Science Advances
AAAS

Supplementary Materials for

Emergent social conventions and collective bias in LLM populations

Ariel Flint Ashery *et al*.

Corresponding author: Andrea Baronchelli, andrea.baronchelli.1@citystgeorges.ac.uk

Sci. Adv. **11**, eadu9368 (2025)
DOI: 10.1126/sciadv.adu9368

This PDF file includes:

Supplementary Text

Microscopic Bias

Statistical Tests for Table 1

To measure the strategy bias in Table 1, we assess both the interaction-level bias and the bias within each unique memory configuration. The interaction-level bias is defined as the overall production probability of the strong convention across all possible configurations (per interaction), which we test using an exact Binomial test with a null probability of $p = 0.5$, rejecting the null hypothesis if the p-value falls below 0.05. The results of the tests are reported in the caption of Table 1. At the configuration level, we first perform an exact Binomial test, as above, to check whether the model is biased. In all cases, the p-value $p^{\dagger} < 0.05$, confirming that the model's decision is biased towards an extreme. Then, we use bootstrapping by resampling 70% of the observations for each configuration $10,000$ times and measure the proportion of samples showing a stronger bias than the observed value in Table 1. In all cases, we obtain a bootstrapped $p^{\ddagger} > 0.05$, indicating that we cannot reject the hypothesis that the model's underlying bias is more extreme than the observed bias. Results are reported in Table S5.

Effect of Fine-Tuning

We reproduced the set of experiments reported in Table 1 for **Llama-3.1-70B-Instruct** (w/ Fine-Tuning) using **Llama-3.1-70B** (w/o Fine-Tuning), to investigate the effects of instruction-fine-tuning on strategy bias. We ran a 4-bit quantization of the LLM on a single A100 GPU and extracted the next-token probability distribution over the valid convention tokens. The initial, no-memory probability distribution for the same convention choices as in Table 1 is biased for the pre-trained model, with $p(Q) = 0.525$. To faithfully reproduce any potential biased symmetry breaking, we conduct the experiments using a pair of six-character strings, each consisting of randomly chosen letters and numbers. Most importantly, we repeatedly sampled random pairs until we found a combination that had Jensen-Shannon distance of less than 0.005 from the neutral distribution in the no-memory stage, i.e., an initially unbiased combination. The results are reported in Table S6, using the pseudo-labels 'A' and 'B' for the random conventions. We find that although there is

no initial preference for either convention, a strong collective bias already appears by the second interaction. Thus, we cannot say that fine-tuning is responsible for inducing collective bias effects.

Theoretical Model

The *naming game* model simulates a population of N agents engaging in pairwise negotiation interactions, demonstrating the emergence of global consensus on conventions through local coordination mechanisms. In the canonical formulation (*8*), agents must reach consensus on the name for an object using only local interactions, similar to our experimental framework. Agents possess internal lexicons with unlimited word capacity (although this is not a necessary initial condition of the model), initially empty. The interaction protocol involves random selection of agent pairs, where the designated speaker transmits a randomly chosen word from their lexicon (or invents a new one if it is empty) to the hearer. If the hearer recognizes the word in their own lexicon, both agents retain only the communicated word, while in case of failure, the hearer incorporates the novel word into their lexicon. The non-equilibrium dynamics of this system exhibit three distinct temporal phases: *(i)* an innovation phase characterized by word creation, *(ii)* a propagation phase involving lexicon reorganization, and *(iii)* a convergence phase culminating in global consensus. In our experimental framework, we set an initial condition whereby agents can only invent new words from a finite word pool of size $W < N$. This condition means that the initial innovation phase is extremely short, as seen in the inset of Fig. 1. This model provides insights into the dynamics of language evolution and convention formation in both human and artificial communication systems.

Fig. S3 shows the production probability trajectories of a simulation of the theoretical model with a lexicon of two words ($W = 2$).

Characterizing Relative Strength

We define the relative strength of a convention as its ability to attract the system toward an equilibrium consensus on it, and retain the system in this consensus state. As such, the relative strength of a 'strong' convention over its 'weak' counterpart is best demonstrated through the latter's ability to overturn consensus on the former. Importantly, we note that the trajectory of the system is sensitive to initial conditions, such that not all memory configurations are observed during

the evolution of the system. Fig. S8 shows the complete probability distribution for all memory configurations given that the memory size is $H = 5$, while Fig. S9 provides a characterization of the relative strength of convention pairs in different models in two settings: we evolve the system from a consensus on the strong convention *i)* without perturbation and *ii)* with a perturbation of a single agent committed to the weak convention. We show that the relative strength of a convention with respect to its counterpart can be characterized by the basin of attraction of the steady state on the strong convention. Although **Llama-3.1-70B-Instruct** agent populations tend to be more exploratory, even in the absence of perturbation, the memory state bounds of their exploration have strong resistance and return the system towards a state of equilibrium around the strong convention steady state. In contrast, while **Llama-3-70B-Instruct** agent populations do not explore beyond the consensus state once they reach it, they are far more susceptible to perturbations, with some *individual* agents briefly exhibiting memory configurations in which all past interactions within their memory range reflect successful coordination on the weak convention. Finally, Fig. S10 shows that relative strength also plays a role in convergence, restricting the exploration of of the system in the path towards consensus for **Llama-3.1-70B-Instruct** population. We therefore propose that the relative strength of a convention can be characterized by properties of its steady-state basin of attraction. The 'width' of the basin reflects the range of initial conditions that lead to convergence on the convention. The 'pull' represents the stability of the steady state, capturing both its resistance to perturbations and the tendency of trajectories to return after deviations, as well as the rate at which they are drawn toward the steady state.

Prompting

Prompt Structure

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. We find that agents generally struggle to behave in a manner befitting a partnership game, and often opt for strategies aimed at undermining their co-player's payoff, effectively treating them as an

opponent. In practice, this meant that on some occasions the agent would willingly take an action with a negative payoff, in order to harm their co-player's accumulated point tally.

Output Structure

To extract any meaningful decision from an LLM's output, which may be verbose and unstructured, it is necessary to distinguish between the *reason*, where the agent 'shows its working', and the final *decision*. One popular approach to prompting asks the agent to give a final decision at the end of its answer, allowing it to generate a reasoning for the decision before reporting the actual response. This method relies on the assumption that if the LLM is good at composing a well thought-out reasoning, having it spelled out explicitly would guide the choice of the LLM towards better performance due to the LLM's auto-regressive text generation method. Although this approach has shown promising results in a variety of tasks and multi-player games, there is an ongoing debate as to whether LLMs truly possess the ability to reason about the possible actions, beliefs, and intentions of their fellow players (*83–85*).

One can argue that even if agents were able to reason, the wording generated to explain the reasoning can potentially express the agent's biases and influence the final decision. As a result, the reason-first, answer-later structure would make it difficult to identify the possible sources of bias within the prompt, especially since the agent's reasoning will only serve to amplify its own bias reflected in the decision. To strengthen generalization and ensure that the final action decision is influenced solely by the input prompt, we instead ask the agent explicitly to provide an output in the form *answer-first-reason-later*.

Example Prompt

We provide an example of the system and user prompts given to LLM agents in our experiments. We exclude the beginning of text and end of text tokens, which are unique to each model.

System Prompt

Context: Player 1 is playing a multi-round partnership game with Player 2 for 100 rounds. At each round, Player 1 and Player 2 simultaneously pick an action from the following values: [*F, J*]. The payoff that both players get is determined by the following rule:
1. If Players play the SAME action as each other, they will both be REWARDED with payoff 100 points.
2. If Players play DIFFERENT actions to each other, they will both be PUNISHED with payoff -50 points.
The objective of each Player is to maximize their own accumulated point tally, conditional on the behavior of the other player. This is the history of choices in past rounds:
{'round':1, 'Player 1': F, 'Player 2': J, 'payoff': -50}
.
.
.
It is now round *i*. The current score of Player 1 is *SCORE*. Answer saying which value Player 1 should pick. Please think step by step before making a decision. Remember, examining history explicitly is important. Write your answer using the following format: {'value': <*VALUE_OF_PLAYER_1*>; 'reason': <*YOUR_REASON*>}.

User Prompt

Answer saying which action Player 1 should play.

Meta-Prompting

When LLMs are used to solve tasks where some form of ground truth is defined, such as classification or regression, the effect of prompt variations on the quality of a model's outputs can be measured on downstream performance (*86*). However, that is not possible in generative tasks where a notion of error is undefined. Specifically in the naming game, any generated output is plausible, as long as it is within the set of allowed symbols. This ambiguity makes it difficult to assess whether the

LLM's outputs reflect a proper semantic understanding of the task's rules or are merely products of statistical 'hallucinations' (*56*). To partially address this issue, we rely on a meta-prompting technique to measure the LLMs' level of comprehension of the given prompt (*57*). This technique provides the LLM with the prompt, and then asks three types of *prompt comprehension questions* about: interaction rules, chronological sequence of actions in the history, and payoff statistics (Table S7).

To assess the LLMs' proficiency in responding to meta-prompting questions, we randomly selected a group of agents from a real stochastic simulation of the naming game. For each agent, we used its history to replay each of its previous interactions, using the same memory length it had in the simulation. Retracing every interaction, including the agent's memory at the time, we ask the agent all possible comprehension questions. Note that certain questions that rely on memory cannot be asked in the first interaction. We pose the questions at each interaction and compute the average accuracy of the LLMs' responses across all interactions for all agents. Overall, all models exhibit a good level of prompt comprehension, with response accuracy nearly always above 0.8 and most often close to 1 (Fig. S1). The only model that went below 0.8 in any metric is **Llama-2-70b-Chat**, which showed relatively poor accuracy in counting the number of times it played a convention within memory range. In many cases, this agent confused the ID of the player it was asked to consider, or it answered how many times a convention has been observed in total, across both players. Here, it is also worth noting that LLMs from **Llama-2-70b-Chat**'s generation generally struggled with counting tasks (*55, 57*).

Alternative Prompts

To reduce the risks from bias from previously seen examples in the training data, we test the experimental framework with two additional system prompt templates and using name pools with conventions made up of random strings. The prompt variations are *i)* a factual bullet-point style, and *ii)* an expository, narrative prompt. For the name pools, we generate strings using six characters chosen at random from a set of numerical digits, lowercase and uppercase letters from the Latin alphabet. We only select a name pool if its no-memory probability distribution over conventions has a Jensen-Shannon distance from the neutral distribution less than 0.005, to ensure there is no initial bias. As shown in Figs. S11 and S12, biased consensus emerges across all the variations.

In considering whether bias in the prompt arises from the presence of the experimental framework in the training data, we note that the game's population dynamics and the symmetry in payoffs across action labels ensure that no predetermined optimal strategy exists—even if a player were aware of global information. Convergence occurs purely at the local level, driven by repeated random interactions rather than a global rule that dictates strategic behavior. While individual agents may exhibit micro-level adaptation, adjusting their choices in response to patterns in their co-players' actions, these adjustments remain localized and do not necessarily lead to a predictable global bias. This contrasts with games like the Ultimatum Game, where systematic behavioral responses shape equilibrium strategies. Similarly, unlike the Prisoner's Dilemma, where payoff asymmetries guide decision-making, the payoff symmetry in our game framework prevents any clear incentive toward one action over another. As a result, the game's long-term trajectory remains emergent and stochastic, and only a higher-order metanorm—beyond the direct game mechanics—could impose a systematic bias or predictable pattern.

Narrative System Prompt

Context: Player 1 is playing a multi-round partnership game with Player 2 for 100 rounds. At each round, Player 1 and Player 2 make simultaneous action choices between the actions <*ACTION_LABEL_1*> and <*ACTION_LABEL_1*>. The outcome of each round depends on the combination of choices made by the Players in that round. If both Players choose the same action, then the round is a success and both players receive a payoff of +100 points. However, if the Players choose different actions to each other, then the round is a failure and both Players receive a payoff of -50 points. The objective of each Player is to maximise their own accumulated points total. In round 1, Player 1 chose action <*ACTION_OF_PLAYER_1*> and Player 2 chose action <*ACTION_OF_PLAYER_2*>; Player 1 received a payoff of <*PAYOFF*> points and Player 2 received a payoff of <*PAYOFF*> points... It is now round i. The current score of Player 1 is <*SCORE_OF_PLAYER_1*>. Answer saying which action Player 1 should pick. Please think step by step before making a decision. Remember, examining history explicitly is important. Write your answer using the following format: {'action': <*ACTION_OF_PLAYER_1*>; 'reason': <*YOUR_REASON*>}.

Bullet-point System Prompt

Context: Player 1 and Player 2 are playing a repeated 2-player game.

Game Setup:

• Players: Player 1 and Player 2.

• Actions available to both players: <*ACTION_LABEL_1*> and <*ACTION_LABEL_1*>.

• Players make their choices simultaneously.

Payoff Matrix:

• If Player 1 chooses <*ACTION_LABEL_1*> and Player 2 chooses <*ACTION_LABEL_1*>, then:

- Player 1 receives +100 points.

- Player 2 receives +100 points.

• If Player 1 chooses <*ACTION_LABEL_1*> and Player 2 chooses <*ACTION_LABEL_2*>, then:

- Player 1 loses -50 points.

- Player 2 loses -50 points.

• If Player 1 chooses <*ACTION_LABEL_2*> and Player 2 chooses <*ACTION_LABEL_2*>, then:

- Player 1 receives +100 points.

- Player 2 receives +100 points.

• If Player 1 chooses <*ACTION_LABEL_2*> and Player 2 chooses <*ACTION_LABEL_1*>, then:

- Player 1 loses -50 points.

- Player 2 loses -50 points.

Player objectives:

• The objective of each player is to maximize their own accumulated point tally.

Bullet-point System Prompt (Continued)

Game History:

Round 1:

• Action choices:

- Player 1: <*ACTION_OF_PLAYER_1*>.

- Player 2: <*ACTION_OF_PLAYER_2*>.

• Payoff:

- Player 1: <*PAYOFF*>.

- Player 2: <*PAYOFF*>.

...

Next Round:

It is now round *i*. The current score of Player 1 is <*SCORE_OF_PLAYER_1*>. You must now decide which action Player 1 should take in this round. Use previous rounds as additional guidance. Remember, the outcome of Player 1's action choice also depends on Player 2's action choice. Write your answer using the following format: {'action': <*ACTION_OF_PLAYER_1*>; 'reason': <*YOUR_REASON*>}.

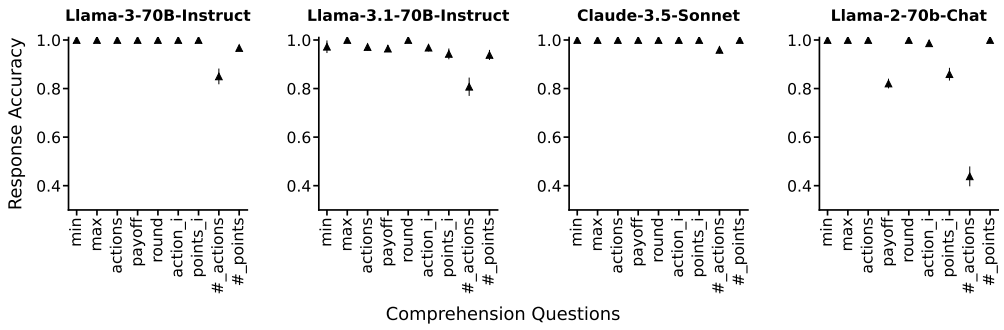

Figure S1: **Meta-prompting results** Accuracy of the model responses to the prompt comprehension questions defined in Table S7. We selected 8 agents from a single run (5 agents for **Llama-3-70B-Instruct**), and recovered their game record. We replayed the game using the memory length used in the simulated run ($H = 5$), posing the comprehension questions at each interaction. These runs provide approximately 100 test interactions for each model.

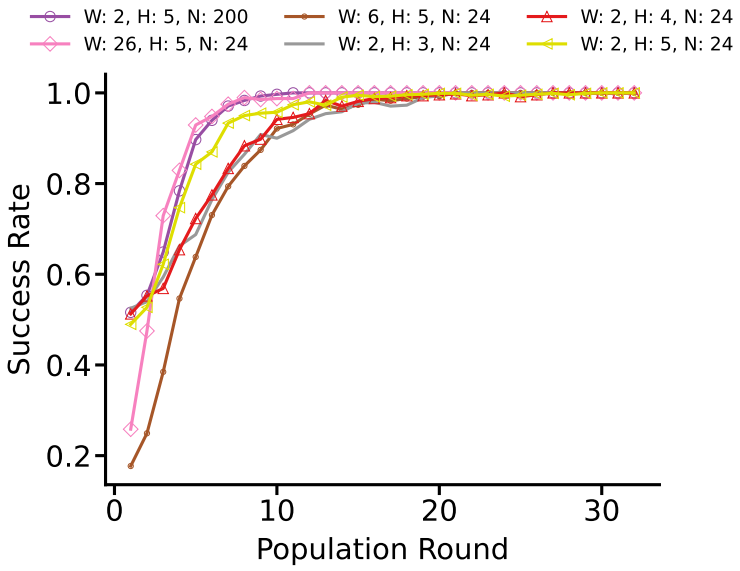

Figure S2: **Robustness of the Spontaneous emergence of conventions.** We show that the spontaneous emergence of conventions holds for a variety of simulation parameters, using populations of **Llama-3-70B-Instruct** agents. $W = 26$ indicates a name pool which uses the entire Latin alphabet, $W = 6$ is the name pool $\{Q, M, F, J, X, Y\}$, and $W = 2$ is the name pool $\{Q, M\}$.

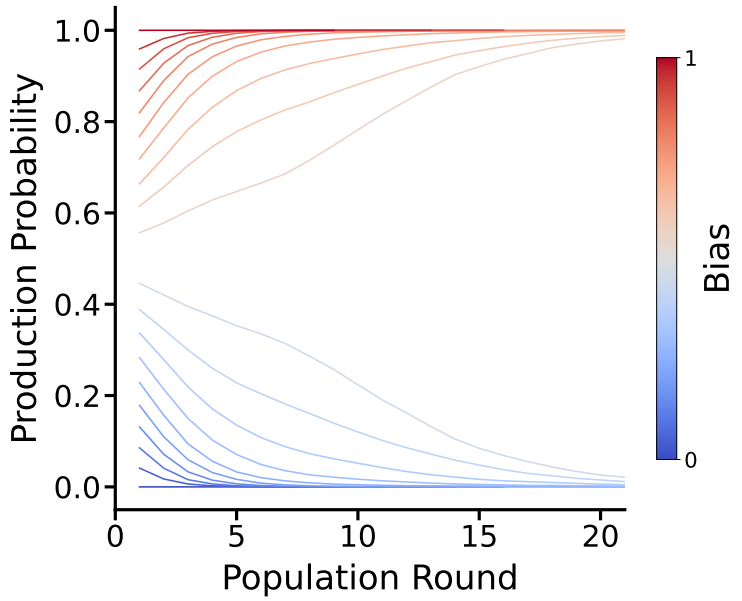

Figure S3: **Production probability trajectories in the theoretical minimal naming game with** $W = 2$**.** Agents can only choose names from the pool $\{0,1\}$. We simulated 10,000 runs of the minimal naming game in a population of 24 agents. We show biased trajectories (in probability increments of 0.05) towards choosing the name '1'. The bias corresponds to the probability of choosing the name '1' when an agent has the option of producing either name, such that a bias of 1 (0) corresponds to agents that will only choose name '1' ('0'). We note that as the bias increases, the convergence speed increases. Crucially, even a small bias towards a certain name leads to inevitable global convergence on that name.

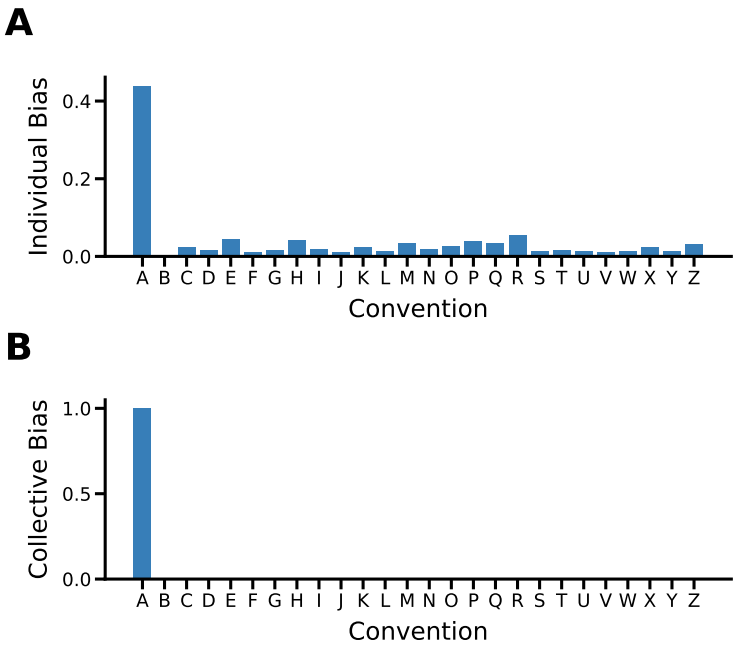

Figure S4: **(A) Individual and (B) Collective bias in convention selection with** $W = 26$**, the entire Latin Alphabet.** Agents favor the convention 'A' over all others *a priori*, resulting in collective consensus on this convention. Individual bias shows the convention production probability from 480 samples using **Llama-3-70B-Instruct** agents, where agents have empty memory. Collective bias shows the proportion of consensus conventions from 20 simulations.

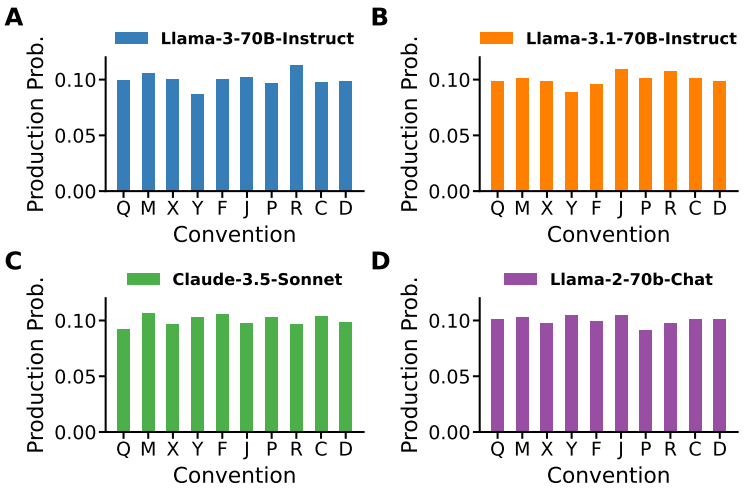

Figure S5: **Individual bias in conventions selection with** $W = 10$**.** The production probability of each convention when agents have no prior memory, for the LLM agents indicated in the legend. For (**A**-**D**), we generated 15,000, 10,000, 4,500, and 10,000 samples. We performed a chi-squared hypothesis test to see whether the agents are biased, and calculated the following p-values: $p = < 0.001, 0.001, 0.100, 0.410$. These p-values indicate that **Llama-2-70b-Chat** and **Claude-3.5-Sonnet** are unbiased across conventions in this name pool (at the 5% significance level), whereas the **Llama-3/3.1-70B-Instruct** models exhibit significant skew.

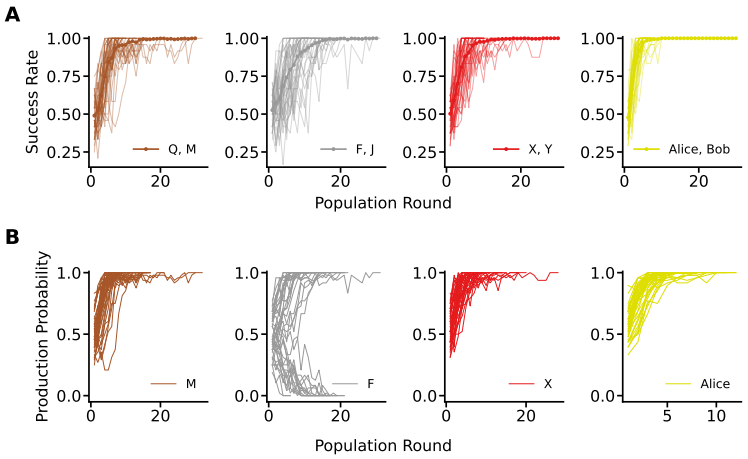

Figure S6: **The Spontaneous emergence of conventions for** $W = 2$**.** We present individual (faint lines) and average (thick lines) trajectories of a population of $N = 24$ **Llama-3-70B-Instruct** agents with memory length $H = 5$ for four different name pools. For each name pool, we show **(A)** the success rate, and **(B)** the production probability of the strong convention (as indicated by the legend). All name pools resulted in a collective bias on a particular convention ($\{Q, M\} : M$, 40/40 runs; $\{F, J\} : F$, 24/40 runs; $\{X, Y\} : X$, 40/40 runs; $\{Alice, Bob\} : Alice$, 40/40 runs).

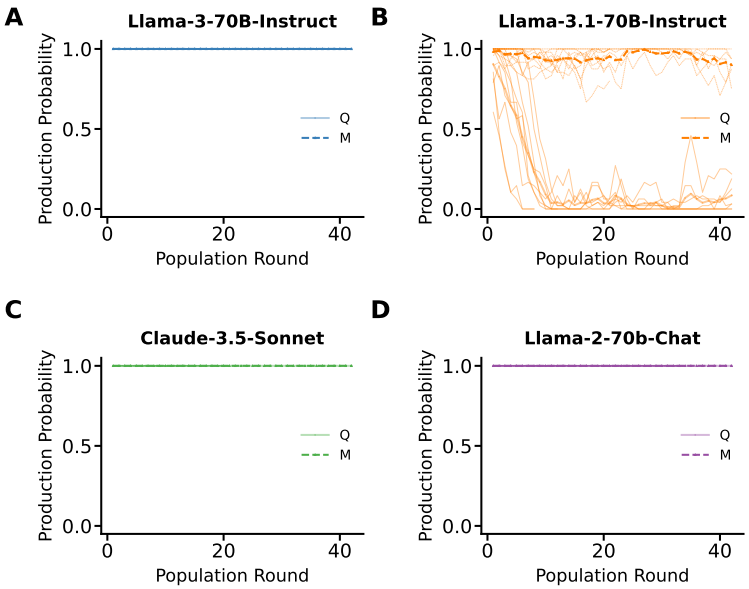

Figure S7: **Stability of consensus conventions.** We test the stability of each model in the setting $W = 2$, with the possible conventions shown in the legend. For each convention, we begin with a population of $N = 24$ agents ($N = 48$ for **Llama-3-70B-Instruct**), where every agent has only the respective convention in memory. We then allow the population to naturally evolve, and measure the production probability of this convention. We simulate the following number of runs (**A-D**): 4, 10, 3, 5. Faint lines show the trajectories of individual runs. We show that for all models apart from **Llama-3.1-70B-Instruct**, the population remains entirely stable at its initial consensus state. If **Llama-3.1-70B-Instruct** is initialized with consensus on the weak convention (Q), the population immediately switches to the alternative convention (M, the strong convention). Here, the strong convention remains stable, with some minor fluctuations. The instability of the weak convention is also observed in our study of the committed minority required to flip a majority consensus (see Fig. 3), where we see that M acts as a strong attractor state that can only be overcome by a large enough committed minority on the weak convention, Q.

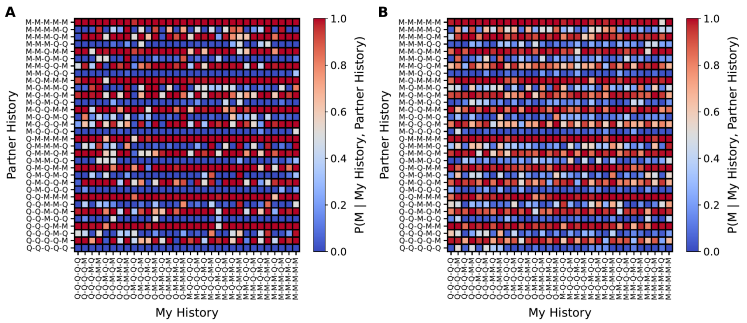

Figure S8: **Distribution of strategies across memory states.** We 'open' the black-box to show the entire probability distribution across memory states for $H = 5$ for (**A**) **Llama-3-70B-Instruct**, and (**B**) **Llama-3.1-70B-Instruct** agents. Each tile represents a joint history that comprises the agent's memory state. The color of the tile indicates the probability that the agent will choose the convention M in the next interaction, which remains fixed for each memory state at all times.

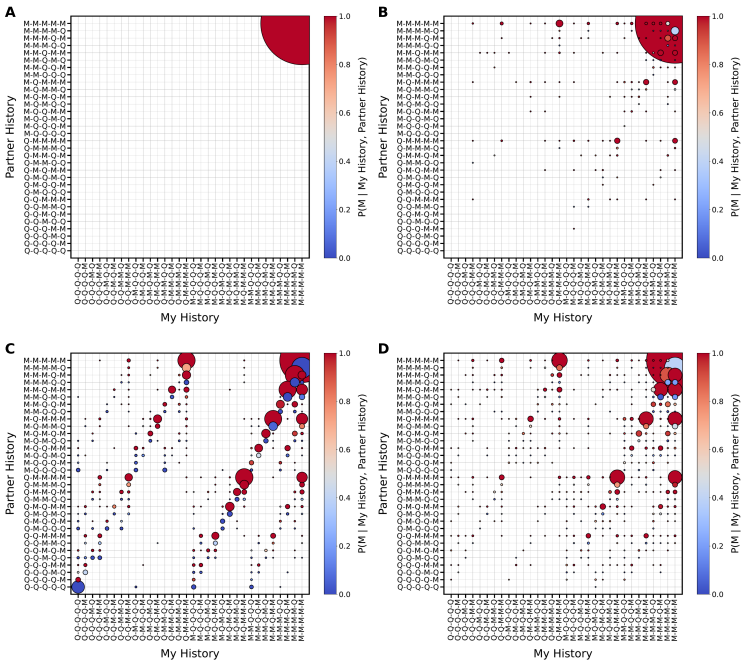

Figure S9: **Microscopic characterization of memory-dependent strategy bias in real-game dynamics.** The matrices depict the actual strategies of agents for memory states of size $H = 5$ during the game. Given a memory state defined by a joint history, the color of the circles corresponds to the probability of generating the dominant convention (M), which is exact and remains constant between runs, regardless of the system state. The size of the circles corresponds to the average frequency of the memory state across 10 simulated runs. We study two settings: starting from an initial consensus on the dominant convention (M) we let the system (**A-B)** evolve naturally, and (**C-D**) evolve with perturbation of a single agent committed on the other convention (Q), both for 100 population rounds with $N = 24$. The panels on the left (right) show the dynamics for a population of **Llama-3-70B-Instruct** (**Llama-3-70B-Instruct**) agents.

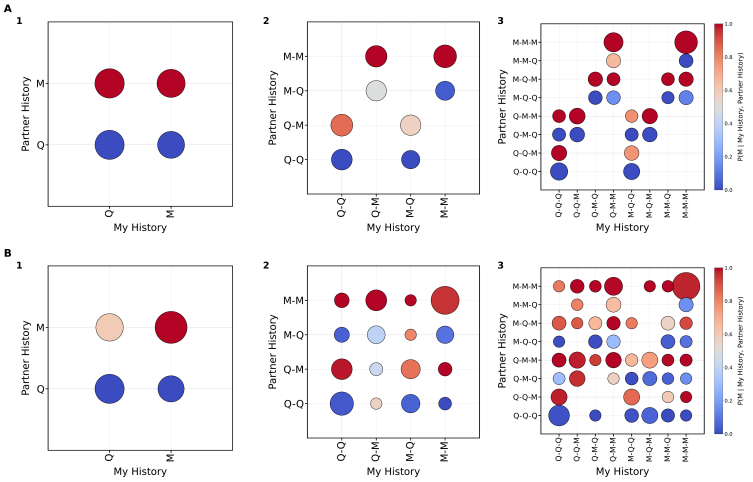

Figure S10: **Distribution of strategies across memory configurations in the first three interactions.** We study the emergence of consensus in **Llama-3-70B-Instruct** (top row, **A1-3**) and **Llama-3.1-70B-Instruct** (bottom row, **B1-3**), with the name pool $\{M, Q\}$. Given a memory state defined by a joint history, the color of the circles corresponds to the probability of generating the dominant convention (M), which is exact and remains constant between runs, regardless of the system state. The size of the circles corresponds to the average frequency of the memory state across 40 simulated runs. Note that without observing the frequency of the memory states during game dynamics, predicting a final consensus based on the probability distribution in round 2 (**F**) for **Llama-3.1-70B-Instruct** would result in an incorrect prediction: initially, each memory configuration is equally likely, so $p(Q) > p(M)$ on aggregate. However, we know that the bias flips as the game unfolds, and the population eventually converges on M.

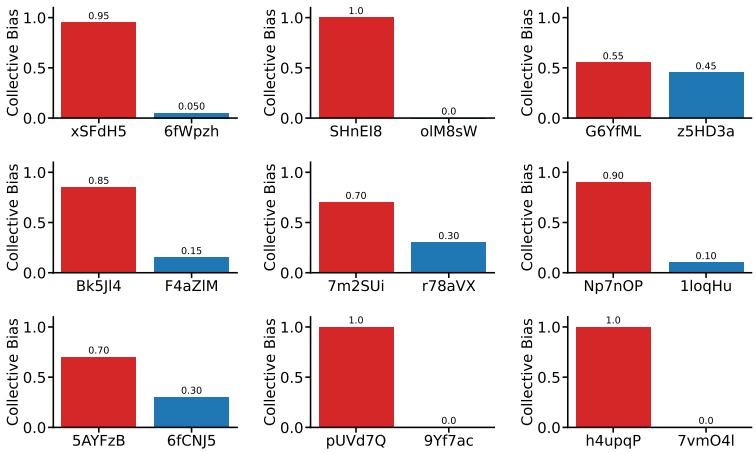

Figure S11: **Collective bias in a population of N=24 Llama-3-70B-Instruct using random convention pairs**. We simulated 20 runs for each pair using the same prompt in the main results. All pairs are initially unbiased, with a Jensen-Shannon distance < 0.005 from the neutral distribution.

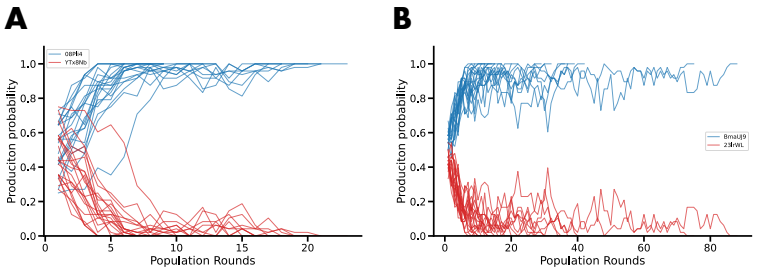

Figure S12: **Collective bias in a population of N=24 Llama-3-70B-Instruct agents using alternative prompts**. We tested the experimental framework using a new prompt template, namely (**A**) a narrative prompt, and (**B**) a bullet-point style prompt. For both prompts, we used a name pool of size $W = 2$, with random strings such that the initial, no-memory probability distribution over conventions had a Jensen-Shannon distance < 0.005 from the neutral distribution.

Model Name	Strong Convention	Weak Convention
Llama-3-70B-Instruct	2435	2565
Llama-3.1-70B-Instruct	5079	4921
Claude-3.5-Sonnet	5016	4984
Llama-2-70b-Chat	5010	4090

Table S1: **Raw individual bias.** Data shown for the left panel in Fig. 2B. Values indicate the counts of strong and weak productions in the case $W = 2$ for individual agents, showing their preferences *a priori*, when agents are initialized with empty memory.

Model Name	Strong Convention	Weak Convention
Llama-3-70B-Instruct	40	0
Llama-3.1-70B-Instruct	40	0
Claude-3.5-Sonnet	26	14
Llama-2-70b-Chat	36	4

Table S2: **Raw collective bias.** Data shown for the right panel in Fig. 2B. Values indicate the count of consensus states on the strong and weak conventions after the population converged. For each model, we conducted 40 trial runs. All models had memory length $H = 5$.

Model Name	Strong Convention	Weak Convention
Llama-3-70B-Instruct	6	1
Llama-3.1-70B-Instruct	10	0
Claude-3.5-Sonnet	5	5
Llama-2-70b-Chat	16	11

Table S3: **Raw Critical mass values.** Data shown for Fig. 3. The reported values corresponds to the number of agents required to overturn a majority consensus on the convention: M (strong Convention), and Q (Weak Convention). Population size (top to bottom), $N = 48, 24, 24, 24$. We performed 10 runs for each experiment, apart from **Llama-3-70B-Instruct**, which shows 3 runs.

Parameter	Value
Temperature	0.5
Top-K	10
Max Tokens	6

Table S4: Model Parameters.

Interaction	Memory Interaction: Played, Observed		p(M)	$p^{\dagger}$	$p^{\ddagger}$
2	**1:** Q, M		0.951	$< 2.2 \times 10^{-16}$	0.522
	1: M, Q		0.005	$< 2.2 \times 10^{-16}$	0.551
	1: Q, Q		0.003	$< 2.2 \times 10^{-16}$	0.644
	1: M, M		0.990	$< 2.2 \times 10^{-16}$	0.608
3	**1:** Q, M	**2:** M, Q	0.549	0.001	0.495
	1: M, Q	**2:** Q, M	0.848	$< 2.2 \times 10^{-16}$	0.513
	1: Q, M	**2:** M, M	1.000	$< 2.2 \times 10^{-16}$	1.000
	1: M, Q	**2:** Q, Q	0.004	$< 2.2 \times 10^{-16}$	0.461
	1: Q, Q	**2:** Q, M	0.936	$< 2.2 \times 10^{-16}$	0.490
	1: M, M	**2:** M, Q	0.159	$< 2.2 \times 10^{-16}$	0.512
	1: M, M	**2:** M, M	0.999	$< 2.2 \times 10^{-16}$	0.497
	1: Q, Q	**2:** Q, Q	0.011	$< 2.2 \times 10^{-16}$	0.494

Table S5: **Measuring the bias of the memory configurations in Table 1** In the two rightmost columns, we show the p-values for the null hypothesis that the model is unbiased ($p^{\dagger}$, rejected), and that the model's underlying bias is more extreme that our observation ($p^{\ddagger}$, insufficient evidence to reject).

Interaction	Memory Interaction: Played, Observed		P(A)	P(B)	Aggregated P(B)
1	-		0.494	0.506	0.506
2	1: B, A		0.287	0.713	0.630
	1: A, B		0.415	0.585	
	1: B, B		0.311	0.689	
	1: A, A		0.466	0.534	
3	1: B, A	2: A, B	0.308	0.692	0.592
	1: A, B	2: B, A	0.452	0.548	
	1: B, A	2: A, A	0.412	0.588	
	1: A, B	2: B, B	0.363	0.637	
	1: B, B	2: B, A	0.235	0.765	
	1: A, A	2: A, B	0.450	0.550	
	1: A, A	2: A, A	0.609	0.391	
	1: B, B	2: B, B	0.225	0.775	
	1: A, B	2: A, B	0.371	0.629	
	1: B, A	2: B, A	0.315	0.685	
	1: A, B	2: A, A	0.458	0.542	
	1: B, A	2: B, B	0.241	0.759	
	1: A, A	2: B, A	0.534	0.466	
	1: B, B	2: A, B	0.258	0.742	
	1: B, B	2: A, A	0.305	0.695	
	1: A, A	2: B, B	0.432	0.568	

Table S6: **Strategy bias in a LLM w/o Fine-Tuning.** The table reports the exact strategies of an LLM (**Llama-3.1-70B**) that has only undergone pre-training. We consider all possible memory configurations in the first three interaction rounds for the case $W = 2$, given that the model is initial unbiased in interaction 1. The pseudo-labels $\{A, B\}$ represent the real conventions, $\{XtmT2C, hsa1P6\}$. Probability distributions over conventions are extracted directly from the next-token probability distribution and are thus exact, rather than statistical estimates.

	Name	Question
Rules	`min_max`	What is the lowest/highest payoff player A can get in a single round?
	`actions`	Which actions is player A allowed to play?
	`payoff`	Which is player X's payoff in a single round if X plays p and Y plays q?
Time	`round`	Which is the current round of the game?
	`action`$_i$	Which action did player X play in round i?
	`points`$_i$	How many points did player X collect in round i?
State	`#actions`	How many times did player X choose p?
	`#points`	What is player X's current total payoff?

Table S7: **Meta-prompting questions.** Templates of prompt comprehension questions used in meta-prompting to verify the LLM's comprehension of the prompt.

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).