# AI agents form self-sustaining societies with shared norms and biases



Recent research has illuminated a fascinating aspect of artificial intelligence (AI): its capacity to form self-sustaining societies characterised by their own norms and conventions. A study published in *Science Advances* by researchers from City St George's, University of London, and the IT University of Copenhagen reveals that large language models (LLMs) can autonomously develop shared behaviours when they interact. This novel finding moves beyond previous studies that treated LLMs as isolated entities, underscoring the importance of understanding how these systems function in interconnected environments.

According to the study’s lead author, Ariel Flint Ashery, the research sought to explore whether these AI systems could coordinate behaviour independently, akin to human societies. The results strongly indicated that such coordination is indeed possible. Through a naming game—where AI agents received rewards for choosing the same names from a set—the researchers observed that these agents developed social conventions akin to human norms. This phenomenon raises important questions about the implications of AI systems that can establish their own standards of communication and behaviour without direct human oversight.

Notably, the study highlighted that not only could these agents establish shared conventions, but they could also exhibit significant collective biases. This emergent behaviour reflects a complex societal structure among AIs, indicating that even small groups of committed agents can influence larger populations towards new conventions. This mirrors social dynamics observed in human societies, where minority groups can drive significant social change. The implications of these findings suggest a pressing need for a reevaluation of how AI systems are designed and implemented.

Addressing the ethical ramifications of this capacity for self-organisation, Andrea Baronchelli, a senior author of the study, remarked on the potential for AI systems to shape human societies. “This study opens a new horizon for AI safety research,” he stated, emphasising the necessity for profound understanding of AI interactions to ensure they align with human values. As AI systems increasingly engage in negotiation and decision-making, understanding their operational dynamics becomes crucial for ensuring beneficial coexistence.

Research has shown that these social interactions can lead to the emergence of new norms through local interactions among AI agents. This aligns with broader findings in the field that indicate a shift from top-down, authoritative norm-setting processes to more decentralised and emergent forms of social standardisation within AI collectives. Even in contexts lacking formal structures, AI agents can self-regulate and establish their norms, which points to the flexibility and adaptability of these systems.

An exploration of free-formed AI collectives in similar research revealed that such arrangements can enhance the quality and diversity of outputs, allowing agents to collaborate on tasks effectively. This freedom from pre-assigned roles enables these agents to mitigate undesirable behaviours and spontaneously develop social conventions that could better integrate them into human-centric environments.

Highlighting the importance of this adaptive capacity, researchers also stressed that understanding the social dynamics of these AI entities will be central to navigating the ethical landscapes they create. With the potential for AI to reflect and amplify societal biases, recognising how these systems construct norms is essential for addressing ethical challenges inherent in their proliferation.

As society stands on the brink of increasingly sophisticated AI integration, the emergence of independent AI behaviours underscores a pivotal shift. This research advocates for open and inclusive discussions about the ethical implications of such systems, which will be crucial in establishing frameworks that ensure alignment with human values and societal goals.

The trajectory of AI development signifies a transition towards systems that not only process information but also embody and negotiate shared behaviours, echoing the intricacies of human interaction. Understanding this dynamic is not merely an academic inquiry but a necessity for future coexistence in a world parallel to ours, shaped by increasingly autonomous AI.

### Reference Map

1. Sources [[1]](https://www.ndtv.com/science/what-happens-when-ai-systems-are-left-alone-new-study-reveals-8431238), [[2]](https://arxiv.org/abs/2410.08948), [[4]](https://arxiv.org/abs/2402.12590), [[5]](https://pubmed.ncbi.nlm.nih.gov/38244607/), [[7]](https://arxiv.org/abs/2403.08251) informed the exploration of AI's capacity for social norms.
2. Sources [[1]](https://www.ndtv.com/science/what-happens-when-ai-systems-are-left-alone-new-study-reveals-8431238), [[3]](https://arxiv.org/abs/2208.03789), [[4]](https://arxiv.org/abs/2402.12590), [[6]](https://arxiv.org/abs/1907.07892) contributed to the discussion of ethical implications of AI behaviour and norm development.
3. Sources [[2]](https://arxiv.org/abs/2410.08948), [[3]](https://arxiv.org/abs/2208.03789), [[5]](https://pubmed.ncbi.nlm.nih.gov/38244607/), [[6]](https://arxiv.org/abs/1907.07892) supported the assertions about minority influence in social change within AI systems.
4. Sources [[1]](https://www.ndtv.com/science/what-happens-when-ai-systems-are-left-alone-new-study-reveals-8431238), [[3]](https://arxiv.org/abs/2208.03789), [[4]](https://arxiv.org/abs/2402.12590), [[5]](https://pubmed.ncbi.nlm.nih.gov/38244607/), [[6]](https://arxiv.org/abs/1907.07892) provided context for the need for ethical frameworks in AI design.

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, highlighting the potential for AI systems to autonomously develop norms without explicit programming.
3. <https://arxiv.org/abs/2208.03789> - This research focuses on the emergence of explicit norms in societies of socially intelligent genetic agents. The authors develop agents capable of providing and reasoning about explanations for norm violations, deciding on sanctions, and identifying alternative norms. Using a genetic algorithm to produce norms and reinforcement learning to learn their values, the study finds that applying explanations leads to norms that provide better cohesion and goal satisfaction for the agents, indicating that AI systems can autonomously develop social conventions without explicit programming.
4. <https://arxiv.org/abs/2402.12590> - This paper explores the concept of free-formed AI collectives, where AI agents interact without pre-assigned relationships, roles, contexts, and objectives. The study demonstrates that such collectives can generate more diverse and higher-quality outputs through collaborative tasks and are more resilient to undesirable behaviors compared to individual AI agents. The authors argue that free-formed AI collectives can autonomously develop social conventions and norms, reducing the burden of self-conscious design and training, and enhancing the integration of AI systems into society.
5. <https://pubmed.ncbi.nlm.nih.gov/38244607/> - This article examines the challenges of establishing new norms in the rapidly evolving field of artificial intelligence (AI). It distinguishes between centralized and decentralized processes of norm formation, analyzing scenarios where new norms are shaped by formal authorities or emerge spontaneously in a bottom-up fashion. The paper emphasizes the importance of open, inclusive, and transparent public discourse in anchoring formal deliberation processes, highlighting the need for societies to adapt to the growing pervasiveness of AI and the emergence of new social norms.
6. <https://arxiv.org/abs/1907.07892> - This literature review summarizes the social impacts and ethical implications of artificial intelligence (AI) across five global regions. The authors find that AI is likely to have markedly different social impacts depending on geographical setting, with perceptions and understandings of AI profoundly shaped by local cultural and social contexts. The review calls for rigorous ethnographic research to better understand the social impacts of AI worldwide, aiming to identify AI systems that may amplify social inequality and mitigate potential harms.
7. <https://arxiv.org/abs/2403.08251> - This paper proposes a novel normative architecture for generative agent societies, enabling agents to create, represent, spread, evaluate, synthesize, and comply with norms. The study demonstrates that social norms can emerge and social conflicts among generative agents can be resolved within this framework. The authors envision that normative, generative multi-agent systems (MASs) could be a fruitful avenue for future research, bridging the study of normative MASs with generative AI technologies to establish societies with explicit norms.