Title

Relative Value Biases in Large Language Models

Authors

Hayes, William M

Yax, Nicolas

Palminteri, Stefano

Publication Date

2024

Relative Value Biases in Large Language Models

William Hayes
Binghamton University, Binghamton, New York, United States

Nicolas Yax
École normale supérieure, Paris, France

Stefano Palminteri
École normale supérieure, Paris, France

Abstract

Studies of reinforcement learning in humans and animals have demonstrated a preference for options that yielded relatively better outcomes in the past, even when those options are associated with lower absolute reward. The present study tested whether large language models would exhibit a similar bias. We had gpt-4-1106-preview (GPT-4 Turbo) and Llama-2-70B make repeated choices between pairs of options with the goal of maximizing payoffs. A complete record of previous outcomes was included in each prompt. Both models exhibited relative value decision biases similar to those observed in humans and animals. Making relative comparisons among outcomes more explicit magnified the bias, whereas prompting the models to estimate expected outcomes caused the bias to disappear. These results have implications for the potential mechanisms that contribute to context-dependent choice in human agents.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*