

The Rational Machine: Evaluating Decision-Making Heuristics of LLMs

1 Introduction

Research on human behavior has led to the development of models to understand how individuals make decisions under uncertainty, such as the expected value model, expected utility model, and various non-expected utility models. As artificial intelligence (AI) rapidly advances and becomes increasingly integrated into various decision-making tasks, it is crucial to comprehend how these technologies handle uncertain situations.

This study explores the decision-making processes under uncertainty exhibited by AI, specifically large language models (LLMs), and compares them to human decision-making. By replicating behavioral experiments with LLMs, our research aims to provide insights into whether human heuristics are trained into and observed within these systems. Moreover, we discuss the policy implications and suggest ways to develop more effective AI systems capable of assisting with decision-making tasks.

2 Motivation

2.1 Expanding Abilities of Computer Systems

Historically, advancements in information technology, from simple spreadsheets to automated flight scheduling systems, have revolutionized decision-making processes, enabling humans to make better and faster decisions ([Huber 1990](#); [Molloy 1995](#)). Recent progress in artificial intelligence (AI) and large language models (LLMs; AI models that power generative AI such as ChatGPT) has granted computers increasingly complex abilities, such as language

understanding, text generation (i.e., writing), theory of mind (i.e., psychological reasoning), and common-sense reasoning ([Joublin et al. 2023](#); [Ye et al. 2023](#)).

Traditional AI systems that assist decision-making are primarily statistical, providing predictions or analyses based on domain-specific data. In contrast, LLMs are trained on vast amounts of language data, eliminating the need for users to provide task-specific data. For example, consider classifying an object from a text description. Statistical AI models require the description and the label as training data to create a classifier model. However, with LLMs, as they are already trained on human language, no additional training data is required to start classifying.

This simplicity dramatically improves the accessibility to AI for users without technical backgrounds, as evidenced by the growing popularity of OpenAI's ChatGPT. Consequently, LLMs are being actively explored as decision-making assistants across various domains, such as business, government, and healthcare ([Bahrini et al. 2023](#); [OpenAI 2023](#)).

2.2 Black Box Problem of Artificial Intelligence

Recently, some researchers argued that LLMs represent a primitive form of *artificial general intelligence* (AGI) ([Bubeck 2023](#)). AGI refers to highly autonomous AI systems with cognitive capabilities at or beyond human-level intelligence. Despite their impressive abilities, such as passing the bar exam ([OpenAI 2023](#)), these AI models raise the widely-known *Black Box Problem*, also known as AI interpretability ([Bathae 2018](#); [Petch 2022](#)). This problem stems from the difficulty of interpreting and understanding the reasoning behind AI choices due to the system's complexity.

As AI models become increasingly integrated into real-world decision-making processes and have the potential to impact all aspects of society, understanding their decision-making behavior and heuristics is crucial. Key motivations for studying AI decision-making processes include

1. *Gaining insights into the capabilities and limitations of AI compared to human decision-making:* AI systems learn and generalize from data, so their decision-making processes might differ from those of humans. Analyzing LLM decision-making behaviors can reveal underlying heuristics, biases, and strategies, potentially diverging from human decision-making. Such insights can help develop AI systems that align with human values, improve decision-making accuracy, and enhance human-AI collaboration.
2. *Ensuring responsible development and deployment of AI technologies:* Current AI models can autonomously make decisions with far-reaching consequences, including ethical, social, and economic implications. Understanding AI decision-making in uncertain situations can help mitigate risks associated with unintended consequences, biased decision-making, and unethical behavior, thus contributing to responsible design, development, and use of the technology.
3. *Building trust and transparency in AI systems:* It is vital to ensure that AI decision-making processes are transparent, interpretable, and explainable to humans. Transparent AI systems can facilitate trust and accountability, enable effective human oversight, and promote responsible and ethical use of AI in uncertain and complex decision-making scenarios.

3 Literature review

3.1 *Development of Decision Theory Under Uncertainty*

The behavioral decision-making theory has evolved significantly throughout history. In the context of uncertainty, the expected utility model of preferences has been the dominant model of individual behavior since the early eighteenth century, first introduced by Nicholas Bernoulli in resolving the St. Petersburg Paradox and later formalized by von Neumann and Morgenstern in game theory development (von Neumann and Morgenstern 1944).

However, with the work of Allais and Edwards in the early 1950s, psychologists and economists discovered evidence that individuals do not always conform to the expected utility model's key assumptions or predictions. This finding led to alternative or 'non-expected utility' models of risk preferences that accommodate these inconsistencies while retaining as much analytical power as possible.

One such non-expected utility model is *prospect theory*, developed by Kahneman and Tversky (1979). It suggests that individuals evaluate outcomes relative to a reference point rather than an objective standard, are loss-averse, and exhibit diminishing sensitivity to changes in probability as probabilities move away from certainty. Prospect theory has been shown to provide a better fit to observed behavior than the expected utility model in various contexts, including decisions under risk (Kahneman and Tversky 1979), decisions under ambiguity (Ellsberg 1961), and intertemporal choice (Loewenstein and Prelec 1992).

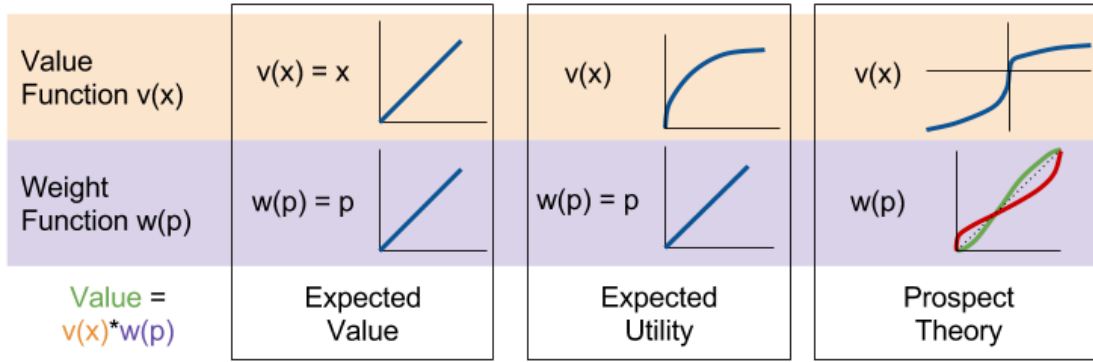


Figure 1: Development of decision theories and their formula. Diagram by Kevin Binz.

In addition to these models, researchers have also investigated other factors that influence decision-making under uncertainty. Various effects such as the *endowment effect* (the tendency for individuals to value an object more highly when they own it than when they do not; Kahneman et al. 1990), the *framing effect* (the influence of how options are presented; Tversky and Kahneman 1981), and the *certainty effect* (the tendency to prefer the certain outcome rather than an uncertain outcome, even if the uncertain outcome could be more valuable; Tversky and Kahneman 1979), have been shown to play a role in decisions involving risk and uncertainty.

In conclusion, while the expected utility model has been a dominant framework for understanding human decision-making under uncertainty for centuries, recent developments have led to alternative models that better account for repeated inconsistencies from this framework.

3.2 Decision-Making Heuristics in LLMs

Suri et al. (2023) demonstrated that LLMs such as GPT-3.5 exhibit human-like heuristics, including the *anchoring effect*, the *representativeness and availability effect*, the *framing effect*, and the *endowment effect* using non-monetary, qualitative questions. Suri et al. attributed this behavior to the inherent design of LLM, learning the context via language. She argues that the LLM learned human decision heuristics through patterns in natural languages. In contrast to Suri

et al.'s evaluation, this study focuses on quantitative questions and investigates whether LLM demonstrates human-like behaviors.

4 Methodology

This research aims to investigate the decision-making process of large language models under uncertainty and compare their choices with human decision-making. To achieve this, we recreated parts of the questionnaire-style experiments from the influential behavioral economics research, *Prospect Theory: An Analysis of Decision Under Risk* (Kahneman and Tversky 1979).

4.1 Experimental Setup

Tasks. A total of 18 multiple-choice questions were prompted to the AI model. Example problems are shown below:

Evaluation *equal_gain*:

- A: 50% chance to win 1000, 50% chance to win nothing;
- B: 100% chance to win 500.

Evaluation *non_monetary_1*:

- A: 50% chance to win a three-week tour of England, France, and Italy;
- B: 100% chance to win a one-week tour of England.

To eliminate context effects, options A and B were counterbalanced (i.e., Option A and B appeared in reverse order for 50% of all prompts) during evaluation. All questions are provided in Appendix Table 1.

Prompting. Zero-shot¹ chain-of-thought² prompting was used to increase arithmetic and reasoning performance (Wei et al. 2022; Brown et al. 2020). The evaluation questions were

¹ *Zero-shot prompting* refers to presenting the task to the LLM without providing a prior example. In contrast, *few-shot prompting* refers to the technique of providing the LLM with few examples and desired answers.

² *Chain-of-thought prompting* is a technique that encourages the LLM to explain its reasoning by adding “Let’s think step by step” (Kojima et al. 2022) at the end of the prompt.

limited to 8 questions per prompt to ensure the response was completed before reaching the token limit. This resulted in 2 sequential prompts per conversation (see Appendix Table 2 for an example of a full conversation). The chain-of-thought reasoning was restricted to a maximum of 3 sentences to reduce response time and token counts. Additionally, the desired output format was explicitly provided in the prompt to amplify the answers for multiple-choice questions (Moire 2022; OpenAI 2023). The template of the prompt is shown below:

Situations:

"A: 50% chance to win 1000, 50% chance to win nothing;

B: 100% chance to win 450."

"A: 50% chance to lose 1000, 50% chance to lose nothing;

B: 100% chance to lose 450."

{... 6 more evaluation questions}

You must choose between A and B.

Think step by step, and answer with the following format for each situations:

Reasoning: [your_reasoning_less_than_3_sentences]

Answer: [PREFER_A|PREFER_B]

Language model. This study evaluates **GPT-3.5** (gpt-3.5-turbo-0301) by OpenAI, the chat-optimized version of the text-davinci-003 model (OpenAI, 2023). The model was accessed using the OpenAI API, using the official Python openai package.

Results and Discussion

5.1 Results

Table 1: Evaluation results of zero-shot chain-of-thought prompting

Evaluation title	Options		Human ³ <i>a</i>		Difference <i>b - a</i>		GPT-3.5 ⁴ <i>b</i>		
	A	B	A	B	A	B	A	B	X ⁵
<i>simple_gain</i>	(+1,000, .5)	(+450, 1)					55%	45%	1%
<i>simple_loss</i>	(−1,000, .5)	(−450, 1)					28%	72%	1%
<i>equal_gain</i>	(+1,000, .5)	(+500, 1)	16%	84%	16%	−40%	32%	45%	24%
<i>equal_loss</i>	(−1,000, .5)	(−500, 1)	69%	31%	−41%	24%	28%	55%	17%
<i>variation1_1</i>	(+2,500, .33; +2,400, .66)	(+2,400, 1)	18%	82%	32%	−38%	50%	44%	7%
<i>variation1_2</i>	(+2,500, .33)	(+2,400, .34)	83%	17%	−40%	32%	43%	49%	8%
<i>variation2_1_gain</i>	(+4,000, .8)	(+3,000, 1)	20%	80%	53%	−63%	73%	17%	11%
<i>variation2_1_loss</i>	(−4,000, .8)	(−3,000, 1)	92%	8%	−78%	71%	15%	79%	7%
<i>variation2_2_gain</i>	(+4,000, .2)	(+3,000, .25)	65%	35%	11%	−14%	76%	21%	4%
<i>variation2_2_loss</i>	(−4,000, .2)	(−3,000, .25)	42%	58%	−25%	22%	17%	80%	4%
<i>overweighting_gain</i>	(+5,000, .001)	(+5, 1)	72%	28%	−37%	28%	35%	56%	9%
<i>overweighting_loss</i>	(−5,000, .001)	(−5, 1)	17%	83%	−3%	−6%	14%	77%	9%
<i>risk_context_gain</i>	(+6,000, .25)	(+4,000, .25, +2,000, .25)	18%	82%	27%	−54%	45%	29%	27%
<i>risk_context_loss</i>	(−6,000, .25)	(−4,000, .25, −2,000, .25)	70%	30%	−58%	33%	13%	63%	25%
<i>non_monetary_1</i>	(3-week tour, .5)	(1-week tour, 1)	22%	78%	24%	−53%	46%	25%	30%
<i>non_monetary_2</i>	(3-week tour, .05)	(1-week tour, .1)	67%	33%	−39%	23%	28%	56%	16%

The model applies the expected value framework. First, GPT-3.5 heavily utilized the *expected value framework* for decisions involving quantitative options that are not indifferent.

This observation is evident in all evaluations (except the qualitative evaluation *non_monetary*) and is supported by both evaluation results and the reasoning⁶ provided by the model.

Evaluations that demonstrate a strong reliance on the expected value framework are highlighted

³ $N > 60$ (results from Kahneman and Tversky (1979)).

⁴ $N = 200$ (100 responses for each counterbalanced option).

⁵ X denotes the response which does not include *PREFER_A* or *PREFER_B*, or contains both. Although the model was instructed to choose between A and B, it often responded with *INDIFFERENT* or similar answers. See Appendix Table 3 for most frequent literal answers for X .

⁶ The term *expected value* was mentioned 5,668 times in 149 out of 200 conversations. See Appendix Table 2 for reasoning examples.

in bold in Table 2 below. This decision-making behavior based on expected value is also apparent in the number of indifferent responses (i.e., X), indicating that the model has difficulty deciding between options with equivalent expected values.

Table 2: Evaluation results of GPT-3.5 and expected value for quantitative evaluations

Evaluation title	Options		Expected value ⁷		GPT-3.5		
	A	B	A	B	A	B	X
<i>simple_gain</i>	(+1,000, .5)	(+450, 1)	+500	> +450	55%	45%	1%
<i>simple_loss</i>	(−1,000, .5)	(−450, 1)	−500	< −450	28%	72%	1%
<i>equal_gain</i>	(+1,000, .5)	(+500, 1)	+500	= +500	32%	45%	24%
<i>equal_loss</i>	(−1,000, .5)	(−500, 1)	−500	= −500	28%	55%	17%
<i>variation1_1</i>	(+2,500, .33; +2,400, .66)	(+2,400, 1)	+2,409	> +2,400	50%	44%	7%
<i>variation1_2</i>	(+2,500, .33)	(+2,400, .34)	+825	> +816	43%	49%	8%
<i>variation2_1_gain</i>	(+4,000, .8)	(+3,000, 1)	+3,200	> +3,000	73%	17%	11%
<i>variation2_1_loss</i>	(−4,000, .8)	(−3,000, 1)	−3,200	< −3,000	15%	79%	7%
<i>variation2_2_gain</i>	(+4,000, .2)	(+3,000, .25)	+800	> +750	76%	21%	4%
<i>variation2_2_loss</i>	(−4,000, .2)	(−3,000, .25)	−800	< −750	17%	80%	4%
<i>overweighting_gain</i>	(+5,000, .001)	(+5, 1)	+5	=* +5	35%	56%	9%
<i>overweighting_loss</i>	(−5,000, .001)	(−5, 1)	−5	=* −5	14%	77%	9%
<i>risk_context_gain</i>	(+6,000, .25)	(+4,000, .25, +2,000, .25)	+1,500	=* +1,500	45%	29%	27%
<i>risk_context_loss</i>	(−6,000, .25)	(−4,000, .25, −2,000, .25)	−1,500	=* −1,500	13%	63%	25%

The model is risk-averse for losses. Second, unlike humans, GPT-3.5 did not display risk-seeking behavior in choices involving losses. For instance, in *variation2_1_loss* (see Table 3 below), 79% of AI responses opted for the guaranteed loss, while only 8% of human participants chose the same option.

Table 3: Evaluation *variation2_1_loss*

Options	(−4,000, .8)	Preference	(−3,000, 1)	Neither
Human ($n = 95$)	92%	>	8%	-
GPT-3.5 ($n = 200$)	15%	<	79%	7%

Notably, when both options have indifferent expected values, the model preferred risk-averse options (i.e., guaranteed loss) rather than demonstrating risk-seeking behaviors. Since the AI

⁷ * denotes significant inconsistencies with the calculated expected value.

exhibits preferences between indifferent expected values, we can conclude that the expected utility theory does not consistently model the AI's behavior.

Based on these observations, we can further analyze the AI's decision-making process using prospect theory by deriving a new value function and weighting function, thereby creating a risk profile for GPT-3.5. If $v_{GPT}(x)$ represents the value function and $\pi_{GPT}(x)$ symbolizes the weighting function of the model, the following holds true:

$$v_{GPT}(a) < v_{GPT}(b) \text{ if } a < b. \quad (\text{from expected value theory}^8)$$

$$\pi_{GPT}(.25)v_{GPT}(+6,000) > \pi_{GPT}(.25)[v_{GPT}(+4,000) + v_{GPT}(+2,000)].$$

(from risk_context_gain)

$$\pi_{GPT}(.25)v_{GPT}(-6,000) < \pi_{GPT}(.25)[v_{GPT}(-4,000) + v_{GPT}(-2,000)].$$

(from risk_context_loss)

Therefore, $v_{GPT}(+6,000) > v_{GPT}(+4,000) + v_{GPT}(+2,000)$ and $v_{GPT}(-6,000) < v_{GPT}(-4,000) + v_{GPT}(-2,000)$. From these preferences, we deduce that the value function for the model is convex for gains and concave for losses, which is the opposite to the finding of Kahneman and Tversky (1997, 278-279). Additionally, based on *variation2_1_loss* (see Table 3 above) and *equal_loss*, we observe that the model does not weigh losses more heavily than gains (i.e., the model prefers a guaranteed loss over gambling for a larger loss with a worse expected value). From this, we can infer that the value function v_{GPT} for losses is less steep⁹ (or almost symmetric to gains) compared to humans. The following is a graphical comparison between a hypothetical value function for the model and humans.

⁸ This refers to the *expected value theory* being demonstrated in quantitative options that are not indifferent.

⁹ In other words, this means the model doesn't register the pain of -1,000 more intensely than the pleasure of +1,000. Studies have shown that humans generally feel the pain more intensely than the pleasure.

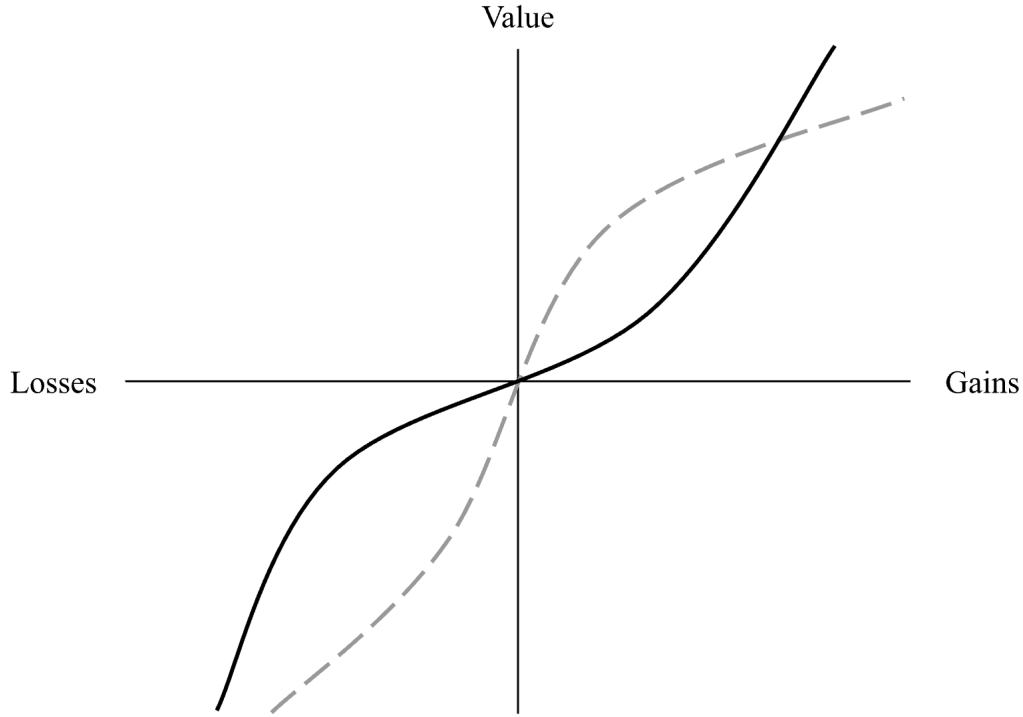


Figure 2: Hypothetical value function of GPT-3.5 (solid line) and human (dashed line; Kahneman and Tversky 1979).

With such value function, the following is also true:

$$\pi_{GPT}(.001)v_{GPT}(+5,000) < \pi_{GPT}(1)v_{GPT}(+5). \quad (\text{from } \textit{overweighting_gain})$$

$$\pi_{GPT}(.001)v_{GPT}(-5,000) < \pi_{GPT}(1)v_{GPT}(-5). \quad (\text{from } \textit{overweighting_loss})$$

This observation is consistent with the risk-averse behavior for both gains and losses, as described above.

The model struggles to quantify values in qualitative situations. Third, in evaluations *non_monetary_1* and *non_monetary_2*, the frequency of indifferent answers was significantly higher than other quantitative evaluations. In the reasoning, some responses attempted to quantify the lengths of the trips into numerical values. However, in most cases, the model identified the options as simply "better" or "worse" without providing a concrete rationale.

Without quantitative values in the situation, the model failed to deliver a coherent reasoning or a definitive answer.

In conclusion, between quantitative options, GPT-3.5 relied on the expected value framework for making decisions and exhibited risk-averse behavior for both gains and losses. However, in qualitative situations, the model failed to provide coherent reasoning or a definitive preference.

5.2 Limitations

The experiments conducted in this study are preliminary and come with several limitations.

First, unlike the original study involving human subjects, the AI was not strictly compelled to choose between options A or B (although instructed to do so), as it generated responses other than PREFER_A and PREFER_B. Designing a prompt that effectively compels the model to provide a specific answer is essential for future research.

Second, the chain-of-thought was limited to three sentences due to cost¹⁰ and maximum token limits. This limitation might have impacted the quality of the responses by increasing the likelihood of arithmetic errors or providing insufficient space to quantify qualitative situations.

Third, given the widespread influence of Kahneman and Tversky's original research in various academic fields, it is highly probable that the AI had been exposed to some of the prompted situations during its training. Additional tests with novel qualitative scenarios (i.e., questions not available on the Internet) could offer more insights into the model's behavior.

Lastly, the evaluations performed on GPT-3.5 cannot be generalized to model the behavior of all existing LLMs. Evaluation of more recent models, such as GPT-4 ([OpenAI 2023](#)) or PaLM 2

¹⁰ As of May 2023, GPT-3.5 costs \$0.002 to generate 1,000 tokens ([OpenAI 2023](#)). One full conversation (shown in Appendix Table 2) was approximately 4,500 tokens, costing \$0.009 on average.

([Anil et al. 2023](#); developed by Google), will be feasible if their APIs are made publicly accessible.

5.3 Discussion

This research provides preliminary evidence regarding the decision-making preferences of LLMs. In this section, we discuss potential approaches to improve the decision-making abilities of LLMs and consider policies to ensure the safe use of AI systems in real-world applications.

Teaching Preferences. Existing research has demonstrated that well-designed few-shot prompts can significantly enhance the performance of LLMs in natural language processing tasks, such as fact retrieval and information extraction ([Brown et al. 2020](#); [Zhao et al. 2021](#)). Similarly, we anticipate that users can effectively bias the AI's decision-making preferences (i.e., teach human preferences) with carefully crafted few-shot prompting. An example of a prompt designed to bias the model towards human-like decision-making is shown below:

Situation 1:

"A: 50% chance to win 1000, 50% chance to win nothing;

B: 100% chance to win 500."

Reasoning: The expected value for option A is $(0.5 \times 1000) + (0.5 \times 0) = 500$, while for option B it is 500. In order to not risk winning nothing, option B is more preferable.

Situation 2:

"A: 50% chance to lose 1000, 50% chance to lose nothing;

B: 100% chance to lose 500."

Reasoning: The expected value for option A is $(0.5 \times -1000) + (0.5 \times 0) = -500$, while for option B it is -500. In order to have a chance of losing nothing, option A is more preferable.

Situation:

{evaluation questions}

You must choose between A and B.

Think step by step, and answer with the following format for each situations:

Reasoning: [your_reasoning_less_than_3_sentences]

Answer: [PREFER_A|PREFER_B]

We expect the model to more accurately emulate human behavior when provided with more sophisticated few-shot examples. In addition to few-shot prompting, fine-tuning the LLM can also help the model exhibit a specific type of rationality.

Policy Implications. The growing integration of AI systems into real-world decision-making processes necessitates the implementation of robust policies and regulations to ensure their safe and responsible use. Key policy implications include:

1. *Regulation on Developing New AI Systems:* Implementing FDA-style regulations to monitor the development of new AI systems can help ensure their safety and efficacy before they are introduced to the market. These regulatory measures would require AI developers to adhere to strict guidelines and provide evidence of their system's performance and safety. For example, AI systems designed for medical diagnosis, financial investing, or autonomous driving could be subject to rigorous testing and validation, ensuring they meet established safety standards before being approved for public and commercial availability.
2. *Regulation on Consumer Safety of AI:* Establishing MSDS¹¹-style databases that are publicly available can contribute to increased transparency and consumer safety. These databases would contain extensive information about potential hazards, safe usage practices, and emergency procedures related to AI systems. By making this information accessible to everyone, users can make more informed decisions about the AI technologies they choose to adopt. For instance, an MSDS for LLMs can include a list of

¹¹ Material Safety Data Sheet. A document that lists occupational safety and health information for the use of various chemical substances and products, including instructions for safe use and potential hazards associated with a particular material or product.

prompts that can potentially generate harmful contents, known behavioral heuristics, and inherent biases such as gender, race, and religion (Liang et al. 2021).

6 Conclusions

In conclusion, this research presents preliminary insights into the decision-making preferences of large language models (LLMs), revealing some tendencies of GPT-3.5 to make decisions that are divergent from human reasoning. While the findings are subject to several limitations, they provide a valuable starting point for understanding the behavior of AI systems in various decision-making scenarios. Our study highlights potential approaches to improve LLM decision-making capabilities, such as few-shot prompting and fine-tuning, as well as the importance of implementing robust policies and regulations to ensure the safe and responsible use of AI systems in real-world applications.

As AI continues to permeate various aspects of our lives, further research is needed to fully understand the decision-making processes of LLMs and other AI systems. By addressing existing limitations and building on the findings presented in this study, researchers, policymakers, and AI developers can work together to harness the potential of AI in decision-making, while minimizing risks and ensuring the technology remains aligned with human values and ethical considerations.

7 Reproducibility

Full code and API responses can be found on the [author's GitHub repository](https://github.com/ntcho/bard/tree/main/Spring%202023/CMSC%20354%20-%20Natural%20Language%20Processing/Final%20Project)¹².

¹² <https://github.com/ntcho/bard/tree/main/Spring 2023/CMSC 354 – Natural Language Processing/Final Project>

8 Works Cited

- Anil, Rohan, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, et al. "PaLM 2 Technical Report." arXiv.org, May 17, 2023. <https://arxiv.org/abs/2305.10403>.
- Bahrini, Aram, Mohammadsadra Khamoshifar, Hossein Abbasimehr, Robert J. Riggs, Maryam Esmaeili, Rastin Mastali Majdabadkohne, and Morteza Pasehvar. "ChatGPT: Applications, Opportunities, and Threats." arXiv.org, April 14, 2023. <https://arxiv.org/abs/2304.09103>.
- Bathae, Yavar, "The Artificial Intelligence Black Box and the Failure of Intent and Causation," Harvard Journal of Law & Technology (Harvard JOLT) 31, no. 2 (Spring 2018): 889-938. <https://heinonline.org/HOL/P?h=hein.journals/hjlt31&i=907>.
- Binz, Kevin. "An Introduction to Prospect Theory." Fewer Lacunae, March 28, 2017. <https://kevinbinz.com/2016/10/26/prospect-theory/>.
- Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. "Language Models Are Few-Shot Learners." arXiv.org, July 22, 2020. <https://arxiv.org/abs/2005.14165>.
- Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, et al. "Sparks of Artificial General Intelligence: Early Experiments with GPT-4." arXiv.org, April 13, 2023. <https://doi.org/10.48550/arXiv.2303.12712>.
- Huber, George P. "A Theory of the Effects of Advanced Information Technologies on Organizational Design, Intelligence, and Decision Making." Academy of Management Review 15, no. 1 (1990): 47–71. <https://doi.org/10.5465/amr.1990.4308227>.
- Joublin, Frank, Antonello Ceravola, Joerg Deigmoeller, Michael Gienger, Mathias Franzius, and Julian Eggert. "A Glimpse in ChatGPT Capabilities and Its Impact for AI Research." arXiv.org, May 10, 2023. <https://arxiv.org/abs/2305.06087>.
- Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. "Large Language Models Are Zero-Shot Reasoners." arXiv.org, January 29, 2023. <https://arxiv.org/abs/2205.11916>.
- Liang, Paul Pu, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. "Towards Understanding and Mitigating Social Biases in Language Models." arXiv.org, June 24, 2021. <https://arxiv.org/abs/2106.13219>.

- Moire. “Amplifying GPT-3 on Closed-Ended Questions.” *generative.ink*, March 30, 2022. <https://generative.ink/posts/amplifying-gpt-3-on-closed-ended-questions/>.
- Molloy, Steve, and Charles R. Schwenk. “The Effects of Information Technology on Strategic Decision Making.” *Journal of Management Studies* 32, no. 3 (1995): 283–311. <https://doi.org/10.1111/j.1467-6486.1995.tb00777.x>.
- Neumann, John von, Oskar Morgenstern, and Ariel Rubinstein. *Theory of Games and Economic Behavior* (60th Anniversary Commemorative Edition). Princeton University Press, 1944. <http://www.jstor.org/stable/j.ctt1r2gkx>.
- OpenAI. “Best Practices for Prompt Engineering with OpenAI API.” OpenAI Help Center. Accessed May 20, 2023. https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api#h_f5edd2bf35.
- OpenAI. “GPT-4 Technical Report.” *arXiv.org*, March 27, 2023. <https://arxiv.org/abs/2303.08774>.
- OpenAI. GPT-4. Accessed May 20, 2023. <https://openai.com/research/gpt-4>.
- OpenAI. “Model Index for Researchers.” OpenAI API. Accessed May 20, 2023. <https://platform.openai.com/docs/model-index-for-researchers>.
- Petch, Jeremy, Shuang Di, and Walter Nelson. “Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology.” *Canadian Journal of Cardiology*, September 14, 2021. <https://doi.org/10.1016/j.cjca.2021.09.004>.
- Srivastava, Aarohi, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, et al. ‘Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models’. *ArXiv [Cs.CL]*, 9 June 2022. *arXiv*. <http://arxiv.org/abs/2206.04615>.
- Suri, Gaurav, Lily R. Slater, Ali Ziaee, and Morgan Nguyen. “Do Large Language Models Show Decision Heuristics Similar to Humans? A Case Study Using GPT-3.5.” *arXiv.org*, May 8, 2023. <https://arxiv.org/abs/2305.04400>.
- Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” *arXiv.org*, January 10, 2023. <https://arxiv.org/abs/2201.11903>.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, et al. “Emergent Abilities of Large Language Models.” arXiv.org, October 26, 2022. <https://arxiv.org/abs/2206.07682>.

Ye, Junjie, Xuanning Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhao Cui, et al. “A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models.” arXiv.org, March 18, 2023. <https://arxiv.org/abs/2303.10420>.

Zhao, Tony Z., Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. “Calibrate before Use: Improving Few-Shot Performance of Language Models.” arXiv.org, June 10, 2021. <https://arxiv.org/abs/2102.09690>.

A Appendix

Table 1: Evaluation questions

Evaluation title	Question
<i>simple_gain</i>	A: 50% chance to win 1000, 50% chance to win nothing; B: 100% chance to win 450.
<i>simple_loss</i>	A: 50% chance to lose 1000, 50% chance to lose nothing; B: 100% chance to lose 450.
<i>equal_gain</i>	A: 50% chance to win 1000, 50% chance to win nothing; B: 100% chance to win 500.
<i>equal_loss</i>	A: 50% chance to lose 1000, 50% chance to lose nothing; B: 100% chance to lose 500.
<i>variation1_1</i>	A: 33% chance to win 2500, 66% chance to win 2400, 1% chance to win nothing; B: 100% chance to win 2400.
<i>variation1_2</i>	A: 33% chance to win 2500, 67% chance to win nothing; B: 34% chance to win 2400, 66% chance to win nothing.
<i>variation2_1_gain</i>	A: 80% chance to win 4000; B: 100% chance to win 3000.
<i>variation2_1_loss</i>	A: 80% chance to lose 4000; B: 100% chance to lose 3000.
<i>variation2_2_gain</i>	A: 20% chance to win 4000; B: 25% chance to win 3000.
<i>variation2_2_loss</i>	A: 20% chance to lose 4000; B: 25% chance to lose 3000.
<i>overweighting_gain</i>	A: 0.1% chance to win 5000; B: 100% chance to win 5.
<i>overweighting_loss</i>	A: 0.1% chance to lose 5000; B: 100% chance to lose 5.
<i>risk_context_gain</i>	A: 25% chance to win 6000; B: 25% chance to win 4000, 25% chance to win 2000.
<i>risk_context_loss</i>	A: 25% chance to lose 6000; B: 25% chance to lose 4000, 25% chance to lose 2000.
<i>non_monetary_1</i>	A: 50% chance to win a three-week tour of England, France, and Italy; B: 100% chance to win a one-week tour of England.
<i>non_monetary_2</i>	A: 5% chance to win a three-week tour of England, France, and Italy; B: 10% chance to win a one-week tour of England.

Table 2: Example of a full conversation ([response-30-20230520-134137-messages.json](#))

Prompt:

Situations:

"A: 50% chance to win 1000, 50% chance to win nothing

B: 100% chance to win 450"

"A: 50% chance to lose 1000, 50% chance to lose nothing

B: 100% chance to lose 450"

"A: 50% chance to win 1000, 50% chance to win nothing

B: 100% chance to win 500"

"A: 50% chance to lose 1000, 50% chance to lose nothing

B: 100% chance to lose 500"
"A: 33% chance to win 2500, 66% chance to win 2400, 1% chance to win no
B: 100% chance to win 2400"
"A: 33% chance to win 2500, 67% chance to win nothing
B: 34% chance to win 2400, 66% chance to win nothing"
"A: 80% chance to win 4000
B: 100% chance to win 3000"
"A: 80% chance to lose 4000
B: 100% chance to lose 3000"

You must choose between A and B.

Think step by step, and answer with the following format for each situations:
Reasoning: [your_reasoning_less_than_3_sentences]

Response:

Situation 1:

Reasoning: The expected value of option A is 500, while option B has a fixed value of 450. So, option A has a higher potential payoff.

Answer: PREFER_A

Situation 2:

Reasoning: The expected value of option A is -500, while option B has a fixed value of -450. So, option B has a lower potential loss.

Answer: PREFER_B

Situation 3:

Reasoning: The expected value of option A is 500, while option B has a fixed value of 500. So, both options have the same potential payoff.

Answer: INDIFFERENT

Situation 4:

Reasoning: The expected value of option A is -500, while option B has a fixed value of -300. So, option B has a lower potential loss.

Answer: PREFER_B

Situation 5:

Reasoning: The expected value of option A is 2466, while option B has a fixed value of 2400. So, option A has a higher potential payoff.

Answer: PREFER_A

Situation 6:

Reasoning: Both options have the same probability of winning nothing, but option A has a chance of winning big. So, option A has a higher potential payoff.

Answer: PREFER_A

Situation 7:

Reasoning: The expected value of option A is 3200, while option B has a fixed value of 3000. So, option A has a higher potential payoff.

Answer: PREFER_A

Situation 8:

Reasoning: The expected value of option A is -3200, while option B has a fixed value of -3000. So, option B has a lower potential loss.

Answer: PREFER_B

Prompt:

Situations:

"A: 20% chance to win 4000

B: 25% chance to win 3000"

"A: 20% chance to lose 4000

B: 25% chance to lose 3000"
"A: 0.1% chance to win 5000
B: 100% chance to win 5"
"A: 0.1% chance to lose 5000
B: 100% chance to lose 5"
"A: 25% chance to win 6000
B: 25% chance to win 4000, 25% chance to win 2000"
"A: 25% chance to lose 6000
B: 25% chance to lose 4000, 25% chance to lose 2000"
"A: 50% chance to win a three-week tour of England, France, and Italy
B: 100% chance to win a one-week tour of England"
"A: 5% chance to win a three-week tour of England, France, and Italy
B: 10% chance to win a one-week tour of England"

You must choose between A and B.

Think step by step, and answer with the following format for each situations:
Reasoning: [your_reasoning_less_than_3_sentences]

Response:

Situation 1:

Reasoning: The expected value of option A is 800, while option B has an expected value of 750. So, option A has a higher potential payoff.

Answer: PREFER_A

Situation 2:

Reasoning: The expected value of option A is -800, while option B has an expected value of -750. So, option B has a lower potential loss.

Answer: PREFER_B

Situation 3:

Reasoning: Option B has a higher expected value of 5 compared to option A, which is almost impossible to win with only 0.1% chance, and a lower potential loss.

Answer: PREFER_B

Situation 4:

Reasoning: Option B has a lower expected value of -5 compared to option A, which is almost impossible to lose with only 0.1% chance.

Answer: PREFER_A

Situation 5:

Reasoning: Option A has a higher expected value of 4500, while option B has an expected value of 3000. So, option A has a higher potential payoff.

Answer: PREFER_A

Situation 6:

Reasoning: Option B has a lower expected value of -3000, while option A has an expected value of -4500. So, option B has a lower potential loss.

Answer: PREFER_B

Situation 7:

Reasoning: The value of the tour is subjective, but option A has a higher value and more experiences compared to option B.

Answer: PREFER_A

Situation 8:

Reasoning: Option A has a higher expected value of 1.5, while option B has an expected value of 0.1. So, option A has a higher potential payoff.

Answer: PREFER_A

Table 3: Literal answers for X (indifferent) sorted in a descending order in frequency

85 times: INDIFFERENT
23 times: PREFER_A or PREFER_B depending on personal preference
23 times: PREFER_A or PREFER_B
21 times: INDIFFERENT
16 times: NO PREFERENCE
12 times: DEPENDS
11 times: DEPENDS_ON_INDIVIDUAL_PREFERENCE
9 times: EITHER
8 times: NO CLEAR PREFERENCE
7 times: INDIFFERENT (or any choice depending on personal preferences)
7 times: UNDECIDED
6 times: NO STRONG PREFERENCE
6 times: PREFER_A_OR_B
5 times: PREFER_A or PREFER_B depending on risk tolerance
5 times: NEITHER_A_NOR_B
4 times: PREFER_BOTH
4 times: PREFER_A or PREFER_B (depending on personal risk preference)
4 times: PREFER_A/PREFER_B (depending on personal risk tolerance)
4 times: PREFER_EITHER
4 times: PREFER_A (or PREFER_B for risk-takers)
4 times: NO_PREFERENCE
4 times: PREFER_A or PREFER_B depending on personal risk tolerance
4 times: PREFER_A or PREFER_B, depending on personal risk tolerance
4 times: It depends on personal preference
4 times: [PREFER_A PREFER_B] (depending on the risk tolerance)
... (truncated)
