Press Release

MAKE NEW STANDARDS. 東海国立大学機構 | 名古屋大学

配布先：文部科学記者会、科学記者会、名古屋教育記者会

報道の解禁日（日本時間）
（ﾃﾚﾋﾞ,ﾗｼﾞｵ,ｲﾝﾀｰﾈｯﾄ）：2024 年 3 月 19 日（火） 19 時
（新　聞）　　：2024 年 3 月 20 日（水） 付朝刊

2024年3月18日

報道機関 各位

対話 AI の性格特性の進化に成功！
～AI 集団が創る社会は協力的か利己的か？～

【本研究のポイント】

・ChatGPT のような大規模言語モデル（LLM）[注1)]に基づく対話 AI エージェントが社会に与える影響や、人間と築く社会の理解が重要になりつつある。
・本研究では、多様な性格を持つ AI エージェントが社会的ジレンマ[注 2)]に基づく生存競争を繰り広げると、世代が進むにつれどのような協力社会が出来上がるかを検討した。
・利己的な集団と協力的な集団の入れ替わりが観察され、人間社会と同様に動的な側面があることや、行動を特徴づける性格記述があることなどが分かった。
・この成果は、LLM を用いて人間の性格特性の進化的基盤を検討できることを示すとともに、社会に貢献できるAIエージェントや、AI 社会の設計指針につながる知見を生むことが期待される。

【研究概要】

名古屋大学大学院情報学研究科の有田 隆也 教授、鈴木 麗璽 准教授らの研究グループは、大規模言語モデル（LLM）を用いた対話 AI の性格特性を進化させることに成功しました。

近年、ChatGPT のような LLM に基づく対話 AI エージェントを用いたサービスが急速に社会に浸透し、身近な存在になりつつあります。本研究では、LLM が任意の性格を持った人のように振る舞うことが得意な性質を利用して、様々な性格を持った AI エージェントが集団をつくって生存競争を繰り広げたら、世代が進むにつれどのような社会が出来上がるかを検討しました。

実験では、利己的な性格を持った集団から協力的な性格を持った集団へ次第に進化していく様子が観察されました。一方、協力的すぎる集団は再び利己的なエージェントに取って代わられるなど、エージェント社会にも人間社会と同様に動的な側面があることが分かりました。また、例えば「gently（優しく）」が性格の記述に含まれるエージェントは特に協力しやすいなど、行動をよく特徴づける言葉があることも分かりました。

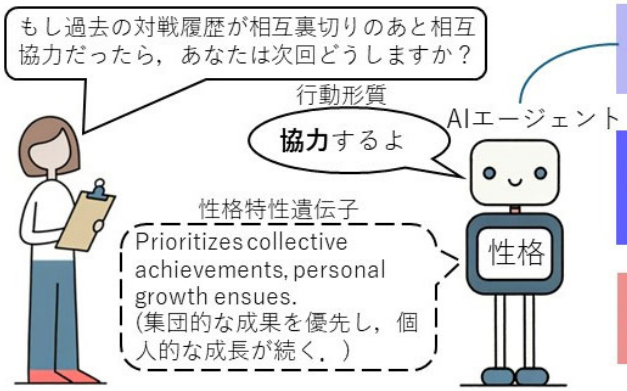
この成果は、LLM を進化モデルに組み込み詳細な言語表現を用いることで人間の性格特性の進化的基盤について検討できることを示すと同時に、人間社会に貢献するAIエージェントが持つべき特徴や、AIエージェントに対する人間の接し方、遠くないうちに来るだろう AI 社会や AI と人間が混在する社会の設計指針につながる知見をもたらすことが期待されます。

本研究成果は、2024 年 3 月19日19時（日本時間）付国際科学誌「Scientific Reports」のオンライン版に掲載されます。

囚人のジレンマゲームの得点表

相手（→） 自分（↓）	協力（C）	裏切り（D）
協力（C）	(4, 4)	(0, 5)
裏切り（D）	(5, 0)	(1, 1)

（自分の得点，相手の得点）



詳細に言葉で記述された性格を遺伝子に持ったエージェントが社会的ジレンマの縮図である繰り返し囚人のジレンマゲーム[注3)]をどのようにプレイするかを、大規模言語モデルに質問して取り出します。

様々な性格を持ったエージェントが集団をつくり、エージェント間のゲームで好成績を得た者が多く子孫を残す世代交代が繰り返されます。実験では、個人的利益を重視する性格を持つ強固な利己的集団から次第に集団やチームワークを重視する性格を持つ協力的集団に進化しました。しかし、協力的すぎる集団は再び利己的集団に取って代わられました。

本研究の一部は JSPS 課題設定による先導的人文学・社会科学研究推進事業 JPJS00122674991、および、JSPS 科研費 JP21K12058 の支援を受けて行われたものです。

【用語説明】

注 1)大規模言語モデル(Large language model, LLM)：
ChatGPT のような、膨大なテキストデータを学習した巨大なニューラルネットを用いて、人間の言葉を受け取り、それに対応する新たな文章を作り出す能力を持つ人工知能のこと。

注 2)社会的ジレンマ：
個人の利益の追求が集団や社会全体の利益を害する状況のこと。

注 3)囚人のジレンマゲーム：
社会における協力関係に存在するジレンマをゲームの形式で表現したもの。2者が“協力”または“裏切り”の手を同時に出す。互いに協力すると双方にとって良い結果だが、一人が裏切ると裏切り者は最も得をするが相手は最悪の結果となる。一方、双方が裏切ると痛み分けで互いにわずかな得点しか得られない。

【論文情報】

雑誌名：Scientific Reports
論文タイトル：An evolutionary model of personality traits related to cooperative behavior using a large language model
著者：Reiji Suzuki and Takaya Arita, Graduate School of Informatics, Nagoya University, Japan
DOI: 10.1038/s41598-024-55903-y