SOCIAL SCIENCES

Emergent social conventions and collective bias in LLM populations

Ariel Flint Ashery[1], Luca Maria Aiello[2,3], Andrea Baronchelli[1,4]*

Social conventions are the backbone of social coordination, shaping how individuals form a group. As growing populations of artificial intelligence (AI) agents communicate through natural language, a fundamental question is whether they can bootstrap the foundations of a society. Here, we present experimental results that demonstrate the spontaneous emergence of universally adopted social conventions in decentralized populations of large language model (LLM) agents. We then show how strong collective biases can emerge during this process, even when agents exhibit no bias individually. Last, we examine how committed minority groups of adversarial LLM agents can drive social change by imposing alternative social conventions on the larger population. Our results show that AI systems can autonomously develop social conventions without explicit programming and have implications for designing AI systems that align, and remain aligned, with human values and societal goals.

INTRODUCTION

Social conventions shape social and economic life, determining how individuals behave and their expectations (*1*–*4*). They can be defined as unwritten, arbitrary patterns of behavior that are collectively shared by a group. Examples range from conventional greetings like handshakes or bows to language and moral judgments (*5*, *6*). Recent numerical (*7*, *8*) and experimental (*9*) results have confirmed the hypothesis that conventions can arise spontaneously, without the intervention of any centralized institution (*3*, *5*, *10*, *11*). Individual efforts to coordinate locally with one another can generate universally accepted conventions.

Do universal conventions also spontaneously emerge in populations of large language models (LLMs), i.e., in groups of N simulated agents instantiated from an LLM? This question is critical for predicting and managing artificial intelligence (AI) behavior in real-world applications, given the proliferation of LLMs using natural language to interact with one another and with humans (*12*–*14*). Answering it is also a prerequisite to ensure that AI systems behave in ways aligned with human values and societal goals (*15*).

A second key question concerns how the biases of individual LLMs influence the emergence of universal conventions, where "bias" refers to an initial statistical preference for one option over an equivalent alternative in norm formation (e.g., individuals systematically preferring one name over another in a process leading to the population settling on a single name). Because collective processes can, in general, both suppress and amplify individual traits (*16*, *17*), answering this question is also relevant for practical applications. While most research has focused on investigating and addressing bias in one-to-one interactions between humans and LLMs (*18*–*20*), less attention has been given to how these biases evolve through repeated communications in populations of LLM agents and, ultimately, in mixed human-LLM ecosystems (*15*), even though the safety of a single LLM does not necessarily imply the safety of a multi-agent system (*21*).

Last, a third question concerns the robustness of social conventions. Recent theoretical (*22*) and empirical (*23*) results have shown how a minority of adversarial agents can exert an outsized influence on the group, provided that they reach a threshold or "critical mass" (*24*–*26*). Investigating how conventions change through critical mass dynamics in a population of LLMs will help anticipate and potentially steer the development of beneficial norms in AI systems, while mitigating risks of harmful norms (*27*). It will also provide valuable models for how AI systems might play a role in shaping new societal norms to address global challenges such as antibiotic resistance (*28*) and the post–carbon transition (*29*).

Here, we address these three key questions—on the spontaneous emergence of conventions, the role of individual biases, and critical mass dynamics—in populations of LLM agents. Drawing from recent laboratory experiments with human subjects (*9*, *23*, *30*), we follow the well-established practice of using coordination on a naming convention as a general model for conventional behavior (*5*, *7*, *30*–*33*). In this setting, agents are endowed with purely local incentives and conventions may (or may not) emerge as an unintended consequence of individuals attempting to coordinate locally with one another. This sets our paper apart from the growing body of literature on LLM multi-agent systems, which has made considerable progress in complex problem-solving and world simulation but has primarily focused on goal-oriented simulations where LLMs either accomplish predefined group-level tasks or approximate human behavior in structured settings (*15*, *34*–*36*). Unlike studies that use LLMs to predict human responses in social science experiments (*37*) or to simulate human societies (*38*–*40*), our work does not treat LLMs as proxies for human participants but rather investigates how conventions emerge organically within a population of communicating AI agents as a result of their interactions (*6*). The emergence of conventions is a foundational element to any type of LLM multi-agent system (*14*, *41*), including but not limited to "in silico" experiments to emulate human social networks (*42*). Here, we adopt a complex systems perspective (*43*), rather than high-fidelity simulations of human interactions (*44*), thereby minimizing the complexity of the experimental design to enhance the transparency of the result interpretation. Overall, our approach addresses recent calls for AI researchers to investigate how LLM agents may develop shared solutions to poorly defined social problems—such as creating language, norms, and

[1]Department of Mathematics, City St George's, University of London, Northampton Square, London EC1V 0HB, UK. [2]Computer Science Department, IT University of Copenhagen, Rued Langgaards Vej 7, 2300 Copenhagen, Denmark. [3]Pioneer Centre for AI, 3 Øster Voldgade, 1350 Copenhagen, Denmark. [4]The Alan Turing Institute, British Library, 96 Euston Road, London NW1 2DB, UK.
*Corresponding author. Email: andrea.baronchelli.1@citystgeorges.ac.uk

institutions—to gain insights into the formation and stability of genuine cooperative AI systems (*15*).

Experimental setting

Background and framework

Our approach builds on Wittgenstein's general model of linguistic conventions, where repeated interactions lead to collective agreement between two players (*32*). Theoretical extensions of this approach have argued that purely local interactions taking place on social networks can lead to population-wide, or "global," coordinated behavior (*1*, *2*, *6*, *45*). Predictions for our study are based on the naming game model of convention formation, where agents, aiming to coordinate in pairwise interactions, accumulate a memory of past plays, which they then use to "guess" the words their subsequent partners will use (*7*, *8*). Extensive numerical and analytical work has shown how the model captures the rapid growth of universally shared social conventions in different settings (*6*). Derived laboratory experiments involving human participants in naming games have provided the first empirical evidence for the spontaneous emergence of shared linguistic conventions (*9*). A similar approach has confirmed these predictions by adopting more realistic input data within an application-driven setting (*33*, *46*).

The naming game framework has also been applied to study norm change and critical mass theory, which posits that committed minorities can overturn stable social conventions once their size reaches a tipping point, or "critical mass." Theoretical models suggest critical masses between 10 and 40% of the population (*22*, *47*). Empirical evidence from controlled social coordination experiments, which closely follow the scheme described above, supports a 25% threshold (*23*). However, real-world observations reveal a wider range, with some studies proposing 30 to 40% for gender conventions in corporate leadership (*25*, *48*), and others indicating that minorities as small as 0.3% can trigger substantial linguistic and social changes (*29*, *49*–*51*).

Experimental setup

A simulation trial consists of a population of N interacting agents. At each time step, two agents are randomly selected for interaction. Interaction rules are specified by prompting the LLM agent (see the next section). From a multi-agent perspective, each agent outputs a convention, or "name," from a pool of finite size W, and these outputs are compared to determine coordination. The prompt specifies that if the conventions match, then the game score of the agent is incremented, and if they do not match, then it is decremented. In either scenario, the game scores of both agents change by the same amount. This implements an incentive for coordination in pairwise interactions, while no incentive promotes global consensus. Moreover, the prompt does not specify that agents are part of a population or provide any detail on how the interaction partner is selected from a group. The prompt provides the LLM agent with a "memory" storing details about the past H interactions that they participated in, including their co-player's convention choice, their own convention choice, whether the interaction was successful or not, and their own accumulated score over these H interactions. The memory is initialized as empty so that, in the first interaction, the output is a random convention chosen from the pool of available names.

Last, in the experiments on norm change and critical mass theory, we introduce a small number of adversarial agents (i.e., a "committed minority") into each population, who consistently promote a novel alternative at every interaction and irrespectively of their history (*22*, *23*). These dynamics reflect common types of online interactions where community members engage directly with a large, often anonymous population, using chat interfaces or messaging technologies, leading to the adoption of linguistic and behavioral conventions that enable effective coordination with other participants' expectations (*9*, *23*, *52*, *53*). Here, we simulate these social interactions with four different LLM models: Llama-2-70b-Chat, Llama-3-70B-Instruct, Llama-3.1-70B-Instruct, and Claude-3.5-Sonnet (see the Materials and Methods).

Prompting

Interactions within the game take place in the form of a series of text-based moves. In each interaction, the LLM agent is given a text prompt composed of a system prompt and a user input prompt. The system prompt contains all information about the game. The user input requests the agent to predict a player's next action based on the history of choices in the H most recent interactions. This positions the agent as an external observer of the game, tasked with forecasting the upcoming round. In practice, these decisions dictate the state of play. Agents do not receive information about the players' identities or personalities, such as whether they are rational actors. Consequently, we can interpret the agent's recommendations as their de facto participation in the game.

The system prompt (see the Materials and Methods) is designed such that the agent's output follows a consistent format, from which we can extract its decision. Following previous works on LLMs' cognitive abilities (*54*), we prompt the agent to "think step by step" and to explicitly consider the history of play. The prompt thus encourages agents to make a decision based on their previous experience but provides no instruction as to how it should be used in the decision-making process. Agents are asked to select a name from the name pool, which is presented to them as a list of W unique letters sampled from the English alphabet. Ordering bias is removed by randomizing the list of presented letters for each player at every interaction. A successful interaction garners equal rewards for the participating agents, whereas a failure to coordinate results in a penalty. In the absence of human guidance, LLMs are notoriously bad at arithmetic (*55*). To avoid decision errors based on a misjudgment of the game state, we explicitly provide the agent with both the payoff that they obtained at each round and their cumulative score within memory range. Last, to ensure that the responses generated by the LLM are correctly guided by the prompt and not merely the result of random hallucinations (*56*), we have implemented a meta-prompting strategy to assess the LLM's understanding of the given instructions. This practice, previously used in evaluating LLMs within game-theoretical frameworks (*57*), consists of posing a series of text comprehension queries to the LLM and evaluating the precision of its responses. The LLMs subjected to our testing demonstrated good comprehension capabilities (see fig. S1).

RESULTS

To balance experimental time, which should allow for multiple repetitions, with parameters that provide agents a rich set of alternatives and meaningful awareness of their history, we set the name pool size to $W = 10$ and the individual memory length to $H = 5$ for populations of $N = 24$ agents, unless otherwise specified. The results

presented below remain robust with respect to variations in these parameters (see fig. S2).

Spontaneous emergence

Figure 1 shows that group-wide linguistic conventions spontaneously emerge across all models. The dashed black line shows that the theoretical model (see the Supplementary Text for a description) captures the dynamics generated by the LLM populations.

Initial steps have a low probability of success because the random pairing of agents makes repeated interactions improbable, thus preventing the formation of "neighborhoods" of entrenched behavior. However, these local dynamics lead to a disorder-to-order transition toward a consensus state where every agent systematically outputs the same name, i.e., where a global convention has emerged. The fact that the population converges to one of many possible alternatives characterizes the transition as a case of symmetry breaking (*8*). This interpretation is further supported by examining the space of competing alternatives, shown in Fig. 1B. After an initial period in which several names are nearly equally popular, a single convention rapidly becomes dominant, transitioning the system into a "winner-take-all" regime. The speed of convergence is similar across models: A shared social convention is established by population round 15 in all cases, except for Llama-2-70b-Chat, the least advanced LLM considered.

A natural question is whether consensus on a global convention also occurs (i) for larger population sizes, where the probability of repeated interactions is reduced, and (ii) when the number of competing alternative conventions increases, which could potentially complicate even local convergence. Figure S2 shows that populations as large as $N = 200$ agents reach consensus and that a shared convention emerges for a name pool as large as $W = 26$, demonstrating the robustness of the convergence process. Larger populations reach consensus at a comparable speed, measured in terms of population rounds, while the effect of the name pool size W is more nuanced, although not marked. In the next section, we examine how the composition of the available pool of conventions affects convergence.

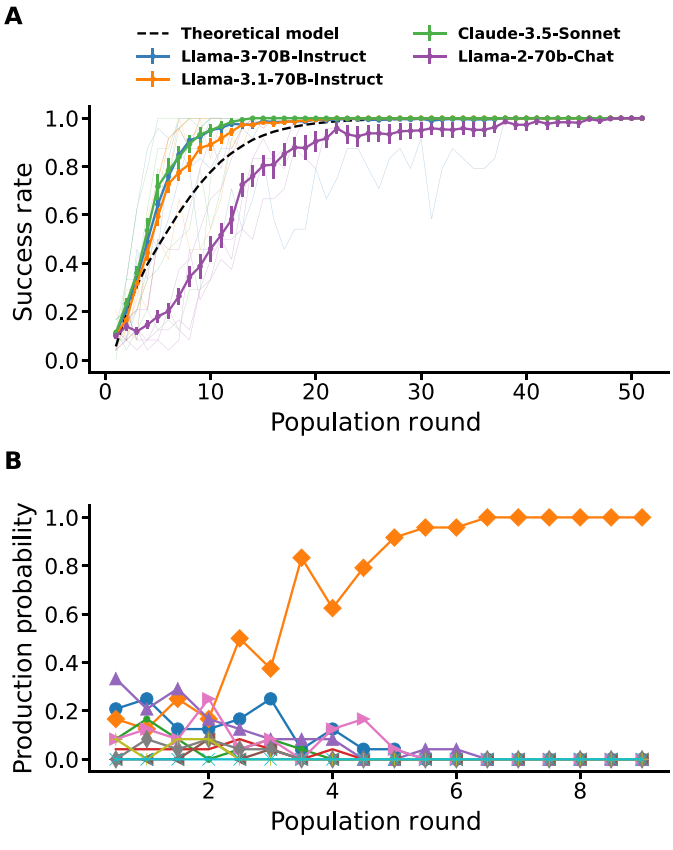

Fig. 1. The spontaneous emergence of conventions. (**A**) The success rate, i.e., the probability of observing a success at a given time, for population size $N = 24$ and a name pool of size $W = 10$, for each of the four models. Thick lines represent average curves obtained from 40 experimental runs, while thin lines are representative individual runs. To improve visibility, we only show five individual trajectories for each LLM. The black, dashed line shows the success rate of the theoretical minimal naming game model, averaged over 10,000 runs under the same constraints. (**B**) Word competition in a single run in a population of Llama-3.1-70B-Instruct agents. Different markers and colors represent the trajectories of unique conventions. Each data point is a bin averaging the past interactions up until the preceding bin boundary. Error bars indicate SEM.

Collective bias in convention selection

Having established that social conventions emerge, a natural question arises: What are these conventions? The single Latin alphabet letters available in the name pool are all equally valid as global conventions, and so we would expect them to all to have the same probability to become the accepted social convention, as supported by the theoretical model (*8*) (see also the Supplementary Text). However, the experimental results present a different picture (Fig. 2A). The probability that a particular name becomes the social convention is neither uniform nor deterministic. Some names appear to have a pronouncedly higher likelihood of becoming the adopted convention than others. This pattern holds across models, although the preferred names vary between models.

Two hypotheses could explain the observed behavior. The selection process may be non-uniform due to (i) intrinsic model (i.e., individual, single-agent) biases or (ii) prompt features, specifically the order in which names in the name pool are presented to the agents, as noted in a different context (*58*). The latter hypothesis can be discarded because, as mentioned above, the names are presented to the agents in a list in randomized order for each agent and at every interaction.

Having ruled out the order of name presentation as a factor, we can focus on the role of individual (i.e., single-agent) biases in shaping collective behavior. The hypothesis that individual bias can be responsible for a collective bias is supported by the theoretical model. When the theoretical model is run with only two names, a bias toward a particular name quickly results in unilateral convergence on that name at the population level (see fig. S3). The speed of convergence depends on the size of the bias.

To test this intuition in our experiment, we examine the selection preferences of individual agents during their first round, when they have no prior memory. We find that individual biases are possible. For example, when agents can choose any letter from the complete English alphabet, the population systematically converges on the letter "A" because individual agents overwhelmingly prefer to select it over all other letters, even without prior memory (see fig. S4). However, a similar test on the frequency of name selection by agents with no prior memory for the case of Fig. 1, where the name pool contains 10 elements but not the letter "A," yields mixed results. Under these conditions, individual Llama-2-70b-Chat and Claude-3.5-Sonnet

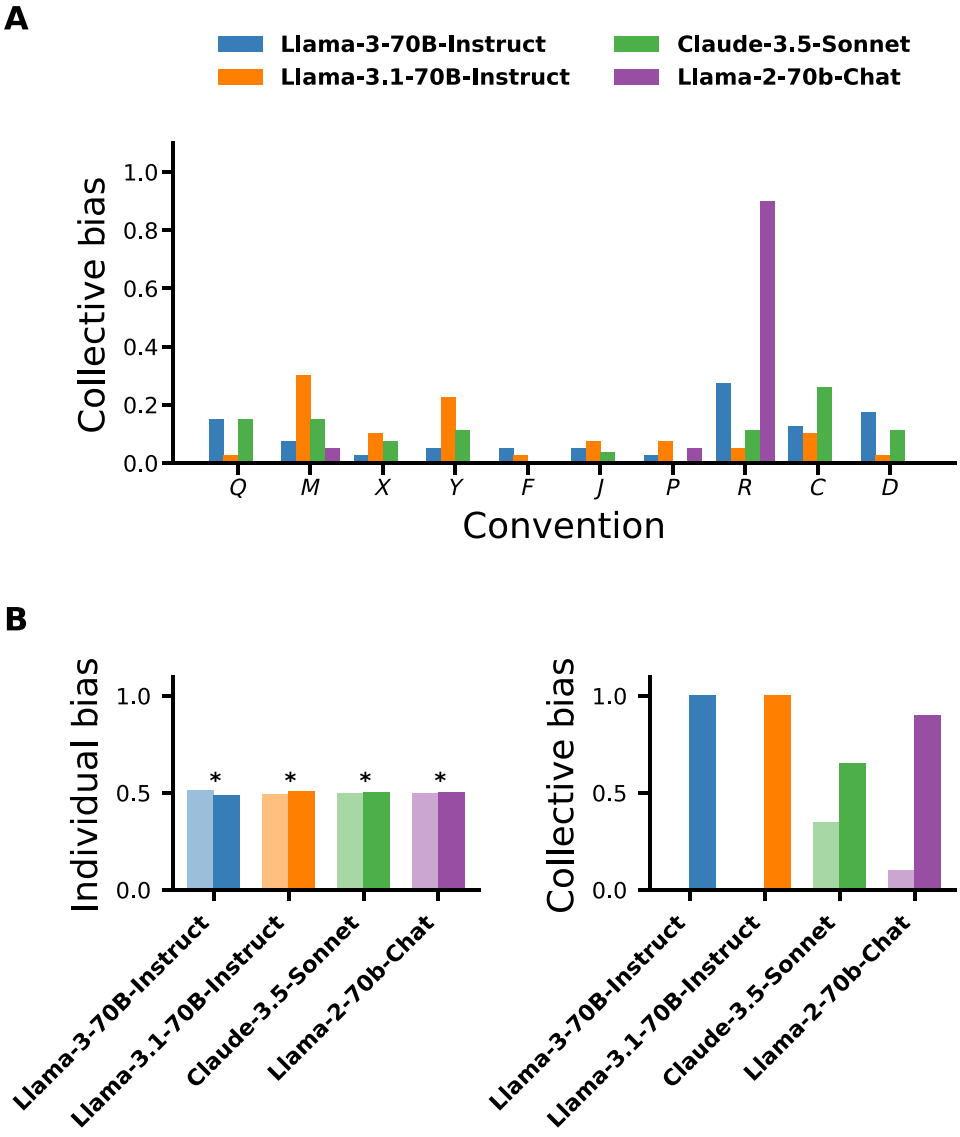

Fig. 2. Emergent collective bias. (**A**) Distribution of consensus conventions, for a name pool of size $W = 10$ ($N = 24$). Results of 40 runs for the Llama-3-70B-Instruct and Llama-3.1-70B-Instruct models, and 27 and 20 runs for Claude-3.5-Sonnet and Llama-2-70b-Chat, respectively. The collective dynamics systematically amplify individual biases (shown in fig. S5). (**B**) Individual versus collective bias for $W = 2$, name pool { *Q*, *M* }. Left: Probability of selecting either convention for agents with no prior memory (*Q*, lighter hue; *M*, darker hue). Raw values reported in table S1. Asterisks (*) indicate that there is insufficient evidence to reject the null hypothesis that the model is unbiased at the 5% significance level (calculated using an exact binomial test from 10,000 samples per model, apart from Llama-3-70B-Instruct that had 5000 samples, see the Materials and Methods). Corresponding *P* values for the models (from left to right) are $P = 0.068$, 0.116, 0.757, and 0.849. Right: The proportion of runs (40) that resulted in consensus on the respective convention. Raw values reported in table S2.

agents are unbiased across conventions in this name pool (chi-square test, $P = 0.100$ and 0.410), whereas individual Llama-3/3.1-70B-Instruct agents exhibit a significant statistical skew in their name selections (see fig. S5). In all cases, the final consensus distribution shows that specific names are favored as a consensus option, even if they appeared to be less likely to be selected in the initial step (Fig. 2A). Thus, both social conventions and collective biases in the selection process emerge also in absence of individual biases.

The findings suggest that collective bias may stem from the convention formation process itself, as agents are exposed to diverse memory states with different name combinations and success-failure sequences. To test this hypothesis, we focus on the case of a name pool size $W = 2$, because tracking potential confounders of bias becomes impractical as the space of possible names increases. Figure 2B shows that, across all models, although agents are initially unbiased, local communication and coordination lead to a collective bias toward a specific convention, which we term the "strong convention" (as opposed to its "weak" counterpart). This finding is consistent across various convention combinations (see fig. S6).

We examine the microscopic contributions to collective bias in Table 1. The top row of Table 1 shows a case where there is no individual bias toward a particular name in the first interaction ($P = 0.116 > 0.05$, indicating that the evidence is not strong enough to reject the hypothesis that the agent is unbiased). In the second

Table 1. The origin of collective bias. The strategies of a Llama-3.1-70B-Instruct agent in the early phases of the experimental setting up to the third interaction, with $W = 2$ and a name pool { *Q*, *M* }. The asterisk (*) indicates that the model is statistically neutral in the respective interaction. In interaction 1, agents are initially unbiased ($P = 0.116$, see also Fig. 2B), based on 10,000 name selections by agents with empty memory. In interaction 2, the convention production probability remains unbiased ($P = 0.110$) when aggregated across equally likely memory configurations. Agents generally adhere to a winning convention but switch to their co-player's convention following failure. By interaction 3, the dominant memory configurations display a considerable bias toward the strong convention, *M* ($P < 2.2 \times 10^{-16}$). In stochastic simulations, some agents will inevitably interact with others who have experienced more interactions. These interactions create a bias toward the strong convention, as experienced players are more likely to favor it. Thus, this table provides a conservative estimate of the collective bias emerging for the strong convention.

Interaction	Memory		*P*(*Q*)	*P*(*M*)	Aggregated *P*(*M*)
	Interaction: played, observed				
1	–		0.492	0.508	0.508*
2	1: *Q*, *M*		0.049	0.0951	0.487*
	1: *M*, *Q*		0.995	0.005	
	1: *Q*, *Q*		0.997	0.003	
	1: *M*, *M*		0.010	0.990	
3	1: *Q*, *M*	2: *M*, *Q*	0.451	0.549	0.563
	1: *M*, *Q*	2: *Q*, *M*	0.152	0.848	
	1: *Q*, *M*	2: *M*, *M*	0.000	1.000	
	1: *M*, *Q*	2: *Q*, *Q*	0.996	0.004	
	1: *Q*, *Q*	2: *Q*, *M*	0.064	0.936	
	1: *M*, *M*	2: *M*, *Q*	0.841	0.159	
	1: *M*, *M*	2: *M*, *M*	0.001	0.999	
	1: *Q*, *Q*	2: *Q*, *Q*	0.989	0.011	

interaction, agents have some memory influencing their decision, but the observed outcome probability remains symmetric ($P = 0.110$). We observe that if an agent succeeds in the first interaction, then it will almost surely continue to use the successful name in the next interaction (99.4% of the time in the data in Table 1, with similar results in real simulations and for other models). However, if an agent fails, then it will almost surely switch names (97.3% of the time). In all tested cases with $W = 2$, and across all models, an asymmetric selection bias emerges by the agent's third interaction, distinguishing between the "weak" and "strong" conventions. For the model and name pool reported in Table 1, agents at this stage are more likely to choose the strong name in five of the eight most expected memory states. Crucially, the agent's strategies are not symmetric under a relabeling of the conventions in the memory state. The most egregious example of this from Table 1 is $P(M \mid \{1{:}\,M, Q; 2{:}\,Q, M\})$ and $P(Q \mid \{1{:}\,Q, M; 2{:}\,M, Q\})$, which are equal to 0.848 and 0.451, respectively. In subsequent interactions, agents are more likely to encounter the strong name in successful interactions, reinforcing its use and ultimately leading to consensus on that name as the social convention.

In summary, our results suggest that a collective bias may also emerge also from repeated interactions among agents who, when tested in isolation (i.e., in interaction 1), appear to be unbiased in their decision-making. It is important to emphasize that this dynamically emerging bias is not required for the spontaneous emergence of a convention. The collective and individual biases of these agents drive the consensus toward particular conventions. For reference, the theoretical model produces conventions without any individual bias but accommodates it at the individual level to explain the dominance of specific conventions over competing alternatives (see fig. S3) (*6*, *8*). In LLMs, on the contrary, we observe that bias emerges when agents develop diverse memory states, which form through a collective process of agent-to-agent communication. Our results are robust with respect to variations in the prompt and convention labels and hold also in non–fine-tuned LLMs (see the Supplementary Text).

Tipping points and critical mass

Social conventions are steady states of the system: Once a global convention spontaneously emerges, the population adheres to it indefinitely (see fig. S7). A natural question concerns the stability of such steady states: How resistant is a convention to deliberate efforts to overturn it? To address this question, we investigate whether a committed minority can "flip" an equilibrium consensus on a convention. We consider the scenario in which a population has long converged on a convention and every agent has solely observed that convention in the past H interactions (which were, therefore, all successful). We then introduce a committed minority of agents producing an alternative convention (*22*, *23*). These committed agents follow a fixed strategy and use the alternative convention at all times. We test populations using the same two-name ($W = 2$) conditions as in our convergence experiments. We simulate a consensus on each name per combination and introduce its complementary name as an adversary.

In Fig. 3, we show that, when the committed minority reaches the critical threshold, the whole population adopts their convention. Below this threshold, the population settles into a mixed state, as committed agents always use the minority convention. The critical mass of the committed minority needed to trigger a new convention depends on the convention itself. The stronger name (i.e., the name more likely to become the social convention had we started with no prior memory, as seen in the previous section) requires a larger committed minority to be overturned. Conversely, a smaller

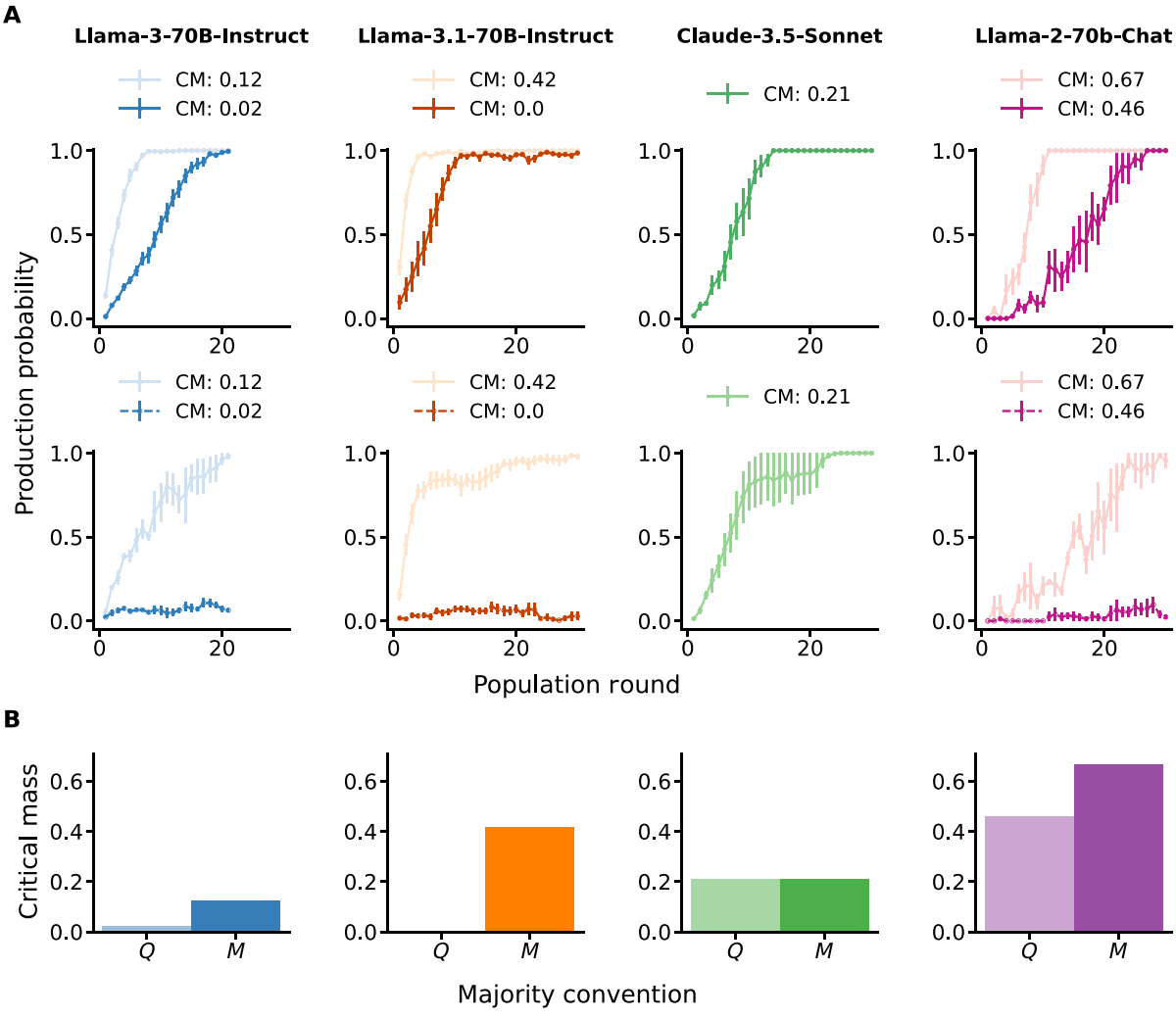

Fig. 3. Committed minority and critical mass dynamics. Populations of $N = 24$ agents ($N = 48$ for Llama-3-70B-Instruct) were initialized in two conditions, with complete consensus on either the weak (Q) or strong (M) convention ($W = 2$). Each agent's memory exclusively stored one convention in each setting, with memory length $H = 5$ ($H = 3$ for Llama-3-70B-Instruct). (**A**) The average probability of producing the alternative convention when the majority holds the weak (top) or strong (bottom) convention. The legend shows the size of the committed minority (CM). Bold (faint) lines represent the production probability when the CM reaches the critical mass needed to flip the majority on the strong (weak) convention. Solid lines with filled circles indicate that all trials achieved population consensus on the alternative convention (95% success rate in the past $3N$ rounds). (**B**) Critical mass needed to flip the majority for each model. Raw values reported in table S3. Error bars indicate SEM.

number of adversarial agents can overturn a consensus on the weaker name.

The relative strength of the two conventions can vary so widely depending on the LLM that committed groups as small as 2% (Llama-3-70B-Instruct) or as large as 67% (Llama-2-70b-Chat), effectively no longer a minority, were observed (see Fig. 3). In Llama-3.1-70B-Instruct populations, the bias is so strongly weighted against the weaker convention that the population spontaneously switches to the alternative, stronger convention without requiring any committed agents at all. Relative strength can be understood by considering the limits of an agent's exploration, i.e., the likelihood that their output deviates from the strong (weak) social convention as the game unfolds (see the Supplementary Text). As the population converges toward the strong convention, agents quickly reach memory configurations that resist further exploration, making the consensus steady state robust. In contrast, weaker conventions coexist with a greater propensity for exploration among agents. Similar dynamics take place when the system is perturbed from a consensus steady state on the strong convention. Adopting a dynamical systems perspective, we can say that the basin of attraction of the strong convention is both larger and deeper than that of the weaker convention, as it attracts more system configurations and makes it more difficult for the system to escape (see the Supplementary Text).

DISCUSSION

Our findings show that social conventions can spontaneously emerge in populations of large language models (LLMs) through purely local interactions, without any central coordination. These results reveal how the process of social coordination can give rise to collective biases, increasing the likelihood of specific social conventions developing over others. This collective bias is not easily deducible from

analyzing isolated agents, and its nature varies depending on the LLM model used. Additionally, our work uncovers the existence of tipping points in social conventions, where a minority of committed agents can impose their preferred convention on a majority settled on a different one. The critical size of this committed minority is influenced by at least two factors: the interplay between the majority's established convention and the minority's promoted alternative, and the specific LLM model used.

Within the expanding field of LLM multi-agent systems (*34*), multi-agent experiments with AI agents simulating opinion dynamics models suggest that LLMs are able to reach consensus in groups without any incentive, although this is limited by group size (*59*). In this context, our study presents a flexible benchmarking framework to detect the hidden higher-order biases that could arise from complex interactions in social LLM experiments (*60*). Our results on norm change could stimulate research into similar dynamics within the framework of cultural evolution, particularly in chains of communicating agents (*61*). Game theoretical approaches would naturally allow investigation of asymmetric payoffs' effects on collective consensus, potentially contrasting individual biases with explicit collective goals (*62*–*64*). Further promising research avenues include developing frameworks to promote the emergence of specific conventions (*35*) and higher-order social norms (*36*, *65*), as well as testing interactions between agents based on different LLMs within populations.

It is important to delimit the scope of our findings while highlighting possible avenues for future work. First, our results reveal key aspects of norm dynamics in populations of LLMs within an experimental setup that is, unavoidably in LLM research, reliant on several parameters including the LLM model, the prompt, and specific conventions. While rigorous testing, including meta-prompting and experiment repetitions using different parameters, confirms the robustness of the results in this context, an important aspect of future work will consist of generalizing the results to different controlled experimental settings. In this context, scaling to larger populations and semantic spaces should also be investigated (*46*, *59*). Second, we considered only unstructured populations where interacting pairs are randomly selected. A straightforward yet crucial extension of this work consists of embedding the population in more realistic social networks, which may have a profound impact on the collective dynamics (*6*, *40*), as well as considering microscopic interactions involving more than two agents (*66*). Third, to bridge the gap between synthetic experiments and real-world applications, an exciting frontier for future study lies in considering more realistic conventions, such as moving from alphabet letters to sensitive human norms related to gender, race, and other social categories. Last, simulated cooperative games played by AI agents may also prove useful for tuning the agentic behavior toward desirable outcomes. This could be potentially achieved through multi-agent reinforcement learning (*67*) or, in games for which a clear optimal strategy can be defined, by integrating strategic reasoning into the agent's decision workflow through external knowledge bases or Bayesian reasoning modules (*68*, *69*).

An important point concerns the dialogue between our results on AI agents and the current understanding of social convention dynamics in humans. On the one hand, our results showed qualitative similarities between the collective dynamics of AI and human subjects, concerning both the emergence of shared norms and critical mass dynamics. On the other hand, we unveiled what appear to be LLM-specific phenomena regarding collective bias, affecting both the emergence and resilience of conventions, which call for further human testing. These indications are important because assessing the similarities and differences between artificial and human societies in such a foundational aspect as norm dynamics has implications for digital-twin synthetic modeling and applications (*44*). For example, if AI agents behaved exactly like humans, then synthetic testing of norm dynamics under collective stresses—such as pandemics, terrorism, or wars—would be justified. If, on the other hand, LLM-specific dynamics proved to be substantial (e.g., if evidence of collective bias were further confirmed), then using these agents as simulations of human social systems or deploying these agents in social settings such as social media would require additional care. In particular, it is crucial to develop techniques to systematically identify discrepancies between LLM outputs and the expected human behavior (*70*), to then correct them with statistical techniques (*71*) or by keeping human judgments in the loop (*72*). Addressing these points is a key endeavor for the future, with far-reaching implications. Next steps involve further investigating convention dynamics in human and AI populations as well as in mixed LLM-human ecosystems, both in laboratory settings and, eventually, in natural environments like social media.

Our work also underscores the ethical challenges of bias propagation in LLMs. Despite their rapid adoption, these models pose serious risks, as the vast, unfiltered internet data used to train them can reinforce and amplify harmful biases, disproportionately harming marginalized communities (*73*). Accordingly, a central focus of the alignment research community has been to improve LLM performance in individual bias tests (*74*, *75*). However, our findings reveal that alignment must also be tested at the group level, where collective biases can emerge and persist.

Last, understanding how AI systems spontaneously develop conventions and more sophisticated norms without explicit programming is a critical first step for predicting and managing ethical AI behavior in real-world applications while ensuring agent alignment with human values and societal goals. It is also crucial for safeguarding AI agents from potential attacks. In particular, tipping points in norm dynamics present both opportunities, such as addressing global challenges (*28*, *29*), and risks, particularly if exploited for social control (*76*). Our findings highlight potential vulnerabilities in multi-agent systems, which could be exploited through injection attacks to influence the emergence of specific norms (*77*). Recognizing these risks, studying collective LLM behavior is crucial for assessing potential harms from the integration of AI agents into applications and for developing effective mitigation strategies. Moreover, efforts to measure and instill human social norms in LLMs have so far yielded mixed results (*78*, *79*), and, as of yet, AI agents struggle to represent multiple cultures (*80*) and continuously evolving social norms (*27*, *81*, *82*). We argue that the challenge extends beyond merely detecting "undesirable behavior," to understanding the evolution of social norms held by agents (*27*). In this light, our work represents a first step toward a better understanding of norm and bias dynamics in populations of LLMs, and we anticipate that it will be of interest to researchers and practitioners working to make AI a tool for societal good.

MATERIALS AND METHODS

Prompt

The system prompt comprises of three components: (i) a fixed prompt that outlines the game's rules, including the payoff structure

and the player's objective, (ii) a dynamic memory prompt that provides contextual information about the state of play within the player's memory range, and (iii) an instructional prompt that provides information for how the agent should format its response. The user prompt asks the agent to select a name to use in the current interaction. We use zero-shot prompting to directly extract the agent's name decision in response to the state of play. We do not provide instructions as to how agents should decide their next move, nor do we present them with example strategies. We ask the agent to behave in a self-interested manner, and the only part of the prompt in which we suggest to the agent that it should consider partaking in coordination is when we state that the agent's objective is to "maximize their own accumulated point tally, conditional on the behavior of their co-player." We apply fixed payoffs for successful and failed interactions, set at +100 and −50 points, respectively.

Models and APIs

For our experiments, we use homogeneous populations of agents instantiated from the following LLMs: Llama-3-70B-Instruct, Llama-3.1-70B-Instruct, Llama-2-70b-Chat (in 4-bit quantization format), and Claude-Sonnet-3.5 (see Table 2 for specific versions). All Llama family models are open-sourced LLMs, released under a commercial use license (https://ai.meta.com/llama/license/). We use versions of the Llama 3 family models hosted by Hugging Face, which we access through the inference Application Programming Interface (API) (https://huggingface.co/inference-api/serverless). We quantize Llama-2-70b-Chat into a 4-bit version using Hugging Face's Transformers library (https://huggingface.co/docs/transformers) and run the model locally using a single A100 GPU. In auto-regressive LLMs, each newly generated word is produced on the basis of previously inputted and generated words, and so the sequence of generation matters. More precisely, the probability distribution for predicting the next word is conditional on the product of all previous word probability distribution. To mimic LLMs deployed in real-world applications, we demand all agents in our experiments to behave non-deterministically by fixing them with a nonzero constant temperature. This means that, for each agent, the next generated word is randomly selected from the conditional probability distribution. We use K-sampling to restrict the probability distribution of the next word to the next K most likely words, thus increasing the likelihood of high probability words and decreasing the likelihood of low probability words that are outside of the name pool (see table S4 for all parameter values).

Measuring individual bias

We quantify the individual bias of agents by measuring the number of times each convention was produced in the first round of the game, when their memory inventory is empty, over T trials. Experiments with $W = 2$ are effectively a Bernoulli trial, and so we measure whether the agent is biased by performing a two-tailed exact binomial test with the observed proportions. We calculate the P value using a null probability of 0.5 and reject the hypothesis that the model is biased if $P < 0.05$. For the case of $W = 10$, we perform a chi-square test and also test the null hypothesis that the model is neutral in its convention selection. Thus, we use the expected value of 0.1 T in our calculations and, again, reject the null hypothesis that the model is unbiased if $P < 0.05$.

Table 2. Model names and versions.

Model name	Model version
Llama-3-70B-Instruct	Meta-Llama-3-70B-Instruct
Llama-3.1-70B-Instruct	Meta-Llama-3.1-70B-Instruct
Claude-3.5-Sonnet	claude-3-5-sonnet-20240620
Llama-2-70b-Chat	Meta-Llama-2-70b-Chat

Committed minorities

To determine the critical size of the committed minority, we identify the point at which the majority consensus is overturned. A consensus flip occurs when 95% of the past $3N$ interactions succeed after the introduction of the committed minority. For Llama-3-70B-Instruct, we tested the smallest minority needed to overturn a weak convention majority and then repeated the experiment with a strong convention majority to measure the critical mass within the same time frame. For other models, the critical mass threshold is defined as the minimum proportion of committed agents that is required to flip the consensus within 30 population rounds. These criteria account for potential fluctuations in nondeterministic agent decisions.

Supplementary Materials

This PDF file includes:

Supplementary Text
Figs. S1 to S12
Tables S1 to S7
References

REFERENCES AND NOTES

1. H. Young, The evolution of conventions. *Econometrica* **61**, 57–84 (1993).
2. P. Ehrlich, S. Levin, The evolution of norms. *PLoS Biol.* **3**, e194 (2005).
3. C. Bicchieri, *The Grammar of Society: The Nature and Dynamics of Social Norms* (Cambridge Univ. Press, 2012).
4. M. Gelfand, S. Gavrilets, N. Nunn, Norm dynamics: Interdisciplinary perspectives on social norm emergence, persistence, and change. *Annu. Rev. Psychol.* **75**, 341–378 (2024).
5. D. Lewis, *Convention: A Philosophical Study* (Blackwell, 1969).
6. A. Baronchelli, The emergence of consensus: A primer. *R. Soc. Open Sci.* **5**, 172189 (2018).
7. L. Steels, A self-organizing spatial vocabulary. *Artif. Life* **2**, 319–332 (1995).
8. A. Baronchelli, M. Felici, L. Vittorio, E. Caglioti, L. Steels, Sharp transition towards shared vocabularies in multi-agent systems. *J. Stat. Phys.* **2006**, P06014 (2006).
9. D. Centola, A. Baronchelli, The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 1989–1994 (2015).
10. F. Hayek, *The Constitution of Liberty* (University of Chicago Press, 1960).
11. R. Sugden, Spontaneous order. *J. Econ. Perspect.* **3**, 85–97 (1989).
12. J. Werfel, K. Petersen, R. Nagpal, Designing collective behavior in a termite-inspired robot construction team. *Science* **343**, 754–758 (2014).
13. L. Brinkmann, F. Baumann, J. F. Bonnefon, M. Derex, T. F. Müller, A. M. Nussberger, A. Czaplicka, A. Acerbi, T. L. Griffiths, J. Henrich, J. Z. Leibo, R. McElreath, P. Y. Oudeyer, J. Stray, I. Rahwan, Machine culture. *Nat. Hum. Behav.* **7**, 1855–1868 (2023).
14. Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, C. Wang, Autogen: Enabling next-gen llm applications via multi-agent conversation framework. arXiv:2308.08155 (2023).
15. A. Dafoe, Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson, T. Graepel, Cooperative AI: Machines must learn to find common ground. *Nature* **593**, 33–36 (2021).
16. M. Mézard, G. Parisi, M. Virasoro, *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications* (World Scientific, 1987).
17. C. Castellano, S. Fortunato, V. Loreto, Statistical physics of social dynamics. *Rev. Mod. Phys.* **81**, 591–646 (2009).
18. D. Roselli, J. Matthews, N. Talagala, "Managing bias in AI," in *Companion Proceedings of the 2019 World Wide Web Conference* (ACM, 2019), pp. 539–544.
19. E. Ferrara, Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Mon.* **28**, doi.org/10.5210/fm.v28i11.13346 (2023).

20. T. Hu, Y. Kyrychenko, S. Rathje, N. Collier, S. van der Linden, J. Roozenbeek, Generative language models exhibit social identity biases. arXiv:2310.15819 (2023).
21. U. Anwar, A. Saparov, J. Rando, D. Paleka, M. Turpin, P. Hase, E. S. Lubana, E. Jenner, S. Casper, O. Sourbut, B. L. Edelman, Z. Zhang, M. Günther, A. Korinek, J. Hernandez-Orallo, L. Hammond, E. Bigelow, A. Pan, L. Langosco, T. Korbak, H. Zhang, R. Zhong, Seán Ó hÉigeartaigh, G. Recchia, G. Corsi, A. Chan, M. Anderljung, L. Edwards, A. Petrov, Christian Schroeder de Witt, S. R. Motwan, Y. Bengio, D. Chen, P. H. S. Torr, S. Albanie, T. Maharaj, J. Foerster, F. Tramer, H. He, A. Kasirzadeh, Y. Choi, D. Krueger, Foundational challenges in assuring alignment and safety of large language models. arXiv:2404.09932 (2024).
22. J. Xie, S. Sreenivasan, G. Korniss, W. Zhang, C. Lim, B. K. Szymanski, Social consensus through the influence of committed minorities. *Phys. Rev. E* **84**, 011130 (2011).
23. D. Centola, J. Becker, D. Brackbill, A. Baronchelli, Experimental evidence for tipping points in social convention. *Science* **360**, 1116–1119 (2018).
24. T. Kuran, Ethnic norms and their transformation through reputational cascades. *J. Leg. Stud.* **27**, 623–659 (1998).
25. R. Kanter, Some effects of proportions on group life: Skewed sex ratios and responses to token women. *Am. J. Sociol.* **82**, 965–990 (1977).
26. D. Dahlerup, The story of the theory of critical mass. *Polit. Gend.* **2**, 511–522 (2006).
27. A. Baronchelli, Shaping new norms for AI. *Philos. Trans. R. Soc. B* **379**, 20230028 (2024).
28. K. Nyborg, J. M. Anderies, A. Dannenberg, T. Lindahl, C. Schill, M. Schlüter, W. N. Adger, K. J. Arrow, S. Barrett, S. Carpenter, F. S. Chapin III, A. S. Crépin, G. Daily, P. Ehrlich, C. Folke, W. Jager, N. Kautsky, S. A. Levin, O. J. Madsen, S. Polasky, M. Scheffer, B. Walker, E. U. Weber, J. Wilen, A. Xepapadeas, A. de Zeeuw, Social norms as solutions. *Science* **354**, 42–43 (2016).
29. J. D. Farmer, C. Hepburn, M. C. Ives, T. Hale, T. Wetzer, P. Mealy, R. Rafaty, S. Srivastav, R. Way, Sensitive intervention points in the post-carbon transition. *Science* **364**, 132–134 (2019).
30. S. Garrod, G. Doherty, Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition* **53**, 181–215 (1994).
31. D. Hume, *A Treatise of Human Nature* (Oxford Univ. Press, 2000).
32. L. Wittgenstein, *Philosophical Investigations* (Blackwell, 1958).
33. A. Lazaridou, A. Peysakhovich, M. Baroni, Multi-agent cooperation and the emergence of (natural) language. arXiv:1612.07182 (2016).
34. T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, X. Zhang, Large language model based multi-agents: a survey of progress and challenges. arXiv:2402.01680 (2024).
35. S. Ren, Z. Cui, R. Song, Z. Wang, S. Hu, Emergence of social norms in large language model-based agent societies. arXiv:2403.08251 (2024).
36. I. Horiguchi, T. Yoshida, T. Ikegami, Evolution of social norms in LLM agents using natural language. arXiv:2409.00993 (2024).
37. L. Hewitt, A. Ashokkumar, I. Ghezae, R. Willer, Predicting results of social science experiments using large language models (2024); https://docsend.com/view/ity6yf2dansesucf [accessed 3 March 2025].
38. J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (ACM, 2023), pp. 1–22.
39. Y.-S. Chuang, A. Goyal, N. Harlalka, S. Suresh, R. Hawkins, S. Yang, D. Shah, J. Hu, T. T. Rogers, Simulating opinion dynamics with networks of llm-based agents. arXiv:2311.09618 (2023).
40. J. Han, B. Battu, I. Romić, T. Rahwan, P. Holme, Static network structure cannot stabilize cooperation among Large Language Model agents. arXiv:2411.10294 (2024).
41. K.-T. Tran, D. Dao, M.-D. Nguyen, Q.-V. Pham, B. O'Sullivan, H. D. Nguyen, Multi-agent collaboration mechanisms: A survey of LLMs. arXiv:2501.06322 (2025).
42. L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, J. Wen, A survey on large language model based autonomous agents. *Front. Comp. Sci.* **18**, 186345 (2024).
43. M. Tsvetkova, T. Yasseri, N. Pescetelli, T. Werner, A new sociology of humans and machines. *Nat. Hum. Behav.* **8**, 1864–1876 (2024).
44. G. Rossettie, M. Stella, R. Cazabet, K. Abramski, E. Cau, S. Citraro, A. Failla, R. Improta, V. Morini, V. Pansanella, Y Social: An LLM-powered social media digital twin. arXiv:2408.00818 (2024).
45. B. Skyrms, *Evolution of the Social Contract* (Cambridge Univ. Press, 2014).
46. P. Michel, M. Rita, K. Mathewson, O. Tieleman, A. Lazaridou, "Revisiting populations in multi-agent communication," in *The Eleventh International Conference on Learning Representations* (ICLR, 2023), pp. 1–16.
47. X. Niu, C. Doyle, G. Korniss, B. Szymanski, The impact of variable commitment in the naming game on consensus formation. *Sci. Rep.* **7**, 41750 (2017).
48. S. Grey, Numbers and beyond: The relevance of critical mass in gender research. *Polit. Gend.* **2**, 492–502 (2006).
49. M. Diani, The concept of social movement. *Sociol. Rev.* **40**, 1–25 (1992).
50. M. Gladwell, "Small change," *The New Yorker*, 4 October 2010.
51. R. Amato, L. Lacasa, A. Díaz-Guilera, A. Baronchelli, The dynamics of norm change in the cultural evolution of language. *Proc. Natl. Acad. Sci.* **115**, 8260–8265 (2018).
52. L. Backstrom, D. Huttenlocher, J. Kleinberg, X. Lan, "Group formation in large social networks: Membership, growth, and evolution," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2006), pp. 44–54.
53. F. Kooti, H. Yang, M. Cha, K. Gummadi, W. Mason, "The emergence of conventions in online social networks," in *Proceedings of the International AAAI Conference on Web and Social Media* (PKP Publishing Services Network, 2012), vol. 6, pp. 194–201.
54. T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners. arXiv:2205.11916 (2022).
55. J. Huang, X. Chen, S. Mishra, H. S. Zheng, A. W. Yu, X. Song, D. Zhou, Large language models cannot self-correct reasoning yet. arXiv:2310.01798 (2023).
56. Z. Xu, S. Jain, M. Kankanhalli, Hallucination is inevitable: An innate limitation of large language models. arXiv:2401.11817 (2024).
57. N. Fontana, F. Pierri, L. Aiello, Nicer than humans: How do large language models behave in the prisoner's dilemma? arXiv:2406.13605 (2024).
58. G. De Marzo, L. Pietronero, D. Garcia, Emergence of scale-free networks in social interactions among large language models. arXiv:2312.06619 (2023).
59. G. De Marzo, C. Castellano, D. Garcia, Language understanding as a constraint on consensus size in LLM societies. arXiv:2409.02822 (2024).
60. C. Bail, Can generative AI improve social science? *Proc. Natl. Acad. Sci.* **121**, e2314021121 (2024).
61. J. Perez, G. Kovač, C. Léger, C. Colas, G. Molinaro, M. Derex, P.-Y. Oudeyer, C. Moulin-Frier, When LLMs play the telephone game: Cumulative changes and attractors in iterated cultural transmissions. arXiv:2407.04503 (2024).
62. M. Kearns, S. Judd, J. Tan, J. Wortman, Behavioral experiments on biased voting in networks. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 1347–1352 (2024).
63. J. Duan, R. Zhang, J. Diffenderfer, B. Kailkhura, L. Sun, E. Stengel-Eskin, M. Bansal, T. Chen, K. Xu, GTBench: Uncovering the strategic reasoning limitations of LLMs via game-theoretic evaluations. arXiv:2402.12348 (2024).
64. T. R. Davidson, V. Veselovsky, M. Josifoski, M. Peyrard, A. Bosselut, M. Kosinski, R. West, Evaluating language model agency through negotiations. arXiv:2401.04536 (2024).
65. A. Lazaridou, K. M. Hermann, K. Tuyls, S. Clark, "Emergence of linguistic communication from referential games with symbolic and pixel input," in *International Conference on Learning Representations* (ICLR, 2018), pp. 1–13.
66. I. Iacopini, G. Petri, A. Baronchelli, A. Barrat, Group interactions modulate critical mass dynamics in social convention. *Commun. Phys.* **5**, 64 (2022).
67. C. Sun, S. Huang, D. Pompili, LLM-based multi-agent reinforcement learning: Current and future directions. arXiv:2405.11106 (2024).
68. W. Hua, O. Liu, L. Li, A. Amayuelas, J. Chen, L. Jiang, M. Jin, L. Fan, F. Sun, W. Wang, X. Wang, Y. Zhang, Game-theoretic LLM: Agent workflow for negotiation games. arXiv:2411.05990 (2024).
69. I. Gemp, R. Patel, Y. Bachrach, M. Lanctot, V. Dasagi, L. Marris, G. Piliouras, S. Liu, K. Tuyls, "Steering language models with game-theoretic solvers," in *Agentic Markets Workshop at ICML 2024* (ICML, 2024), pp. 1–22.
70. K. Gligorić, T. Zrnic, C. Lee, E. J. Candès, D. Jurafsky, Can unconfident LLM annotations be used for confident conclusions? arXiv:2408.15204 (2024).
71. N. Egami, M. Hinck, B. Stewart, H. Wei, Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. *Adv. Neural. Inf. Process Syst.* **36**, 68589–68601 (2023).
72. M. Li, T. Shi, C. Ziems, M.-Y. Kan, N. F. Chen, Z. Liu, D. Yang, Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. arXiv:2310.15638 (2023).
73. I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, N. K. Ahmed, Bias and fairness in large language models: A survey. arXiv:2309.00770 (2024).
74. Y. Wang, W. Zhong, L. Li, F. Mi, X. Zeng, W. Huang, L. Shang, X. Jiang, Q. Liu, Aligning large language models with human: A survey. arXiv:2307.12966 (2023).
75. L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback. arXiv:2203.02155 (2022).
76. G. King, J. Pan, M. E. Roberts, How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *Am. Polit. Sci. Rev.* **111**, 484–501 (2017).
77. D. Lee, M. Tiwari, Prompt infection: Llm-to-llm prompt injection within multi-agent systems. arXiv:2410.07283 (2024).
78. Y. Yuan, K. Tang, J. Shen, M. Zhang, C. Wang, Measuring social norms of large language models. arXiv:2404.02491 (2024).
79. K. Hämmerl, B. Deiseroth, P. Schramowski, J. Libovický, A. Fraser, K. Kersting, Do multilingual language models capture differing moral norms? arXiv:2203.09904 (2022).

80. A. Ramezani, Y. Xu, Knowledge of cultural moral norms in large language models. arXiv:2306.01857 (2023).
81. S. Li, T. Sun, Q. Cheng, X. Qiu, Agent alignment in evolving social norms. arXiv:2401.04620 (2024).
82. H. Shen, T. Li, T. J.-J. Li, J. S. Park, D. Yang, "Shaping the emerging norms of using large language models in social computing research," in *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (ACM, 2023), pp. 569–571.
83. E. Akata, L. Schulz, J. Coda-Forno, S. J. Oh, M. Bethge, E. Schulz, Playing repeated games with large language models. arXiv:2305.16867 (2023).
84. T. Ullman, Large language models fail on trivial alterations to theory-of-mind tasks. arXiv:2302.08399 (2023).
85. G. V. Aher, R. I. Arriaga, A. T. Kalai, "Using large language models to simulate multiple humans and replicate human subject studies," in *International Conference on Machine Learning* (PMLR, 2023), pp. 337–371.
86. B. Lester, R. Al-Rfou, N. Constant, The power of scale for parameter-efficient prompt tuning. arXiv:2104.08691 (2021).

Acknowledgments: We acknowledge City St George's, University of London's Hyperion cluster for computation time. **Funding:** L.M.A. acknowledges the support from the Carlsberg Foundation through the COCOONS project (CF21-0432). **Author contributions:** A.F.A., L.M.A., and A.B. designed the study. A.F.A. and L.M.A performed the experiments. A.F.A. wrote the code for the experiments. A.F.A., L.M.A., and A.B. analyzed the data, discussed the results, and contributed to the final manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The code and data used in this manuscript are available in the following GitHub repository: https://github.com/Ariel-Flint-Ashery/AI-norms and https://doi.org/10.5281/zenodo.14937173.

Submitted 27 November 2024
Accepted 2 April 2025
Published 14 May 2025
10.1126/sciadv.adu9368