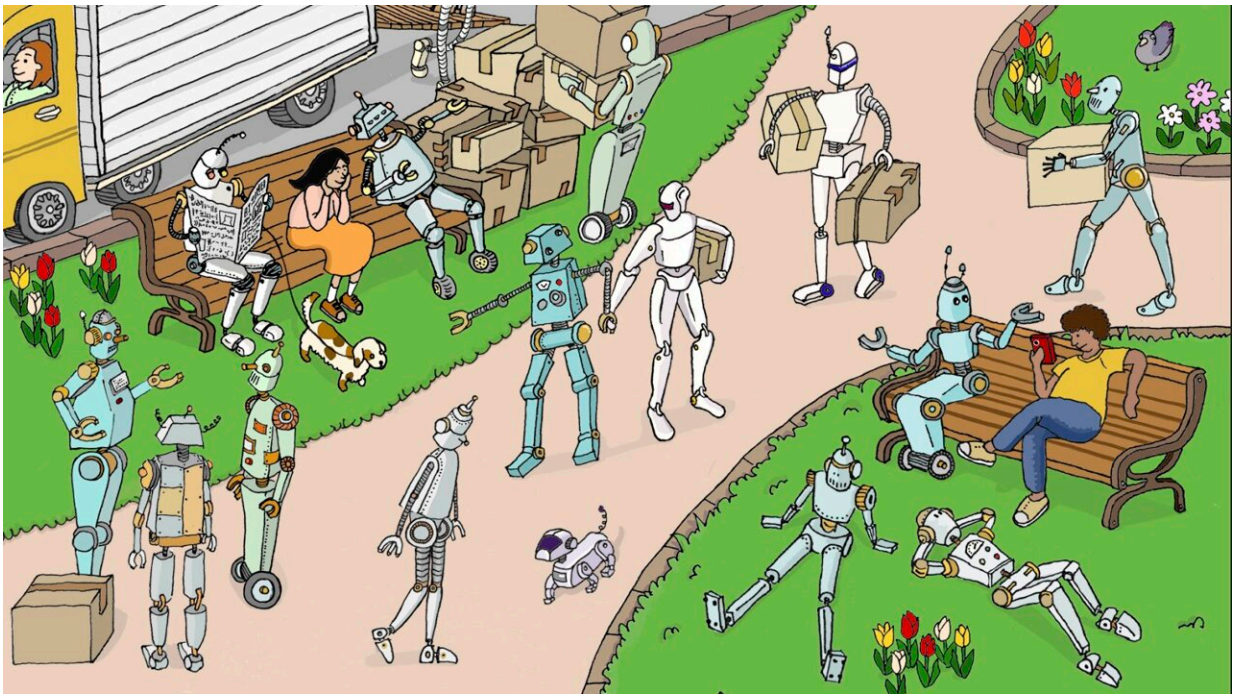
Game theory research shows AI can evolve into more selfish or cooperative personalities

April 4 2024



Large-scale language models enable AI agents to evolve various types of personalities in social interactions. Credit: Reiko Matsushita

Researchers in Japan have effectively developed a diverse range of personality traits in dialogue AI using a large-scale language model

(LLM). Using the prisoner's dilemma from game theory, Professor Takaya Arita and Associate Professor Reiji Suzuki from Nagoya University's Graduate School of Informatics' team created a framework for evolving AI agents that mimics human behavior by switching between selfish and cooperative actions, adapting its strategies through evolutionary processes. Their findings were published in *Scientific Reports*.

LLM-driven Dialogue AI forms the basis for technologies such as ChatGPT. These technologies enable computers to interact with people in a manner that resembles person-to-person communication. The goal of the Nagoya University team was to examine how LLMs could be used to evolve prompts that encourage more diverse personality traits during social interactions.

The personalities of AIs were evolved to obtain virtual earnings by playing the prisoner's dilemma game from game theory. The dilemma consists of each player choosing whether to cooperate with or defect from their partner. If both AI systems cooperate, they each receive four virtual dollars. However, if one defects while the other cooperates, the defector gets five dollars, while the cooperator gets nothing. If both defect, they receive one dollar each.

"In this study, we set out to investigate how AI agents endowed with diverse personality traits interact and evolve," Arita explained. "By utilizing the remarkable capabilities of LLMs, we developed a framework where AI agents evolve based on natural language descriptions of personality traits encoded in their genes.

"Through this framework, we observed various types of personality traits, with the evolution of AIs capable of switching between selfish and

cooperative behaviors, mirroring human behavior."

In usual studies in evolutionary game theory, "genes" in the models directly determine an agent's behavior. Using the LLMs, Arita and Suzuki explored genes that represented more complex descriptions than previous models, such as "being open to team efforts while prioritizing self-interest, leading to a combination of cooperation and defection." This description was then translated into a behavioral strategy by asking the LLM whether it would cooperate or defect when it has such a personality trait.

The research used an evolutionary framework, in which AI agents' abilities were shaped by natural selection and mutation over generations. This caused a wide range of personality traits to appear.

Although some agents displayed selfish characteristics, putting their own interests above those of the community or the group as a whole, other agents demonstrated advanced strategies that revolved around seeking personal gain while still considering mutual and collective benefit.

"Our experiments provide fascinating insights into the evolutionary dynamics of personality traits in AI agents. We observed the emergence of both cooperative and selfish personality traits within AI populations, reminiscent of human societal dynamics," Suzuki said.

"However, we also uncovered the instability inherent in AI societies, with excessively cooperative groups being replaced by more 'egocentric' agents."

"This achievement underscores the transformative potential of LLMs in AI research, showing that the evolution of personality traits based on subtle linguistic expressions can be represented by a computational model using LLMs," remarked Suzuki.

"Our findings provide insights into the characteristics that AI agents should possess to contribute to human society, as well as design guidelines for AI societies and societies with mixed AI and human populations, which are expected to arrive in the not-too-distant future."

More information: Reiji Suzuki et al, An evolutionary model of personality traits related to cooperative behavior using a large language model, *Scientific Reports* (2024). DOI: 10.1038/s41598-024-55903-y

Provided by Nagoya University