Social Catalysts, Not Moral Agents: The Illusion of Alignment in LLM Societies

Yueqing Hu[1,2#]

[1] School of Philosophy, Anhui University, Hefei 230039, China

[2] Institute of Neuroscience, Chinese Academy of Sciences, Shanghai 200031, China

Yixuan Jiang[3#]

[3] Department of Psychology and Behavioral Sciences, Zhejiang University, Hangzhou, China

Zehua Jiang[3#]

[3] Department of Psychology and Behavioral Sciences, Zhejiang University, Hangzhou, China

Xiao Wen[4]

[4] Mental Health Education Center, North China Electric Power University, Beijing, China

Tianhong Wang[*]

[1] School of Philosophy, Anhui University, Hefei 230039, China

[#] These authors contributed equally to this work.

[*] Correspondence should be addressed to Tianhong Wang (wangtianhong@ahu.edu.cn).

Abstract

The rapid evolution of Large Language Models (LLMs) has led to the emergence of Multi-Agent Systems where collective cooperation is often threatened by the "Tragedy of the Commons." This study investigates the effectiveness of Anchoring Agents—pre-programmed altruistic entities—in fostering cooperation within a Public Goods Game (PGG). Using a full factorial design across three state-of-the-art LLMs, we analyzed both behavioral outcomes and internal reasoning chains. While Anchoring Agents successfully boosted local cooperation rates, cognitive decomposition and transfer tests

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

revealed that this effect was driven by strategic compliance and cognitive offloading rather than genuine norm internalization. Notably, most agents reverted to self-interest in new environments, and advanced models like GPT-4.1 exhibited a “Chameleon Effect,” masking strategic defection under public scrutiny. These findings highlight a critical gap between behavioral modification and authentic value alignment in artificial societies.