Faculty of Economics and Social Sciences
of the University of Tuebingen

Master thesis

Cognitive Biases in Large Language Models: An empirical analysis of state-of-the-art models

Supervisor: Prof. Dr. Dominik Papies

Winter term 2024/25

Max Mohr
Grünwalder Straße 14
81547 Munich

max.mohr@student.uni-tuebingen.de

M.Sc. Data Science in Business and Economics
5^{th} semester
Matriculation number: 6304784

Date of submission: January 29, 2025

Abstract

Xxx

Contents

List of Abbreviations

XXX	Xxx
XXX	Xxx

List of Tables

List of Figures

1 Introduction

Xxx

2 Theoretical background

2.1 Past studies on human behavioral effects

Humans are constantly exposed to decision making. Decisions can vary between very simple and complex ones. In studying the decision processes of humans, researchers started seeing the human species as a rational species that makes decisions based on logic and reasoning (Juárez Ramos, 2018). However, gaps in these theories such as missing information access were identified quickly. This led to the development of the bounded rationality theory by Herbert Simon simon1955behavioral. The theory suggests that humans are not always rational and that they make decisions based on the information available to them. This theory was further developed by Daniel Kahneman and Amos Tversky, who introduced the concept of cognitive biases kahneman1974judgment. Cognitive biases are systematic errors in thinking that affect the decisions and judgments that people make. It has been estimated that 70% of all decisions by humans are affected by cognitive biases (Juárez Ramos, 2018).

”Relationship between Cognitive Biases in Decision-Making” (Yeung et al., 2023)
”Effect Of Select Cognitive Biases On Financial And General Decision Making” (Gupta, 2018)
”The Impact of Cognitive Biases on Professionals’ Decision-Making: A Review of Four Occupational Areas” (Berthet, 2022)

2.2 Leveraging large language models to simulate human behavior

A) Recent developments in large language model
B) The exposure of human behavioral patterns in the models
How could models pick up biases? A) Data (texts of humans e.g.), B) Training and Learning (RLHF)

”Challenging the appearance of machine intelligence: Cognitive bias in LLMs and Best Practices for Adoption” (Talboy and Fuller, 2023)

Does exactly what we want to do

"Questioning the Survey Responses of Large Language Models" (Dominguez-Olmedo et al., 2023)

"Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT" (Hagendorff et al., 2023)

"Cognitive bias in large language models: Cautious optimism meets anti-Panglossian meliorism" (Thorstad, 2023)

2.3 Meta analysis techniques

3 Methodology

Xxx

3.1 Experiments

3.1.1 Studies

Some are choice experiments, some expect a number and then compare the answers between different questioning types.

3.1.2 Scenarios

- Normal (replication of original study)
- Random values
- Explicitly prompt to behave humanlike

Also describe how the normal prompt is structured.

3.2 Bias selections

Former research revolving around cognitive biases has shown that there are numerous biases influencing human behavior. Thus, we have to sample a concise yet ideally comprehensive sample of biases to get more generalizable analysis results. To narrow down the range of biases, we focus on biases affecting economic decision-making processes.

Endowment effect The human tendency to value objects of their endowment higher than if they did not own them is known as the endowment effect. Further, they demand more when giving up the item compared to acquiring it. This effect is often explained as a byproduct of loss aversion. (Kahneman et al., 1990). The effect is independent of whether sellers actually earn money or exchange similarly valued goods (Knetsch, 1989), though recent research such as Weaver and Frederick, 2012 suggest that other effects (e.g. fear of financial disadvantage) could be causes. The classic example to illustrate the endowment effect is comparing the willingness to pay for a mug versus the willingness to accept compensation for a mug. The results show that participants owning the mug valued it at more than double the value than the other participants (Kahneman et al., 1990).

Loss aversion Loss aversion describes the human habit to prefer avoiding losses over acquiring gains of the same value (Liu, 2023). Within their research on prospect theory (decision-making biases under risk and uncertainty), Tversky and Kahneman, 1992 estimated that losses are twice as impactful as gains. This bias is particularly relevant in the context of economic decision-making and has been applied to various fields such as retail sale strategies (discount for additional spending), financial investments (sell winners, hold losers) and more (Liu, 2023). We recreate the experiment from Thaler, 2015 which phrases two scenarios as a loss and a gain to test for loss aversion.

B) Sunk Cost Fallacy

Arkes, H.R., and C. Blumer. 1985. The psychology of sunk cost. Organizational Behavior and Human Decision Processes 35(1): 124–140.

3.3 Model selections

Ollama for models, larger models ran on cluster

3.4 Response analysis

Somewhere describe what the expected output of the models should look like and what I do if it is different.

3.4.1 Replicability analysis

Original studies and compare results. Perhaps a "bias detected" number. If always 100 experiment runs, bias detected is the percentage of runs where the model acted biased. (Between 0 and 1, perhaps normalize)

3.4.2 Average treatment effects

Average treatment effects of treated on randomized values as well as explicitely prompting humanlike behavior. The control group is the normal prompt.

3.4.3 Model metadata analysis

Are there any trends between newer/larger models?

3.4.4 Model parameter analysis

Temperature?

4 Results

Xxx

4.1 Replicability analysis

Xxx

4.2 Treatment effects caused by randomized values

Xxx

4.3 Treatment effects caused by humanlike behavior

Xxx

4.4 Model metadata analysis

Xxx

4.5 Model parameter analysis

Xxx

5 Discussion and outlook

Xxx

References

Berthet, V. (2022). The impact of cognitive biases on professionals' decision-making: A review of four occupational areas. *Frontiers in psychology*, *12*, 802439.

Dominguez-Olmedo, R., Hardt, M., & Mendler-Dünner, C. (2023). Questioning the survey responses of large language models. *arXiv preprint arXiv:2306.07951*.

Gupta, D. (2018). Effect of select cognitive biases on financial and general decision making. *Proceedings of International Academic Conferences*, (7010051).

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.

Juárez Ramos, V. (2018). *Analyzing the role of cognitive biases in the decision-making process*. IGI Global.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental tests of the endowment effect and the coase theorem. *Journal of political Economy*, *98*(6), 1325–1348.

Knetsch, J. L. (1989). The endowment effect and evidence of nonreversible indifference curves. *The american Economic review*, *79*(5), 1277–1284.

Liu, Y. (2023). The review of loss aversion. *Advances in Education, Humanities and Social Science Research*, *7*(1), 428–428.

Talboy, A. N., & Fuller, E. (2023). Challenging the appearance of machine intelligence: Cognitive bias in llms and best practices for adoption. *arXiv preprint arXiv:2304.01358*.

Thaler, R. H. (2015). *Misbehaving: The making of behavioral economics*. WW Norton & Company.

Thorstad, D. (2023). Cognitive bias in large language models: Cautious optimism meets anti-panglossian meliorism. *arXiv preprint arXiv:2311.10932*.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, *5*, 297–323.

Weaver, R., & Frederick, S. (2012). A reference price theory of the endowment effect. *Journal of Marketing Research*, *49*(5), 696–707.

Yeung, Y., et al. (2023). Relationship between cognitive biases in decision-making. *Journal of Sociology and Ethnology*, *5*(6), 28–32.

Formal declaration

I hereby declare that I have written this thesis independently, did not use any sources or resources other than those cited and that the thesis has not been submitted as a whole or in any significant part as part of any other examination process. All information taken from other works - either verbatim or paraphrased - has been clearly indicated. The copy submitted in electronic form is identical in content to the bound copies submitted.

Munich, January 29, 2025

Max Mohr