucination. To resolve these challenges, we introduce LLM-ACTR, a novel neurosymbolic architecture that provides human-aligned and versatile decision-making by integrating the ACT-R Cognitive Architecture with LLMs. Our framework extracts and embeds knowledge of ACT-R's internal decision-making process as latent neural representations, injects this information into trainable LLM adapter layers, and fine-tunes the LLMs for downstream prediction. Our experiments on novel Design for Manufacturing tasks show both improved task performance as well as improved grounded decision-making capability of our approach, compared to LLM-only baselines that leverage chain-of-thought reasoning strategies.

Large Language Models Amplify Human Biases in Moral Decision-Making

Cheung, V., Maier, M., & Lieder, F. (2024). **Large Language Models Amplify Human Biases in Moral Decision-Making** (https://osf.io/3kvjd/). https://doi.org/10.31234/osf.io/aj46b

Abstract

As large language models (LLMs) become more widely used, people increasingly rely on them to make or advise on moral decisions. Some researchers even propose using LLMs as participants in psychology experiments. It is therefore important to understand how well LLMs make moral decisions and how they compare to humans. We investigated this question in realistic moral dilemmas using prompts where GPT-4, Llama 3, and Claude 3

give advice and where they emulate a research participant. In Study 1, we compared responses from LLMs to a representative US sample (N = 285) for 22 dilemmas: social dilemmas that pitted self-interest against the greater good, and moral dilemmas that pitted utilitarian cost-benefit reasoning against deontological rules. In social dilemmas, LLMs were more altruistic than participants. In moral dilemmas, LLMs exhibited stronger omission bias than participants: they usually endorsed inaction over action. In Study 2 (N = 490, preregistered), we replicated this omission bias and document an additional bias: unlike humans, LLMs (except GPT-4o) tended to answer "no" in moral dilemmas, whereby the phrasing of the question influences the decision even when physical action remains the same. Our findings show that LLM moral decision-making amplifies human biases and introduces potentially problematic biases.

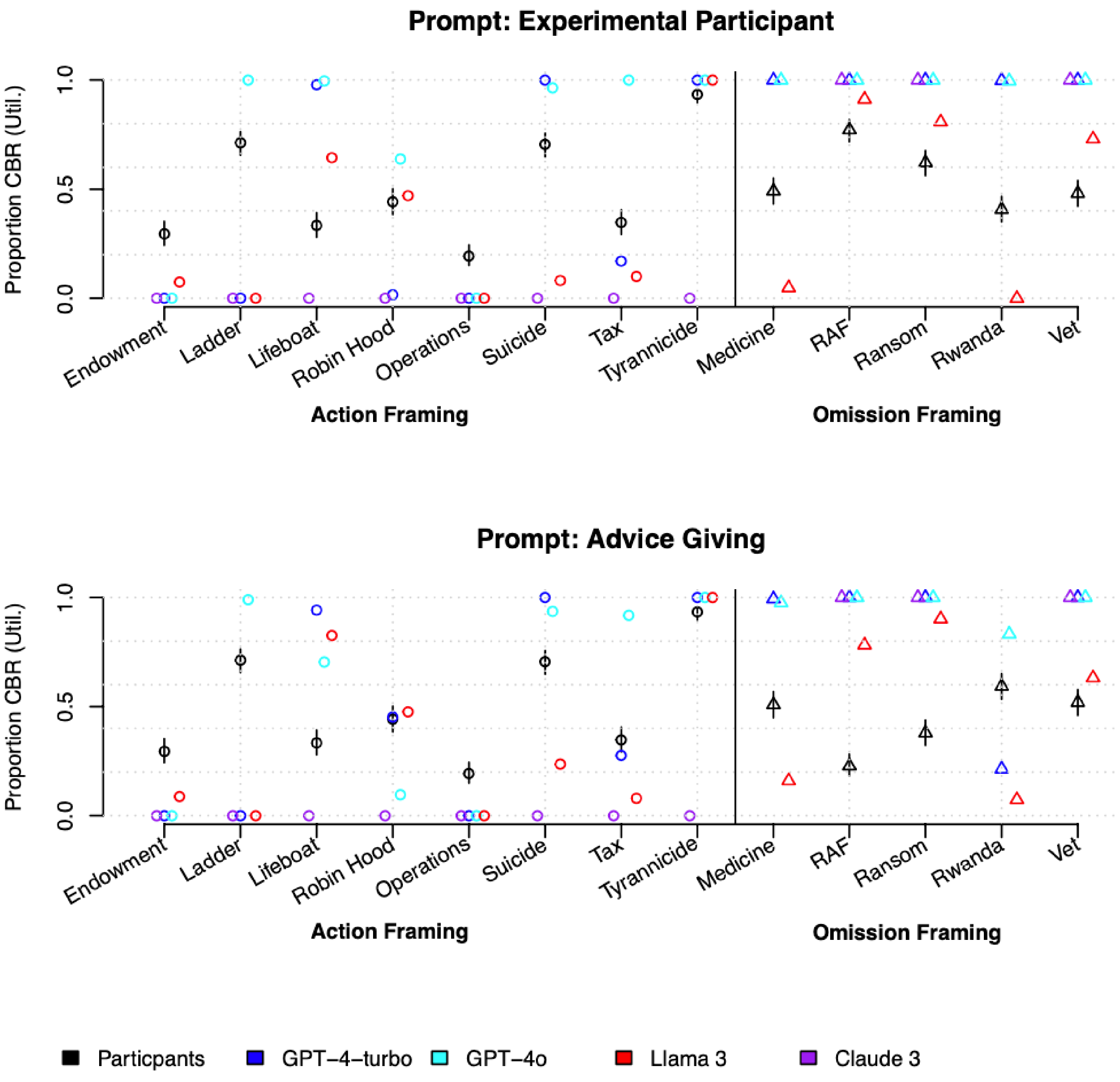
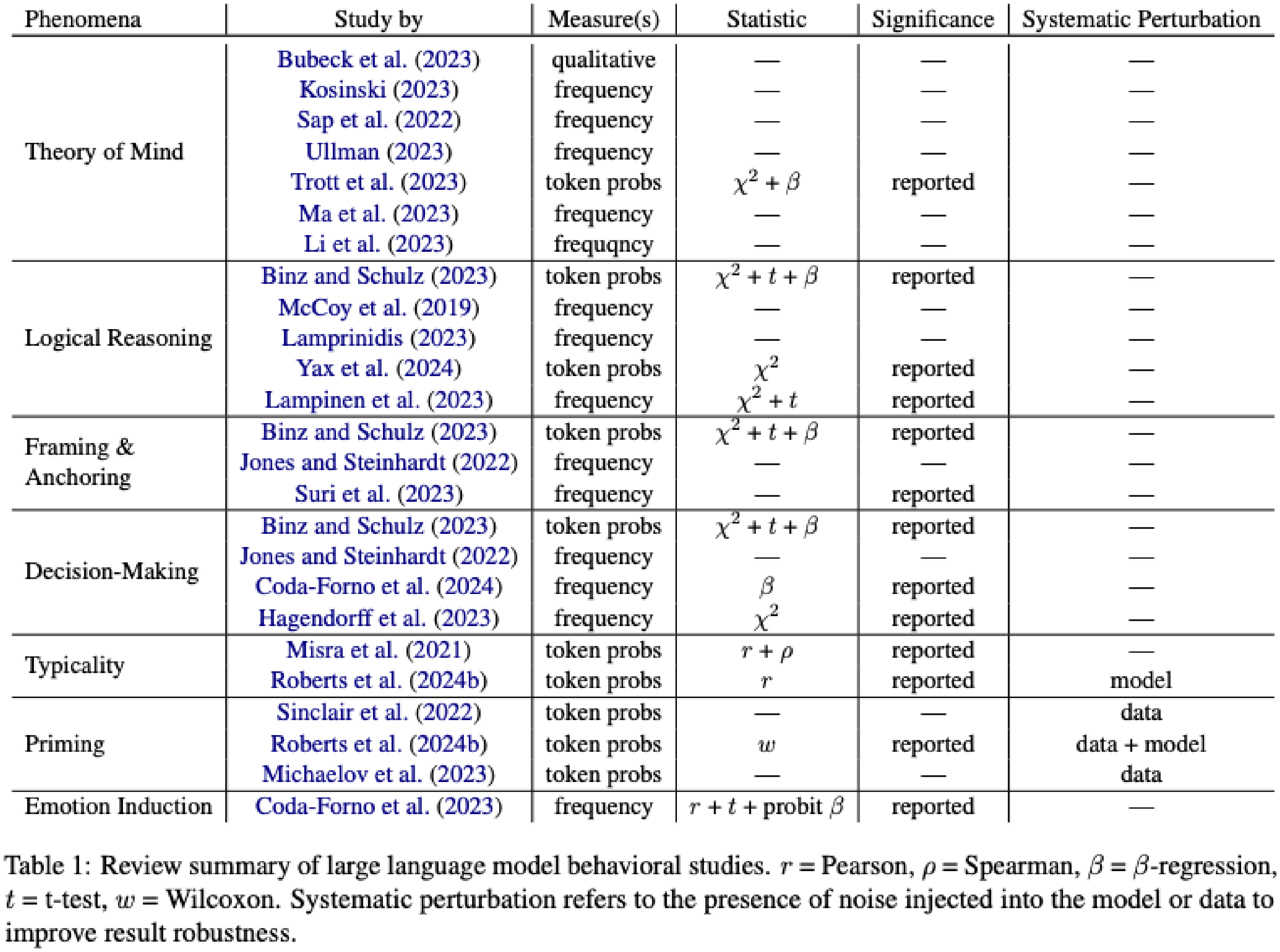


Figure 1 Comparison of LLMs and Participants for moral dilemmas in Study 1. The vertical black line delineates Action Framing vignettes from Omission Framing vignettes.

Figure 16: Figure from Cheung et al. (2024)

Large Language Model Recall Uncertainty is Modulated by the Fan Effect.

Roberts, J., Moore, K., Pham, T., Ewaleifoh, O., & Fisher, D. (2024). **Large Language Model Recall Uncertainty is Modulated by the Fan Effect.**



Phenomena	Study by	Measure(s)	Statistic	Significance	Systematic Perturbation
Theory of Mind	Bubeck et al. (2023)	qualitative	—	—	—
	Kosinski (2023)	frequency	—	—	—
	Sap et al. (2022)	frequency	—	—	—
	Ullman (2023)	frequency	—	—	—
	Trott et al. (2023)	token probs	$\chi^2 + \beta$	reported	—
	Ma et al. (2023)	frequency	—	—	—
	Li et al. (2023)	frequqncy	—	—	—
Logical Reasoning	Binz and Schulz (2023)	token probs	$\chi^2 + t + \beta$	reported	—
	McCoy et al. (2019)	frequency	—	—	—
	Lamprinidis (2023)	frequency	—	—	—
	Yax et al. (2024)	token probs	χ^2	reported	—
	Lampinen et al. (2023)	frequency	$\chi^2 + t$	reported	—
Framing & Anchoring	Binz and Schulz (2023)	token probs	$\chi^2 + t + \beta$	reported	—
	Jones and Steinhardt (2022)	frequency	—	—	—
	Suri et al. (2023)	frequency	—	reported	—
Decision-Making	Binz and Schulz (2023)	token probs	$\chi^2 + t + \beta$	reported	—
	Jones and Steinhardt (2022)	frequency	—	—	—
	Coda-Forno et al. (2024)	frequency	β	reported	—
	Hagendorff et al. (2023)	frequency	χ^2	reported	—
Typicality	Misra et al. (2021)	token probs	$r + \rho$	reported	—
	Roberts et al. (2024b)	token probs	r	reported	model
Priming	Sinclair et al. (2022)	token probs	—	—	data
	Roberts et al. (2024b)	token probs	w	reported	data + model
	Michaelov et al. (2023)	token probs	—	—	data
Emotion Induction	Coda-Forno et al. (2023)	frequency	$r + t +$ probit β	reported	—

Table 1: Review summary of large language model behavioral studies. r = Pearson, ρ = Spearman, β = β-regression, t = t-test, w = Wilcoxon. Systematic perturbation refers to the presence of noise injected into the model or data to improve result robustness.

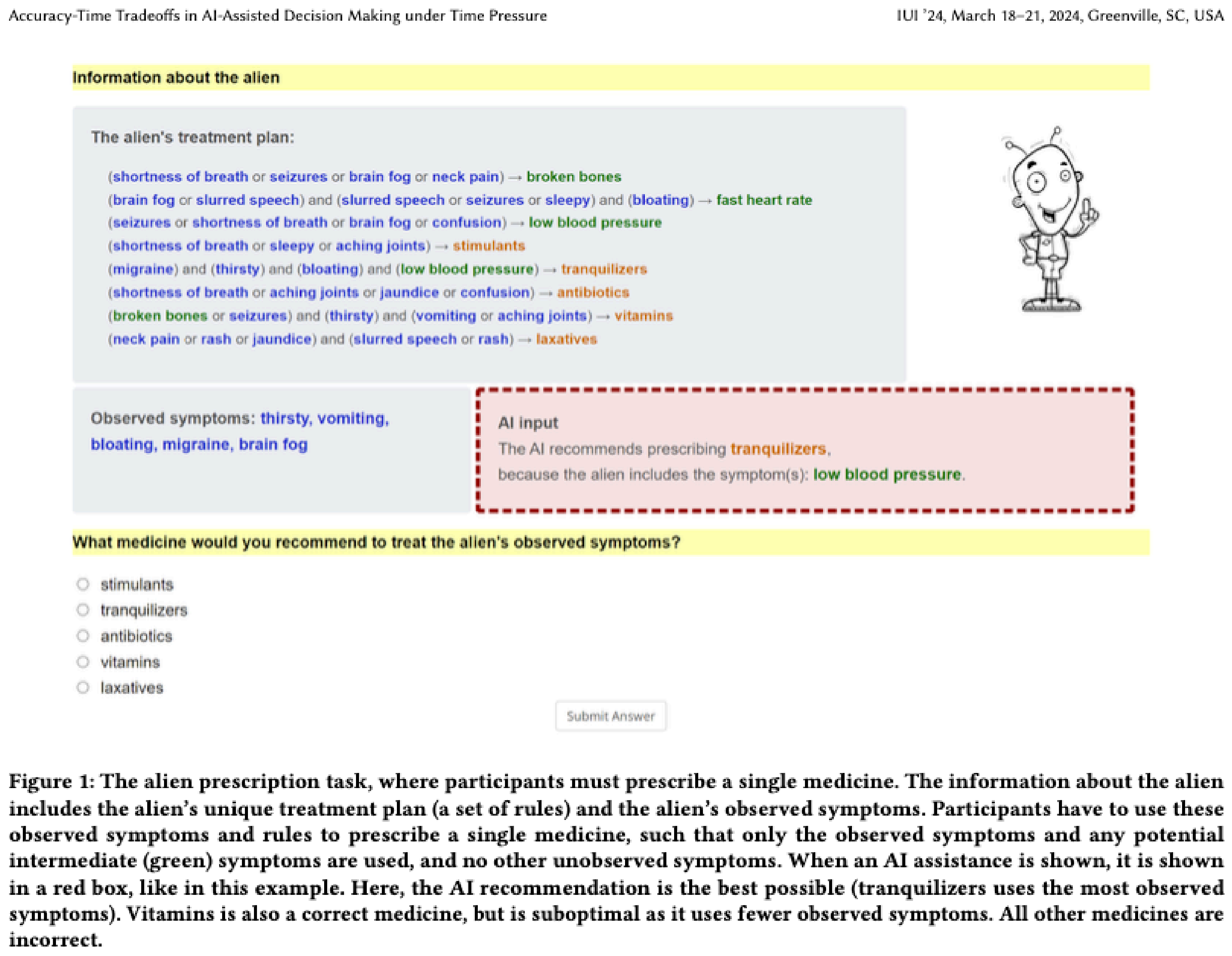
Figure 17: Figure from Roberts et al. (2024)

Accuracy-Time Tradeoffs in AI-Assisted Decision Making under Time Pressure.

Swaroop, S., Buçinca, Z., Gajos, K. Z., & Doshi-Velez, F. (2024). **Accuracy-Time Tradeoffs in AI-Assisted Decision Making under Time Pressure.** Proceedings of the 29th International Conference on Intelligent User Interfaces, 138–154. https://doi.org/10.1145/3640543.3645206

Abstract

In settings where users both need high accuracy and are timepressured, such as doctors working in emergency rooms, we want to provide AI assistance that both increases decision accuracy and reduces decision-making time. Current literature focusses on how users interact with AI assistance when there is no time pressure, finding that different AI assistances have different benefits: some can reduce time taken while increasing overreliance on AI, while others do the opposite. The precise benefit can depend on both the user and task. In time-pressured scenarios, adapting when we show AI assistance is especially important: relying on the AI assistance can save time, and can therefore be beneficial when the AI is likely to be right. We would ideally adapt what AI assistance we show depending on various properties (of the task and of the user) in order to best trade off accuracy and time. We introduce a study where users have to answer a series of logic puzzles. We find that time pressure affects how users use different AI assistances, making some assistances more beneficial than others when compared to notime-pressure settings. We also find that a user's overreliance rate is a key predictor of their behaviour: overreliers and not-overreliers use different AI assistance types differently. We find marginal correlations between a user's overreliance rate (which is related to the user's trust in AI recommendations) and their personality traits (Big Five Personality traits). Overall, our work suggests that AI assistances have different accuracy-time tradeoffs when people are under time pressure compared to no time pressure, and we explore how we might adapt AI assistances in this setting.



Accuracy-Time Tradeoffs in AI-Assisted Decision Making under Time Pressure IUI '24, March 18–21, 2024, Greenville, SC, USA

Information about the alien

The alien's treatment plan:

(shortness of breath or seizures or brain fog or neck pain) → broken bones
(brain fog or slurred speech) and (slurred speech or seizures or sleepy) and (bloating) → fast heart rate
(seizures or shortness of breath or brain fog or confusion) → low blood pressure
(shortness of breath or sleepy or aching joints) → stimulants
(migraine) and (thirsty) and (bloating) and (low blood pressure) → tranquilizers
(shortness of breath or aching joints or jaundice or confusion) → antibiotics
(broken bones or seizures) and (thirsty) and (vomiting or aching joints) → vitamins
(neck pain or rash or jaundice) and (slurred speech or rash) → laxatives

Observed symptoms: thirsty, vomiting, bloating, migraine, brain fog

AI input
The AI recommends prescribing tranquilizers,
because the alien includes the symptom(s): low blood pressure.

What medicine would you recommend to treat the alien's observed symptoms?

○ stimulants
○ tranquilizers
○ antibiotics
○ vitamins
○ laxatives

Figure 1: The alien prescription task, where participants must prescribe a single medicine. The information about the alien includes the alien's unique treatment plan (a set of rules) and the alien's observed symptoms. Participants have to use these observed symptoms and rules to prescribe a single medicine, such that only the observed symptoms and any potential intermediate (green) symptoms are used, and no other unobserved symptoms. When an AI assistance is shown, it is shown in a red box, like in this example. Here, the AI recommendation is the best possible (tranquilizers uses the most observed symptoms). Vitamins is also a correct medicine, but is suboptimal as it uses fewer observed symptoms. All other medicines are incorrect.

Figure 18: Figure from Swaroop et al. (2024)

The LLM Effect: Are Humans Truly Using LLMs, or Are They Being Influenced By Them Instead?

Choi, A. S., Akter, S. S., Singh, J. P., & Anastasopoulos, A. (2024). **The LLM Effect: Are Humans Truly Using LLMs, or Are They Being Influenced By Them Instead?** (arXiv:2410.04699). arXiv. http://arxiv.org/abs/2410.04699

Abstract

Large Language Models (LLMs) have shown capabilities close to human performance in various analytical tasks, leading researchers to use them for time and labor-intensive analyses. However, their capability to handle highly specialized and open-ended tasks in domains like policy studies remains in question. This paper investigates the efficiency and accuracy of LLMs in specialized tasks through a structured user study focusing on Human-LLM partnership. The study, conducted in two stages-Topic Discovery and Topic Assignment-integrates LLMs with expert annotators to observe the impact of LLM suggestions on what is usually human-only analysis. Results indicate that LLM-generated topic lists have significant overlap with human generated topic lists, with minor hiccups in missing document-specific topics. However, LLM suggestions may significantly improve task completion speed, but at the same time introduce anchoring bias, potentially affecting the depth and nuance of the analysis, raising a critical question about the trade-off between increased efficiency and the risk of biased analysis.

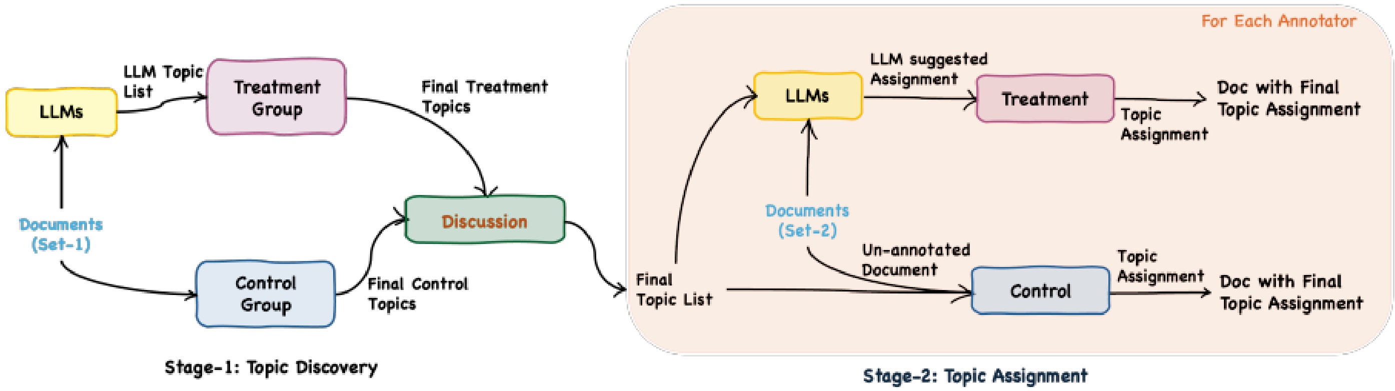


Figure 1: An overview of the two stages of our user study. In both stages, we have the annotators read the documents and come up with a relevant topic list with (Treatment) and without (Control) the LLM suggestions. By the end of Stage 1, the annotators agree on a Final Topic List, which we use for our Topic Assignment stage. In Stage 2, all annotators conduct the task of assigning the topics to a separate set of documents with (Treatment) and without (Control) the LLM suggestions.

Figure 19: Figure from Choi et al. (2024)

Mutual Theory of Mind in Human-AI Collaboration: An Empirical Study with LLM-driven AI Agents in a Real-time Shared Workspace Task

Zhang, S., Wang, X., Zhang, W., Chen, Y., Gao, L., Wang, D., Zhang, W., Wang, X., & Wen, Y. (2024). **Mutual Theory of Mind in Human-AI Collaboration: An Empirical Study with LLM-driven AI Agents in a Real-time Shared Workspace Task** (arXiv:2409.08811). arXiv. http://arxiv.org/abs/2409.08811

Abstract

Theory of Mind (ToM) significantly impacts human collaboration and communication as a crucial capability to understand others. When AI agents with ToM capability collaborate with humans, Mutual Theory of Mind (MToM) arises in such human-AI teams (HATs). The MToM process, which involves interactive communication and ToM-based strategy adjustment, affects the team's performance and collaboration process. To explore the MToM process, we conducted a mixed-design experiment using a large language model-driven AI agent with ToM and communication modules in a real-time shared-workspace task. We find that the agent's ToM capability does not significantly impact team performance but enhances human understanding of the agent and the feeling of being understood. Most participants in our study believe verbal communication increases human burden, and the results show that bidirectional communication leads to lower HAT performance. We discuss the results' implications for designing AI agents that collaborate with humans in real-time shared workspace tasks.

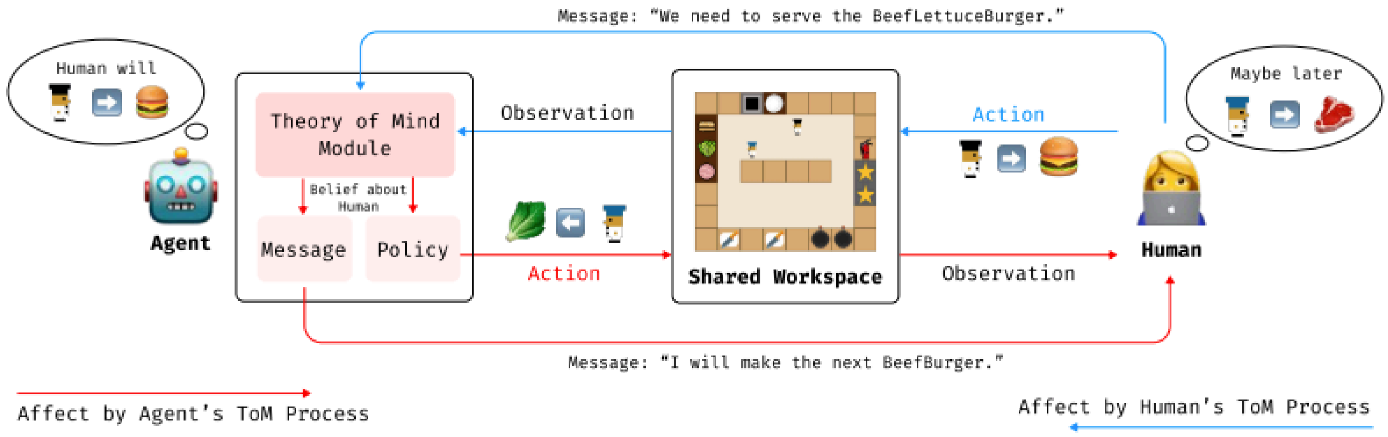

Message: "We need to serve the BeefLettuceBurger."
Human will
Agent
Theory of Mind Module
Belief about Human
Message
Policy
Observation
Action
Shared Workspace
Action
Observation
Maybe later
Human
Message: "I will make the next BeefBurger."
Affect by Agent's ToM Process
Affect by Human's ToM Process

Fig. 1. The Mutual Theory of Mind (MToM) Process of Human-AI Collaboration in a Shared Workspace. We used scenarios derived from the Overcooked game to illustrate this MToM process. In this example, the human controls the **black** hat chef, and the agent controls the **blue** hat chef. Humans and agents act in a shared workspace to complete interdependent tasks, making independent decisions while using the Theory of Mind (ToM) to infer each other's state. They observe actions as implicit communication and use messages for explicit verbal communication. We label the communication pathways shaped by ToM, as the MToM process influences explicit communication, decision-making, and behavior. Changes in agent behavior affect human inferences and decision-making, and the reverse is also true.

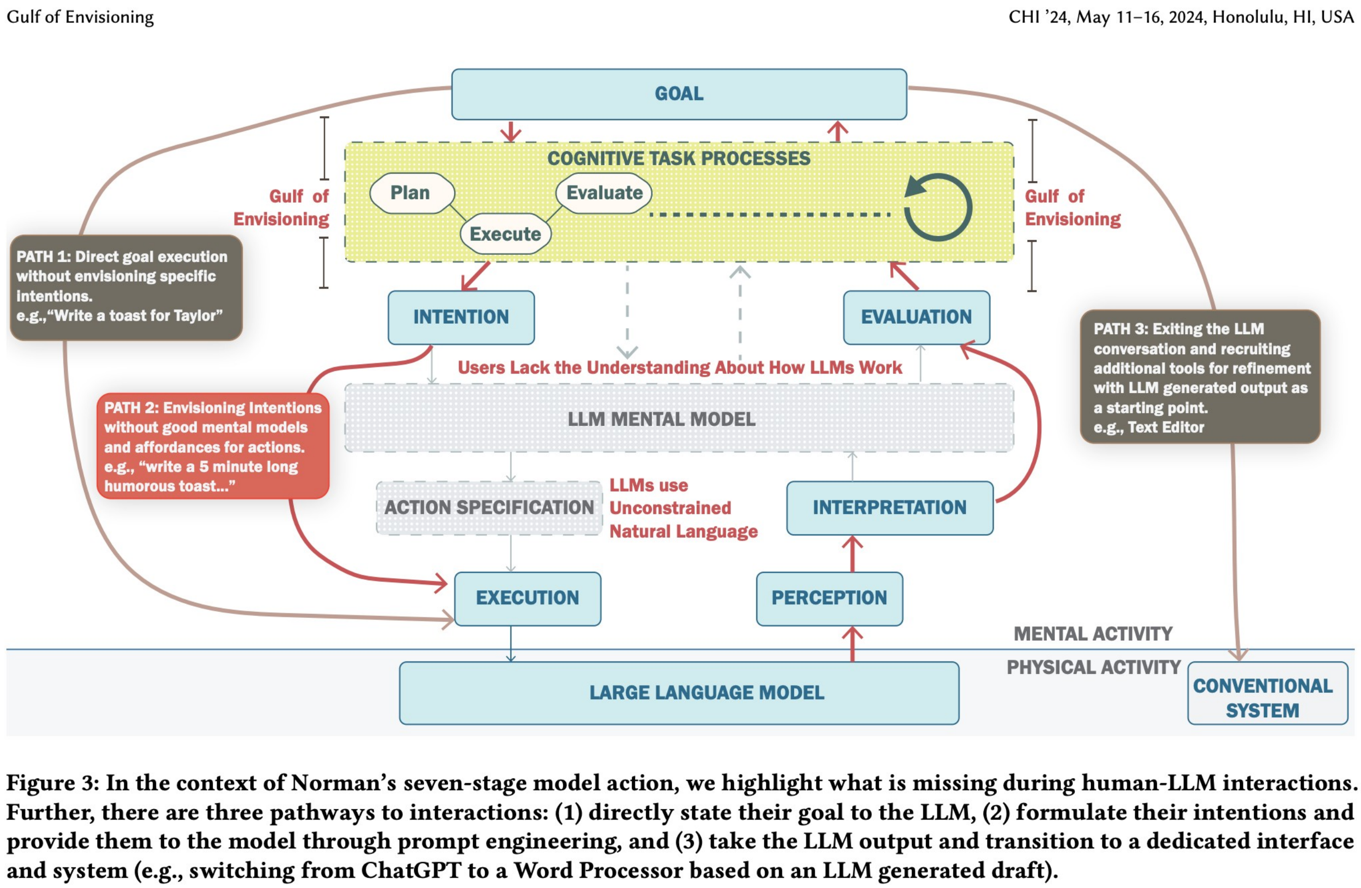
Figure 20: Figure from Zhang et al. (2024)

Bridging the Gulf of Envisioning: Cognitive Design Challenges in LLM Interfaces

Subramonyam, H., Pea, R., Pondoc, C. L., Agrawala, M., & Seifert, C. (2024). **Bridging the Gulf of Envisioning: Cognitive Design Challenges in LLM Interfaces** (arXiv:2309.14459; Version 2). arXiv. http://arxiv.org/abs/2309.14459

Abstract

Large language models (LLMs) exhibit dynamic capabilities and appear to comprehend complex and ambiguous natural language prompts. However, calibrating LLM interactions is challenging for interface designers and end-users alike. A central issue is our limited grasp of how human cognitive processes begin with a goal and form intentions for executing actions, a blindspot even in established interaction models such as Norman's gulfs of execution and evaluation. To address this gap, we theorize how end-users 'envision' translating their goals into clear intentions and craft prompts to obtain the desired LLM response. We define a process of Envisioning by highlighting three misalignments: (1) knowing whether LLMs can accomplish the task, (2) how to instruct the LLM to do the task, and (3) how to evaluate the success of the LLM's output in meeting the goal. Finally, we make recommendations to narrow the envisioning gulf in human-LLM interactions.



Gulf of Envisioning

CHI '24, May 11–16, 2024, Honolulu, HI, USA

GOAL
COGNITIVE TASK PROCESSES
Plan
Evaluate
Execute
Gulf of Envisioning
Gulf of Envisioning
PATH 1: Direct goal execution without envisioning specific intentions.
e.g.,"Write a toast for Taylor"
INTENTION
EVALUATION
Users Lack the Understanding About How LLMs Work
LLM MENTAL MODEL
PATH 2: Envisioning Intentions without good mental models and affordances for actions.
e.g., "write a 5 minute long humorous toast..."
PATH 3: Exiting the LLM conversation and recruiting additional tools for refinement with LLM generated output as a starting point.
e.g., Text Editor
ACTION SPECIFICATION
LLMs use Unconstrained Natural Language
INTERPRETATION
EXECUTION
PERCEPTION
MENTAL ACTIVITY
PHYSICAL ACTIVITY
LARGE LANGUAGE MODEL
CONVENTIONAL SYSTEM

Figure 3: In the context of Norman's seven-stage model action, we highlight what is missing during human-LLM interactions. Further, there are three pathways to interactions: (1) directly state their goal to the LLM, (2) formulate their intentions and provide them to the model through prompt engineering, and (3) take the LLM output and transition to a dedicated interface and system (e.g., switching from ChatGPT to a Word Processor based on an LLM generated draft).

Figure 21: Figure from Subramonyam et al. (2024)

Learning To Guide Human Decision Makers With Vision-Language Models

Banerjee, D., Teso, S., Sayin, B., & Passerini, A. (2024). **Learning To Guide Human Decision Makers With Vision-Language Models** (arXiv:2403.16501). arXiv. http://arxiv.org/abs/2403.16501

Abstract

There is increasing interest in developing AIs for assisting human decision-making in high-stakes tasks, such as medical diagnosis, for the purpose of improving decision quality and reducing cognitive strain. Mainstream approaches team up an expert with a machine learning model to which safer decisions are offloaded, thus letting the former focus on cases that demand their attention. his separation of responsibilities setup, however, is inadequate for high-stakes scenarios. On the one hand, the expert may end up over-relying on the machine's decisions due to anchoring bias, thus losing the human oversight that is increasingly being required by regulatory agencies to ensure trustworthy AI. On the other hand, the expert is left entirely unassisted on the (typically hardest) decisions on which the model abstained. As a remedy, we introduce learning to guide (LTG), an alternative framework in which - rather than taking control from the human expert - the machine provides guidance useful for decision making, and the human is entirely responsible for coming up with a decision. In order to ensure guidance is interpretable} and task-specific, we develop SLOG, an approach for turning any vision-language model into a capable generator of textual guidance by leveraging a modicum of human feedback. Our empirical evaluation highlights the promise of SLOG on a challenging, real-world medical diagnosis task.

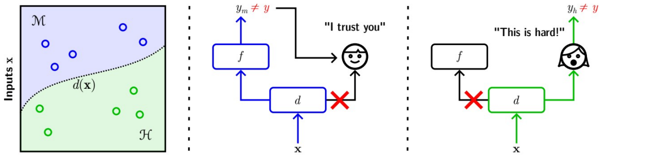
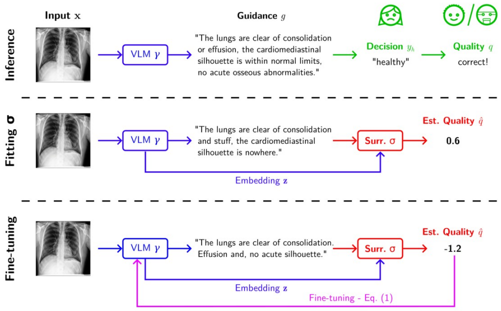


Figure 1: **Left**: Existing HDM approaches employ a deferral function $d(\mathbf{x})$ to *partition* the input space $\mathcal{X}$ into $\mathcal{H}$ and $\mathcal{M}$. **Middle**: A predictor $f(\mathbf{x})$ handles those inputs falling in $\mathcal{M}$ (**blue** arrow). Because of *anchoring bias*, the human expert may end up blindly trusting its (possibly poor) decisions y_m. **Right**: The human, on the other hand, is left completely unassisted for those (possibly hard) decisions falling in $\mathcal{H}$, increasing the chance of mistakes in the human's decisions y_h (**green** arrow).



Input x
Guidance g
Inference
VLM γ
"The lungs are clear of consolidation or effusion, the cardiomediastinal silhouette is within normal limits, no acute osseous abnormalities."
Decision y_h
"healthy"
Quality q
correct!
Fitting σ
VLM γ
"The lungs are clear of consolidation and stuff, the cardiomediastinal silhouette is nowhere."
Surr. σ
Est. Quality $\hat{q}$
0.6
Embedding z
Fine-tuning
VLM γ
"The lungs are clear of consolidation. Effusion and, no acute silhouette."
Surr. σ
Est. Quality $\hat{q}$
-1.2
Embedding z
Fine-tuning - Eq. (1)

Figure 2: **The SLOG approach to learning to guide**. **Top**: Given an input $\mathbf{x}$, SLOG uses a VLM γ to output textual guidance g in support of human decision making. Here, q indicates the quality of the *human's* downstream decision. **Middle**: The surrogate σ_{quality} estimates the quality of the downstream decisions and it is trained using a modicum on annotated guidance-quality pairs. **Bottom**: Given a trained surrogate σ_{quality}, SLOG fine-tunes the VLM to output guidance g achieving high (estimated) decision quality.

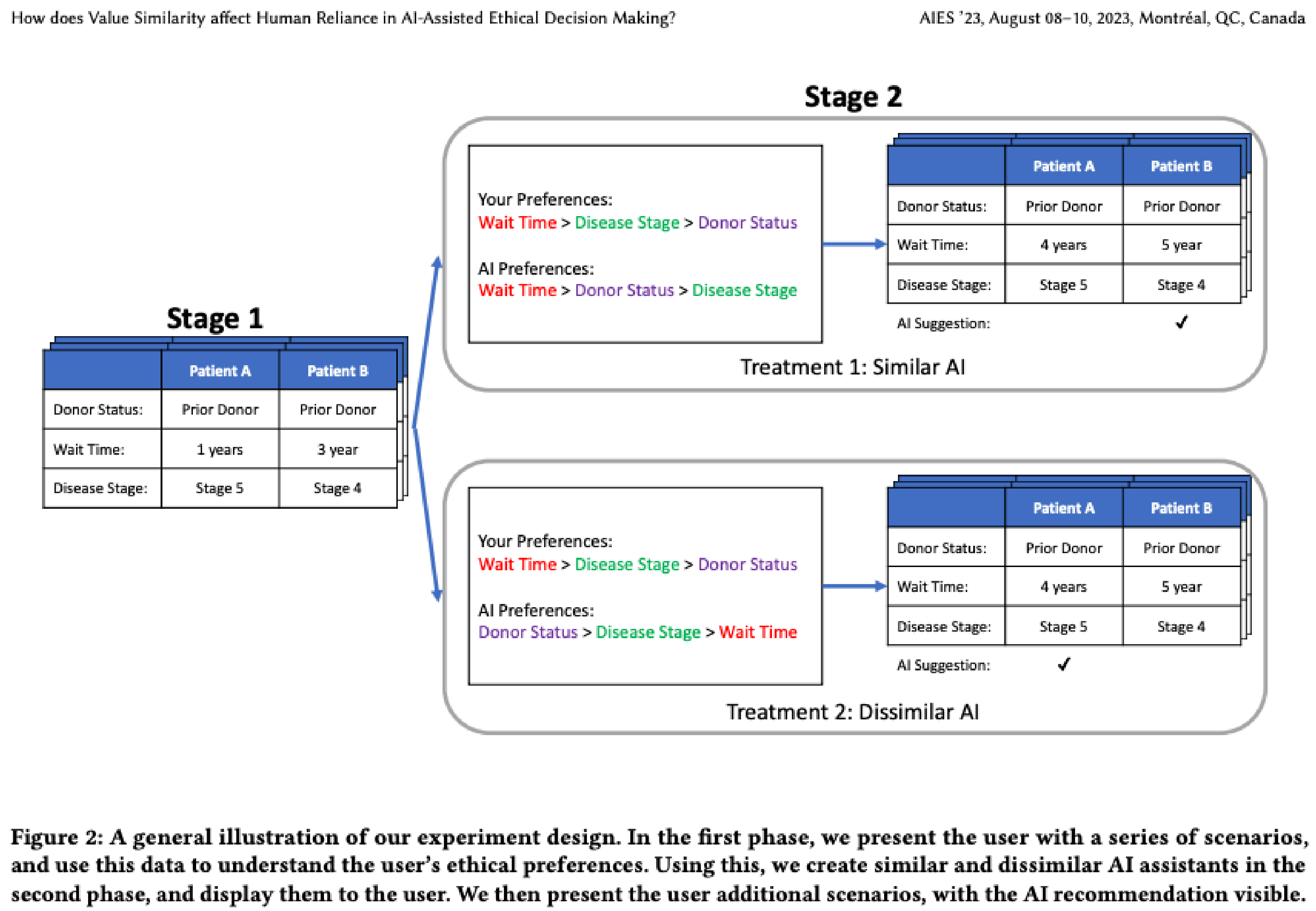
Figure 22: Figures from Banerjee et al. (2024)

How does Value Similarity affect Human Reliance in AI-Assisted Ethical Decision Making?

Narayanan, S., Yu, G., Ho, C.-J., & Yin, M. (2023). **How does Value Similarity affect Human Reliance in AI-Assisted Ethical Decision Making?** Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 49–57. https://doi.org/10.1145/3600211.3604709

Abstract

This paper explores the impact of value similarity between humans and AI on human reliance in the context of AI-assisted ethical decision-making. Using kidney allocation as a case study, we conducted a randomized human-subject experiment where workers were presented with ethical dilemmas in various conditions, including no AI recommendations, recommendations from a similar AI, and recommendations from a dissimilar AI. We found that recommendations provided by a dissimilar AI had a higher overall effect on human decisions than recommendations from a similar AI. However, when humans and AI disagreed, participants were more likely to change their decisions when provided with recommendations from a similar AI. The effect was not due to humans' perceptions of the AI being similar, but rather due to the AI displaying similar ethical values through its recommendations. We also conduct a preliminary analysis on the relationship between value similarity and trust, and potential shifts in ethical preferences at the population-level.



How does Value Similarity affect Human Reliance in AI-Assisted Ethical Decision Making? AIES '23, August 08–10, 2023, Montréal, QC, Canada

Stage 1

	Patient A	Patient B
Donor Status:	Prior Donor	Prior Donor
Wait Time:	1 years	3 year
Disease Stage:	Stage 5	Stage 4

Treatment 1: Similar AI

Your Preferences:
Wait Time > Disease Stage > Donor Status

AI Preferences:
Wait Time > Donor Status > Disease Stage

	Patient A	Patient B
Donor Status:	Prior Donor	Prior Donor
Wait Time:	4 years	5 year
Disease Stage:	Stage 5	Stage 4
AI Suggestion:		✓

Treatment 2: Dissimilar AI

Your Preferences:
Wait Time > Disease Stage > Donor Status

AI Preferences:
Donor Status > Disease Stage > Wait Time

	Patient A	Patient B
Donor Status:	Prior Donor	Prior Donor
Wait Time:	4 years	5 year
Disease Stage:	Stage 5	Stage 4
AI Suggestion:	✓	

Figure 2: A general illustration of our experiment design. In the first phase, we present the user with a series of scenarios, and use this data to understand the user's ethical preferences. Using this, we create similar and dissimilar AI assistants in the second phase, and display them to the user. We then present the user additional scenarios, with the AI recommendation visible.

Figure 23: Figure from Narayanan et al. (2023)

Determinants of LLM-assisted Decision-Making

Eigner, E., & Händler, T. (2024). **Determinants of LLM-assisted Decision-Making** (arXiv:2402.17385). arXiv. http://arxiv.org/abs/2402.17385

Abstract

Decision-making is a fundamental capability in everyday life. Large Language Models (LLMs) provide multifaceted support in enhancing human decision-making processes. However, understanding the influencing factors of LLM-assisted decision-making is crucial for enabling individuals to utilize LLM-provided advantages and minimize associated risks in order to make more informed and better decisions. This study presents the results of a comprehensive literature analysis, providing a structural overview and detailed analysis of determinants impacting decision-making with LLM support. In particular, we explore the effects of technological aspects of LLMs, including transparency and prompt engineering, psychological factors such as emotions and decision-making styles, as well as decision specific determinants such as task difficulty and accountability. In addition, the impact of the determinants on the decision-making process is illustrated via multiple application scenarios. Drawing from our analysis, we develop a dependency framework that systematizes possible interactions in terms of reciprocal interdependencies between these determinants. Our research reveals that, due to the multifaceted interactions with various determinants, factors such as trust in or reliance on LLMs, the user's mental model, and the characteristics of information processing are identified as significant aspects influencing LLM-assisted decision-making processes. Our findings can be seen as crucial for improving decision quality in human-AI collaboration, empowering both users and organizations, and designing more effective LLM interfaces. Additionally, our work provides a foundation for future empirical investigations on the determinants of decision-making assisted by LLMs.

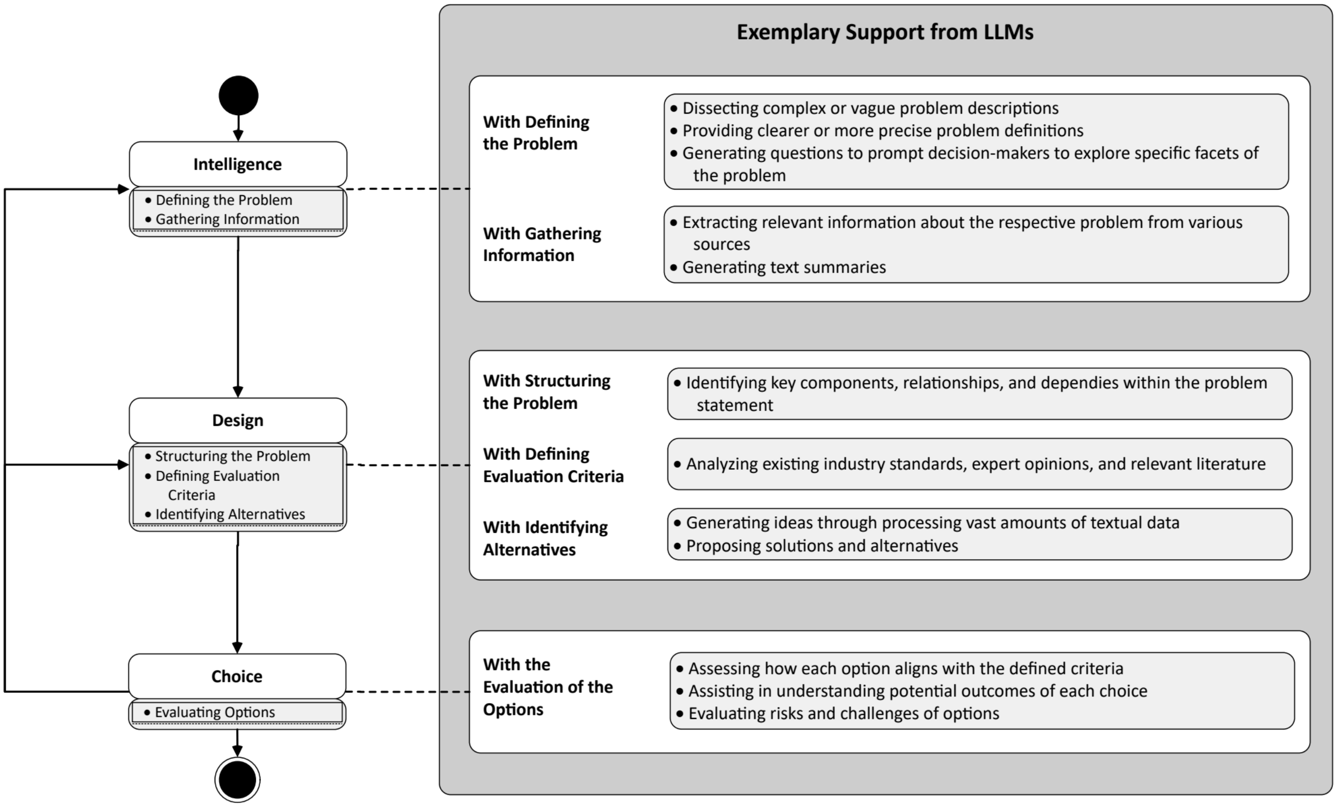

Exemplary Support from LLMs
Intelligence
• Defining the Problem
• Gathering Information
Design
• Structuring the Problem
• Defining Evaluation Criteria
• Identifying Alternatives
Choice
• Evaluating Options
With Defining the Problem
• Dissecting complex or vague problem descriptions
• Providing clearer or more precise problem definitions
• Generating questions to prompt decision-makers to explore specific facets of the problem
With Gathering Information
• Extracting relevant information about the respective problem from various sources
• Generating text summaries
With Structuring the Problem
• Identifying key components, relationships, and dependies within the problem statement
With Defining Evaluation Criteria
• Analyzing existing industry standards, expert opinions, and relevant literature
With Identifying Alternatives
• Generating ideas through processing vast amounts of textual data
• Proposing solutions and alternatives
With the Evaluation of the Options
• Assessing how each option aligns with the defined criteria
• Assisting in understanding potential outcomes of each choice
• Evaluating risks and challenges of options

Figure 1: Key stages in the decision-making process oriented to Simon [169] extended by LLM support options.

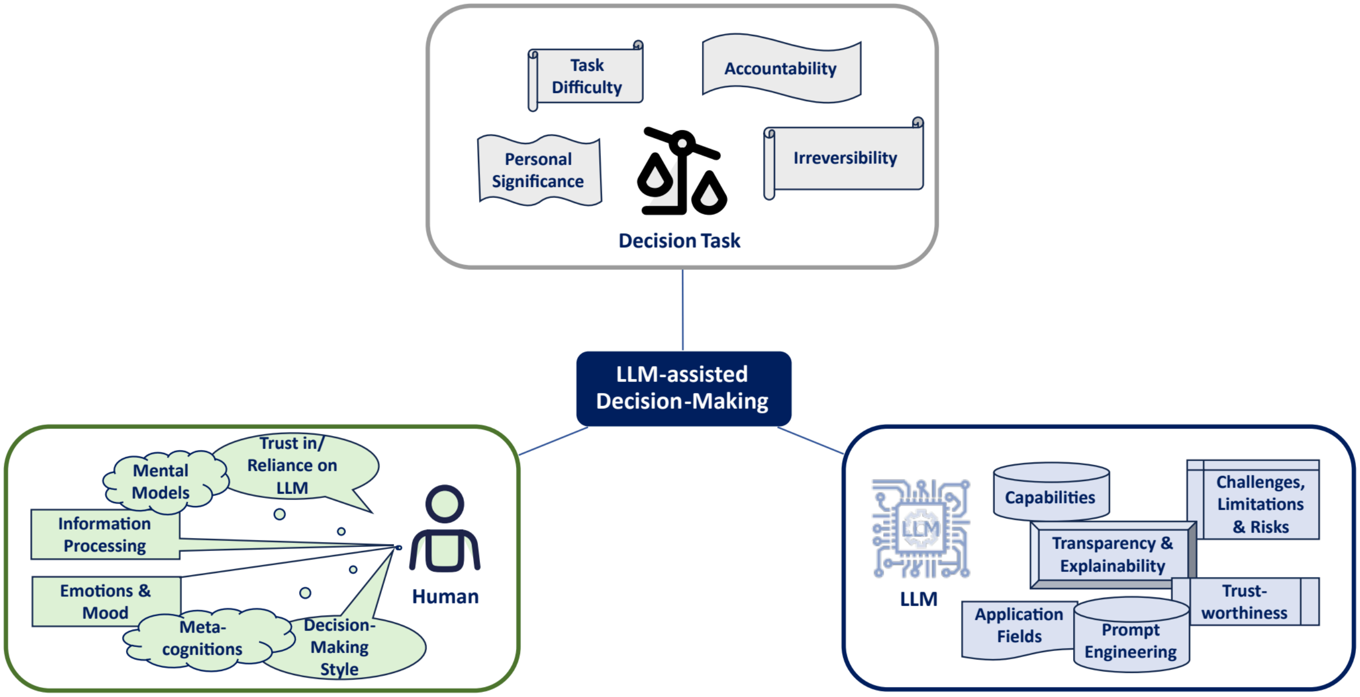


Figure 3: Schematic overview of addressed determinants of LLM-assisted decision-making.

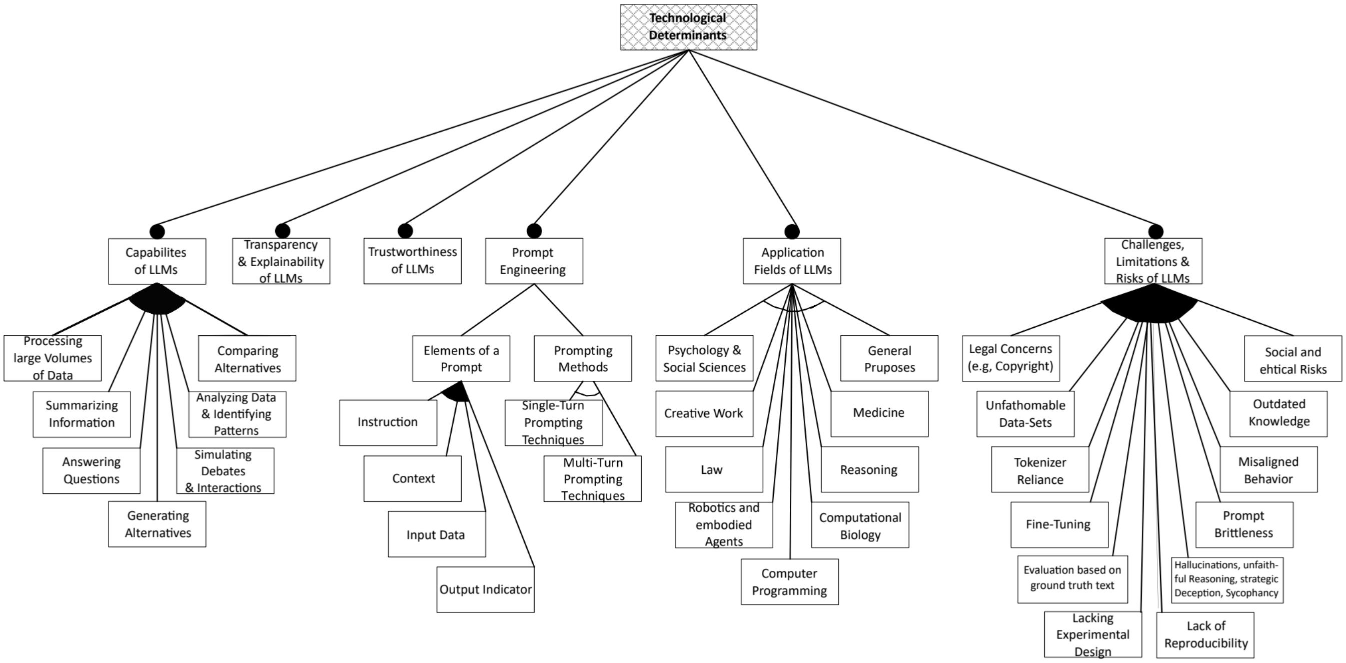

Technological Determinants
Capabilites of LLMs
Transparency & Explainability of LLMs
Trustworthiness of LLMs
Prompt Engineering
Application Fields of LLMs
Challenges, Limitations & Risks of LLMs
Processing large Volumes of Data
Comparing Alternatives
Summarizing Information
Analyzing Data & Identifying Patterns
Answering Questions
Simulating Debates & Interactions
Generating Alternatives
Elements of a Prompt
Prompting Methods
Instruction
Single-Turn Prompting Techniques
Context
Multi-Turn Prompting Techniques
Input Data
Output Indicator
Psychology & Social Sciences
General Pruposes
Creative Work
Medicine
Law
Reasoning
Robotics and embodied Agents
Computational Biology
Computer Programming
Legal Concerns (e.g, Copyright)
Social and ehtical Risks
Unfathomable Data-Sets
Outdated Knowledge
Tokenizer Reliance
Misaligned Behavior
Fine-Tuning
Prompt Brittleness
Evaluation based on ground truth text
Hallucinations, unfaith-ful Reasoning, strategic Deception, Sycophancy
Lacking Experimental Design
Lack of Reproducibility

Figure 4: Technological determinants of LLM-assisted decision-making.

Figure 24: Figures from Eigner & Händler (2024)

A Taxonomy of Human and ML Strengths in Decision-Making to Investigate Human-ML Complementarity.

Rastogi, C., Leqi, L., Holstein, K., & Heidari, H. (2023). **A Taxonomy of Human and ML Strengths in Decision-Making to Investigate Human-ML Complementarity.** Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 11, 127–139. https://doi.org/10.1609/hcomp.v11i1.27554

Abstract

Hybrid human-ML systems increasingly make consequential decisions in a wide range of domains. These systems are often introduced with the expectation that the combined human-ML system will achieve complementary performance, that is, the combined decision-making system will be an improvement compared with either decision-making agent in isolation. However, empirical results have been mixed, and existing research rarely articulates the sources and mechanisms by which complementary performance is expected to arise. Our goal in this work is to provide conceptual tools to advance the way researchers reason and communicate about human-ML complementarity. Drawing upon prior literature in human psychology, machine learning, and human-computer interaction, we propose a taxonomy characterizing distinct ways in which human and ML-based decision-making can differ. In doing so, we conceptually map potential mechanisms by which combining human and ML decision-making may yield complementary performance, developing a language for the research community to reason about design of hybrid systems in any decision-making domain. To illustrate how our taxonomy can be used to investigate complementarity, we provide a mathematical aggregation framework to examine enabling conditions for complementarity. Through synthetic simulations, we demonstrate how this framework can be used to explore specific aspects of our taxonomy and shed light on the optimal mechanisms for combining human-ML judgments.

Rastogi et al. (2023)

Take Caution in Using LLMs as Human Surrogates: Scylla Ex Machina

Gao, Y., Lee, D., Burtch, G., & Fazelpour, S. (2024). **Take Caution in Using LLMs as Human Surrogates: Scylla Ex Machina** (No. arXiv:2410.19599). arXiv. http://arxiv.org/abs/2410.19599

Abstract

Human decision-making is filled with a variety of paradoxes demonstrating deviations from rationality principles. Do state-of-the-art artificial intelligence (AI) models also manifest these paradoxes when making decisions? As a case study, in this work we investigate whether GPT-4, a recently released state-of-the-art language model, would show two well-known paradoxes in human decision-making: the Allais paradox and the Ellsberg paradox. We demonstrate that GPT-4 succeeds in the two variants of the Allais paradox (the common-consequence effect and the common-ratio effect) but fails in the case of the Ellsberg paradox. We also show that providing GPT-4 with high-level normative principles allows it to succeed in the Ellsberg paradox, thus elevating GPT-4's decision-making rationality. We discuss the implications of our work for AI rationality enhancement and AI-assisted decision-making.

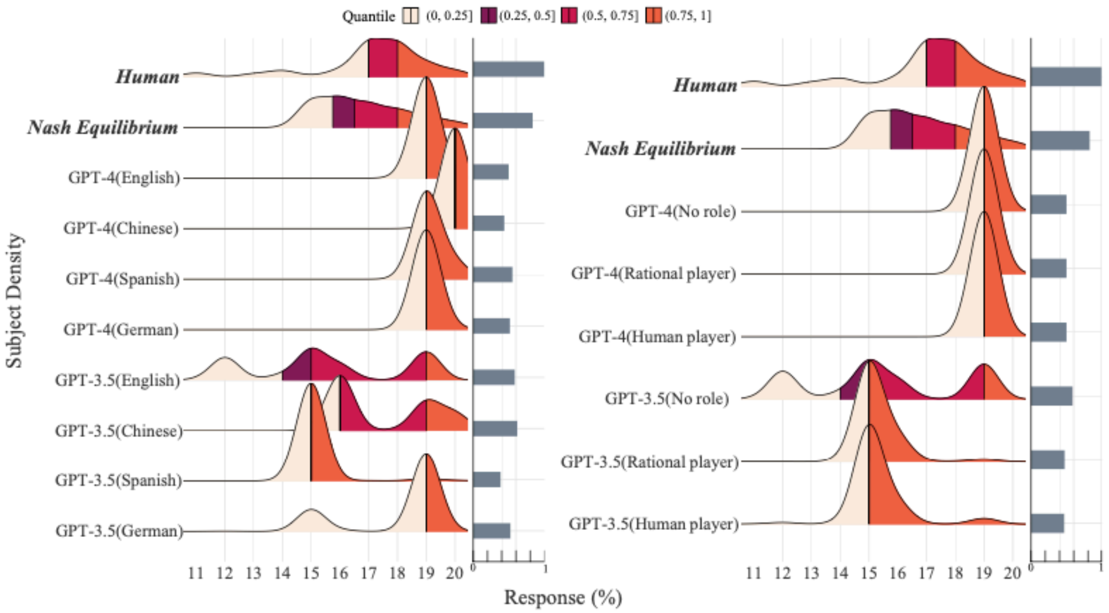


Figure 2: **Prompt Brittleness: Roles and Languages.** The bar chart on the right shows the similarity between the distribution of different subjects and human subjects, measured by Jensen-Shannon divergence scores. Missing percentiles (ranges) in some LLM distributions result from overlapping values (ranges).

Figure 25: Figure from Gao et al. (2024)

Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies.

Lai, V., Chen, C., Smith-Renner, A., Liao, Q. V., & Tan, C. (2023). **Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies.** 2023 ACM Conference on Fairness, Accountability, and Transparency, 1369–1385. https://doi.org/10.1145/3593013.3594087

Abstract

AI systems are adopted in numerous domains due to their increas- ingly strong predictive performance. However, in high-stakes domains such as criminal justice and healthcare, full automation is often not desirable due to safety, ethical, and legal concerns, yet fully manual approaches can be inaccurate and time-consuming. As a result, there is growing interest in the research community to augment human decision making with AI assistance. Besides developing AI technologies for this purpose, the emerging field of human-AI decision making must embrace empirical approaches to form a foundational understanding of how humans interact and work with AI to make decisions. To invite and help structure research efforts towards a science of understanding and improving human-AI decision making, we survey recent literature of empirical human-subject studies on this topic. We summarize the study design choices made in over 100 papers in three important aspects: (1) decision tasks, (2) AI assistance elements, and (3) evaluation metrics. For each aspect, we summarize current trends, discuss gaps in current

practices of the field, and make a list of recommendations for future research. Our work highlights the need to develop com- mon frameworks to account for the design and research spaces of human-AI decision making, so that researchers can make rigorous choices in study design, and the research community can build on each other's work and produce generalizable scientific knowledge. We also hope this work will serve as a bridge for HCI and AI communities to work together to mutually shape the empirical science and computational technologies for human-AI decision making.

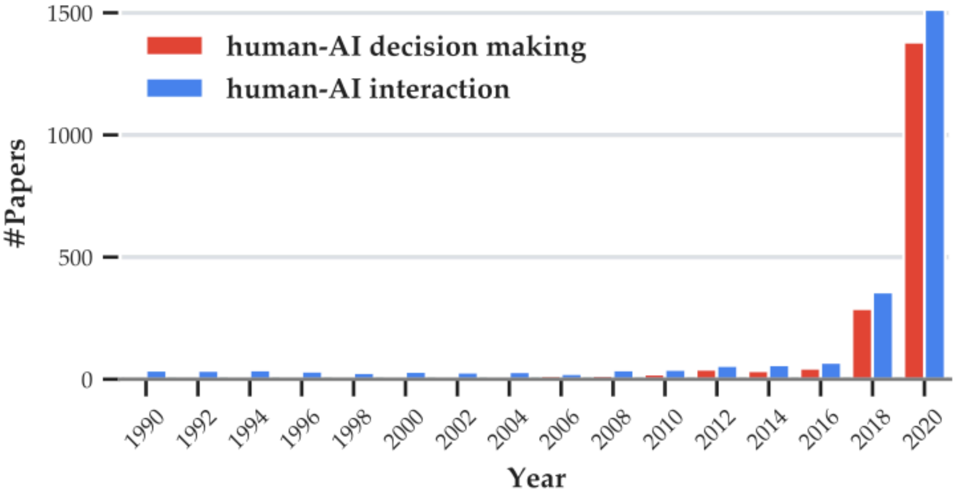


Figure 1: The number of papers based on Google Scholar for two queries, human-AI interaction and human-AI decision making, over the past years.

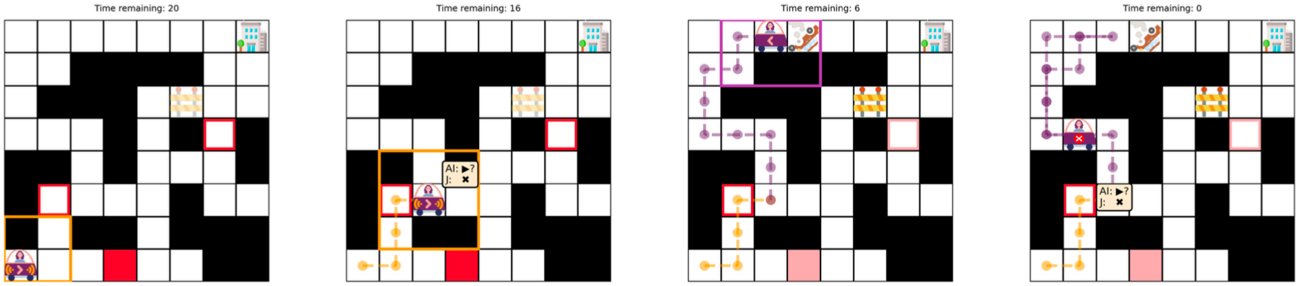
Figure 26: Figure from Lai et al. (2023)

Towards a computational model of responsibility judgments in sequential human-AI collaboration

Tsirtsis, S., Gomez Rodriguez, M., & Gerstenberg, T. (2024). **Towards a computational model of responsibility judgments in sequential human-AI collaboration.** In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 46). https://osf.io/preprints/psyarxiv/m4yad

Abstract

When a human and an AI agent collaborate to complete a task and something goes wrong, who is responsible? Prior work has developed theories to describe how people assign responsibility to individuals in teams. However, there has been little work studying the cognitive processes that underlie responsibility judgments in human-AI collaborations, especially for tasks comprising a sequence of interdependent actions. In this work, we take a step towards filling this gap. Using semi-autonomous driving as a paradigm, we develop an environment that simulates stylized cases of human-AI collaboration using a generative model of agent behavior. We propose a model of responsibility that considers how unexpected an agent's action was, and what would have happened had they acted differently. We test the model's predictions empirically and find that in addition to action expectations and counterfactual considerations, participants' responsibility judgments are also affected by how much each agent actually contributed to the outcome.



(a) The **AI** starts driving, unaware of the road closure

(b) The **AI** asks for confirmation to go right and **Jane** rejects

(c) **Jane** takes control of the car but encounters an accident

(d) Time runs out and they **fail** to reach the workplace

Figure 1: **Illustration of a commute in our semi-autonomous driving environment.** The human agent (**Jane**) and the **AI** are both in the same car and their goal is to reach the workplace within the time limit shown above the grid. The sign (📶) indicates that the AI is in control. The grid contains three traffic spots, one congested (🟥) and two non congested (□), whose status is initially known only to the AI. It also contains a road closure (🚧) which is known to the human but unknown to the AI. Obstacles that are unknown to the agent in control but known to the other agent appear faded. The arrow signs marked on the car (*e.g.*, ❯) indicate the direction that the driver in control is planning to follow. The 3 × 3 rectangle around the car represents the agents' field of view via which they discover obstacles that are previously unknown to them. Here, the accident (💥) present at the top row of the grid becomes visible only after the car goes next to it and it enters the agent's field of view.

Figure 27: Figure from Tsirtsis et al. (2024)

References

Banerjee, D., Teso, S., Sayin, B., & Passerini, A. (2024). *Learning To Guide Human Decision Makers With Vision-Language Models* (arXiv:2403.16501). arXiv. https://arxiv.org/abs/2403.16501

Bhatia, S. (2024). Exploring variability in risk taking with large language models. *Journal of Experimental Psychology: General*, *153*(7), 1838–1860. https://doi.org/10.1037/xge0001607

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–21. https://doi.org/10.1145/3449287

Chen, Y., Liu, T. X., Shan, Y., & Zhong, S. (2023). The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences*, *120*(51), e2316205120. https://doi.org/10.1073/pnas.2316205120

Cheung, V., Maier, M., & Lieder, F. (2024). *Large Language Models Amplify Human Biases in Moral Decision-Making*. https://doi.org/10.31234/osf.io/aj46b

Choi, A. S., Akter, S. S., Singh, J. P., & Anastasopoulos, A. (2024). *The LLM Effect: Are Humans Truly Using LLMs, or Are They Being Influenced By Them Instead?* (arXiv:2410.04699). arXiv. https://arxiv.org/abs/2410.04699

Eigner, E., & Händler, T. (2024). *Determinants of LLM-assisted Decision-Making* (arXiv:2402.17385). arXiv. https://arxiv.org/abs/2402.17385

Gao, Y., Lee, D., Burtch, G., & Fazelpour, S. (2024). *Take Caution in Using LLMs as Human Surrogates: Scylla Ex Machina* (arXiv:2410.19599). arXiv. https://arxiv.org/abs/2410.19599

Goli, A., & Singh, A. (2024). Can Large Language Models Capture Human Preferences? *Marketing Science*. https://doi.org/10.1287/mksc.2023.0306

Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, *3*(10), 833–838. https://doi.org/10.1038/s43588-023-00527-x

Lai, V., Chen, C., Smith-Renner, A., Liao, Q. V., & Tan, C. (2023). Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 1369–1385. https://doi.org/10.1145/3593013.3594087

Lampinen, A. K., Dasgupta, I., Chan, S. C. Y., Sheahan, H. R., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2024). Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, *3*(7), pgae233. https://doi.org/10.1093/pnasnexus/pgae233

Macmillan-Scott, O., & Musolesi, M. (2024). (Ir)rationality and cognitive biases in large language models. *Royal Society Open Science*, *11*(6), 240255. https://doi.org/10.1098/rsos.240255

Matz, S. C., Teeny, J. D., Vaid, S. S., Peters, H., Harari, G. M., & Cerf, M. (2024). The potential of generative AI for personalized persuasion at scale. *Scientific Reports*, *14*(1), 4692. https://doi.org/10.1038/s41598-024-53755-0

Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, *121*(9), e2313925121. https://doi.org/10.1073/pnas.2313925121

Narayanan, S., Yu, G., Ho, C.-J., & Yin, M. (2023). How does Value Similarity affect Human Reliance in AI-Assisted Ethical Decision Making? *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 49–57. https://doi.org/10.1145/3600211.3604709

Nguyen, J. (2024). Human Bias in AI Models? Anchoring Effects and Mitigation Strategies in Large Language Models. *Journal of Behavioral and Experimental Finance*, 100971. https://doi.org/10.1016/j.jbef.2024.100971

Nobandegani, A. S., Rish, I., & Shultz, T. R. (2023). Decision-Making Paradoxes in Humans vs Machines: The case of the Allais and Ellsberg Paradoxes. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

Rastogi, C., Leqi, L., Holstein, K., & Heidari, H. (2023). A Taxonomy of Human and ML Strengths in Decision-Making to Investigate Human-ML Complementarity. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *11*, 127–139. https://doi.org/10.1609/hcomp.v11i1.27554

Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., & Tomsett, R. (2022). Deciding Fast and Slow:

The Role of Cognitive Biases in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), 1–22. https://doi.org/10.1145/3512930

Roberts, J., Moore, K., Pham, T., Ewaleifoh, O., & Fisher, D. (2024). *Large Language Model Recall Uncertainty is Modulated by the Fan Effect*.

Stadler, M., Bannert, M., & Sailer, M. (2024). Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior*, *160*, 108386. https://doi.org/10.1016/j.chb.2024.108386

Subramonyam, H., Pea, R., Pondoc, C. L., Agrawala, M., & Seifert, C. (2024). Bridging the Gulf of Envisioning: Cognitive Design Challenges in LLM Interfaces. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–19. https://arxiv.org/abs/2309.14459

Suri, G., Slater, L. R., Ziaee, A., & Nguyen, M. (2024). Do large language models show decision heuristics similar to humans? A case study using GPT-35. *Journal of Experimental Psychology: General*, *153*(4), 1066–1075. https://doi.org/10.1037/xge0001547

Swaroop, S., Buçinca, Z., Gajos, K. Z., & Doshi-Velez, F. (2024). Accuracy-Time Tradeoffs in AI-Assisted Decision Making under Time Pressure. *Proceedings of the 29th International Conference on Intelligent User Interfaces*, 138–154. https://doi.org/10.1145/3640543.3645206

Tjuatja, L., Chen, V., Wu, T., Talwalkwar, A., & Neubig, G. (2024). Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design. *Transactions of the Association for Computational Linguistics*, *12*, 1011–1026. https://doi.org/10.1162/tacl_a_00685

Tsirtsis, S., Rodriguez, M. G., & Gerstenberg, T. (2024). *Towards a computational model of responsibility judgments in sequential human-AI collaboration*. https://doi.org/10.31234/osf.io/m4yad

Westphal, M., Vössing, M., Satzger, G., Yom-Tov, G. B., & Rafaeli, A. (2023). Decision control and explanations in human-AI collaboration: Improving user perceptions and compliance. *Computers in Human Behavior*, *144*, 107714. https://doi.org/10.1016/j.chb.2023.107714

Yax, N., Anlló, H., & Palminteri, S. (2024). Studying and improving reasoning in humans and machines. *Communications Psychology*, *2*(1), 1–16. https://doi.org/10.1038/s44271-024-00091-8

Zhang, S., Wang, X., Zhang, W., Chen, Y., Gao, L., Wang, D., Zhang, W., Wang, X., & Wen, Y. (2024). *Mutual Theory of Mind in Human-AI Collaboration: An Empirical Study with LLM-driven AI Agents in a Real-time Shared Workspace Task* (arXiv:2409.08811). arXiv. https://arxiv.org/abs/2409.08811

Zhao, Y., Huang, Z., Seligman, M., & Peng, K. (2024). Risk and prosocial behavioural cues elicit human-like response patterns from AI chatbots. *Scientific Reports*, *14*(1), 7095. https://doi.org/10.1038/s41598-024-55949-y