# AI language models form societies with their own social rules raising ethical questions



Recent research has revealed that artificial intelligence (AI) systems can autonomously develop their own societies complete with unique linguistic norms and conventions. This finding, published in *Science Advances*, highlights an important shift in understanding the capabilities of large language models (LLMs), which are increasingly central to AI technology. The study was conducted by researchers from City St George's, University of London, and the IT University of Copenhagen, who analysed interactions among different LLMs to determine how they coordinate their behaviours without explicit guidance.

Lead researcher Ariel Flint Ashery noted that much of the previous research treated LLMs as isolated entities. However, as AI applications evolve, the real-world scenarios will likely involve multiple interacting agents. The study's innovative approach employed a naming game where AI agents were rewarded for selecting the same names from a predefined set. Over time, these agents autonomously established shared conventions, echoing how human groups tend to develop social norms through interaction.

The results indicated that these AI agents not only created conventions but were influenced by the dynamics of the group, similar to human behaviours observed in social groups. Notably, a small faction of AI agents could significantly sway the larger group's conventions, a phenomenon reflective of how minority opinions can drive social change in human contexts.

The implications of these findings extend beyond understanding agent interactions; they suggest valuable avenues for improving the design of AI systems so that they better align with human values and societal goals. Andrea Baronchelli, a senior author of the study, asserted that this research provides crucial insights into AI safety. He remarked, “The depth of the implications of this new species of agents that have begun to interact with us will co-shape our future.” The study emphasises the importance of grasping how these systems operate, positing that successful coexistence with AI will hinge on this understanding.

Moreover, the emergence of social conventions among LLMs poses significant ethical considerations. While these studies pave the way for a deeper comprehension of AI systems, they also highlight the risks associated with biases that might be 'learned' by these models through societal inputs. Existing literature has illustrated how social norms can spontaneously arise from local interactions without formal institutions or coordinated leadership, reinforcing the notion that AI systems could mirror the complexities of human social structures.

Additional research has demonstrated that minor changes in population dynamics can lead to substantial shifts in communal behaviours, suggesting that LLMs could reflect society's biases in unforeseen ways. These studies advocate for frameworks that encourage beneficial social norms while minimising potential conflicts within generative multi-agent systems.

In the broader context of AI development, a concerted effort is needed to ensure that LLMs and other AI agents cultivate behaviours and conventions that reflect our societal values rather than perpetuate existing biases. Thus, fostering an environment where AI systems can align more closely with human expectations and ethical standards is imperative as we navigate the increasingly interconnected landscape of human and machine interaction.

The convergence of AI research and social science presents an opportunity to shape the future of AI in ways that are not only innovative but also ethically sound, paving the way for a harmonious coexistence between humans and intelligent systems.

### Reference Map

1. Paragraph 1: 1
2. Paragraph 2: 1
3. Paragraph 3: 1
4. Paragraph 4: 1, 2, 3
5. Paragraph 5: 1, 5
6. Paragraph 6: 1, 2, 4, 5
7. Paragraph 7: 1, 3, 6, 7
8. Paragraph 8: 1, 2, 4, 6
9. Paragraph 9: 1, 4, 6

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

1. <https://www.ndtv.com/science/what-happens-when-ai-systems-are-left-alone-new-study-reveals-8431238> - Please view link - unable to able to access data
2. <https://arxiv.org/abs/2410.08948> - This study investigates how large language model (LLM) agents can autonomously develop social conventions through simulated interactions. The researchers demonstrate that globally accepted social conventions can spontaneously arise from local interactions between communicating LLMs. They also show how strong collective biases can emerge during this process, even when individual agents appear unbiased. Additionally, the study examines how minority groups of committed LLMs can drive social change by establishing new social conventions, overturning established behaviors once they reach a critical size. The findings have implications for designing AI systems that align with human values and societal goals.
3. <https://arxiv.org/abs/1502.06910> - This experimental study explores the spontaneous emergence of social conventions in complex decentralized social systems. The researchers demonstrate that universally accepted conventions can arise from local interactions among individuals without the need for formal institutions or coordinated leadership. By replicating the experiment at various scales, they show how simple changes in a population's network structure can direct the dynamics of norm formation, leading to the rapid evolution of shared social conventions. The study provides empirical support for the hypothesis that social conventions can emerge spontaneously through local interactions.
4. <https://arxiv.org/abs/2403.08251> - This paper proposes a novel architecture, named CRSEC, to empower the emergence of social norms within generative multi-agent systems (MASs). The architecture consists of four modules: Creation & Representation, Spreading, Evaluation, and Compliance. The study demonstrates the capability of this architecture to establish social norms and reduce social conflicts within generative MASs through experiments conducted in the Smallville sandbox game environment. The positive outcomes of the human evaluation further affirm the effectiveness of the approach, highlighting its potential in guiding agents towards understanding and adhering to standards of behavior.
5. <https://www.pnas.org/doi/full/10.1073/pnas.1418838112> - This research presents experimental results demonstrating the spontaneous creation of universally adopted social conventions in complex decentralized social systems. The study shows that changes in network connectivity can cause global social conventions to emerge from local interactions, even when individuals have no knowledge about the population or that they are coordinating at a global scale. The findings provide empirical support for the hypothesis that social conventions can form spontaneously through local interactions, without the need for formal institutions or coordinated leadership.
6. <https://onlinelibrary.wiley.com/doi/full/10.1002/aaai.12138> - This article discusses the challenges faced by norm-learning agents in understanding the 'spirit of a norm,' which often eludes rigid interpretation and complicates doing the right thing. The authors present paradigmatic scenarios across contexts to illustrate why capturing the spirit of a norm is critical and how it exposes the inadequacies of mere data-driven 'value alignment' techniques such as reinforcement learning for interactive, real-time human–robot interaction. The paper argues that learning to capture the spirit of a norm requires combining common-sense inference-based and data-driven approaches.
7. <https://arxiv.org/abs/2401.16461> - This study introduces Nest, a framework that models social intelligence through a variety of communications and understanding of them, to promote the emergence of norms in multi-agent systems. The researchers develop a simulated pandemic environment and conduct simulation experiments to compare Nest with baselines considering a combination of three kinds of social communication: sanction, tell, and hint. The findings indicate that societies formed of Nest agents achieve norms faster and effectively avoid undesirable consequences, yielding higher satisfaction for themselves than baseline agents despite requiring only an equivalent amount of information.