Collective Bias Emerges Even from Rational Social Learning

Authors: Bufan Gao[1], Xuechunzi Bai[1]*

Affiliations:

[1]Department of Psychology, University of Chicago, Chicago, IL 60637, USA.

*Corresponding author. Email: baix@uchicago.edu

Abstract:

Group decisions can outperform individual ones, but they can also fail. The same mechanism that fuels success -- rationally integrating information from others -- can also create bias. Using hiring in labor markets as a relevant context, we simulate networks of Bayesian learners, generative agents, and human participants making hiring decisions independently or collectively. In all three cases, integrating social information improves efficiency when one option is best, but also produces inequality: agents converge on a few options even though all are equally qualified. Our experiments show that collective bias can emerge purely from individually rational and fully transparent social learning, even without any loss of information. Pooling information compresses experiences, reduces exploration, and amplifies early randomness, revealing a general mechanism of emergent inequality.

One-sentence summary: Collective bias can emerge spontaneously even when everyone is rational and transparent, not despite social learning, but because of it.

Main Text

No individual mind alone can achieve what humanity has accomplished collectively. Learning from other minds enables groups of individuals to better memorize words (*1-2*), estimate objects (*3-5*), identify experts (*6,7*), and solve complex problems (*8-10*). Yet social learning can also lead the collective astray. It can produce conformity to incorrect judgments (*11,12*), spread spurious beliefs (*13*), elevate low-quality cultural trends (*14,15*), and propagate prejudiced attitudes (*16*). In explaining such failures, social learning offers a distinct level of analysis, beyond individual limitations and structural constraints (*17*): an in-between level that focuses on the interactive processes by which individuals learn from one another. At this level, existing accounts often attribute unsuccessful coordination to information loss: people rely on public information while ignore private signals (*15,18*), change information content due to memory constraints (*19-21*), or restrict interactions to like-minded others (*22-23*). These explanations assume collective failures arise from distortions already present during interaction. We argue otherwise: the very act of rational social learning can itself generate collective bias, even in the absence of preexisting conditions. In this view, collective bias is an emergent property.

The labor market offers a concrete and consequential example. Social groups are not equally distributed in the labor market, creating conditions for managers to statistically discriminate against people from different groups when individuating information is lacking (*24-26*). Social scientists tend to understand such inequality as the amplification of bias. At the individual level, managers prefer in-group candidates or rely on imperfect heuristics (*27-29*). At the structural level, initial market conditions exclude certain groups or restrict their access to resources (*30-32*). At the interactive level, worker information is opaque in the market exchanges, or referrals channel information to similar others (*33,34*). While these mechanisms are plausible, they frame inequality as something being amplified rather than created. In contrast, we propose that unequal representation in labor markets can emerge as a result of many rational interactive behaviors, such as decision-makers sharing, integrating, and using the information from their peers. This view parallels the origins of other suboptimal collective dynamics, where uncoordinated microscopic behaviors give rise to systematic patterns (*35-37*). It also points to a new class of interventions at the interactive level: to achieve a fairer labor market, policy should move beyond making information transparent or accurate. Instead, it should design network structures that diffuse valuable information efficiently while preserving independence and diversity.

To isolate the mechanism that rational social learning alone is sufficient to create collective bias, we model a simplified hiring market in which all candidate groups are equally productive and decision-makers are unbiased. In this stylized scenario, managers repeatedly evaluate candidates from several groups, where each decision is costly but informative: managers can learn more about a group by hiring it, but must balance exploring novel options with exploiting what has worked in the past (*38-40*). Managers may rely only on their own experience (asocial learning) or also incorporate the experiences of their peers (social learning). Concretely, we build a multi-agent framework and simulate three types of agents role-playing as hiring managers under one of the two conditions. These agents vary in their levels of rationality and transparency in how they make decisions: First, Bayesian learners are fully specified, rational, and transparent in their belief updating and decision making, ensuring any disparities cannot be attributed to existing bias. Second, large language model based generative agents incorporate extensive prior social knowledge and alignment goals, predicting how fairness-aware agents approach this challenge (*41*). Third, human participants recruited from online platforms provide rare empirical evidence to a literature dominated by non-human simulations (*4-6*). This design allows us to causally identify whether social learning alone can create labor market inequality, and to compare Bayesian models, real-world AI systems, and human participants under matched conditions.

Formally, we model the hiring process as a multi-agent, multi-armed bandit problem (*42;* Fig. 1). Each agent represents a hiring manager who selects from a fixed set of candidate groups, modeled as arms of a multiarmed bandit whose success probabilities are unknown. Each round, agents independently choose one group and observe a binary outcome of success or failure, which they use to update internal beliefs about candidate productivity (Fig. 1B). The objective is to maximize cumulative hiring success over time, which requires navigating the classic exploration–exploitation tradeoff (*43*): exploring uncertain groups may uncover better options but also carries the risk of poor outcomes, while exploiting known groups can yield safer returns. We represent the multi-agent system as an undirected graph, where each node corresponds to an agent and an edge connects two nodes if those agents share information (*42*). When agents are connected, they share their choices and corresponding outcomes from that round, without withholding or changing any information. This shared information is then incorporated into each agent's current beliefs, enabling learning not just from personal experience but also from peers.

We quantify two outcome metrics: efficiency and inequality (SM). Efficiency is defined as the total cumulative rewards earned by agents at the end of the study, reflecting how effectively the group of agents identifies and hires optimal candidates. Inequality is measured as the entropy of the final hiring distribution, which captures how random the decisions are. Lower entropy means hires are concentrated in a few groups, while higher entropy means hires are more evenly spread. To investigate the drivers of these outcomes, we manipulate two system features: reward distribution and market structure. The reward distribution determines whether candidate groups differ in productivity. In the unequal-productivity condition, one group on average is more productive than the others and the success rates are drawn from a uniform distribution between 0.1 and 0.9. In the equal-productivity condition, all groups are on average equally productive with the same success rate (SM). The other key variable, market structure, controls how information flows among agents. In the asocial learning condition, agents are isolated in the graph and update beliefs solely based on personal experiences (Fig. 1C). In the social learning condition, agents are fully connected and incorporate others' choices and outcomes into their own beliefs for future updates (Fig. 1A). This direct manipulation of information flow allows us to isolate the effects of social learning on market inequality from other forms of bias at the interactive level: because all information is shared transparently (*18*), in its original form without revision (*21*), or selective targeting (*23*), any differences between the two markets can therefore be attributed to the simple fact of whether information is shared.

Bayesian Learners. In the first set of experiments, we model agents using an explicitly defined Bayesian updating rule. Each agent uses the Thompson sampling algorithm to make decisions in which the agent samples from the posterior distribution of the success probability for each group and chooses the group with the highest sampled value (*44*). In each network, there are ten Bayesian agents hiring from ten candidate groups over 1000 rounds (SM). In the unequal-productivity condition, social learning substantially improves group-level efficiency (Fig. 2A). Pooling information from other agents yields 2.1% higher cumulative rewards than learning in isolation (b =2.10%, 95% CI [2.01%, 2.18%], p < 0.001). Agents in the social learning condition are able to discover the optimal group approximately 170 rounds earlier than those in the asocial learning condition. These results confirm the wisdom-of-crowds hypothesis (*2*): when one option is better, pooling information boosts efficiency and increases accuracy. However, when all groups are equally productive, learning from each other does not help (Figs. 2D, 2G): although market efficiency was comparable across conditions (*b* = 0.01%, 95% *CI* [-0.09%, 0.10%], *p* = 0.85), agents in the social learning condition create substantially more inequality than those in the asocial learning condition (*b* = –0.371, 95% *CI* [–0.411, –0.331], *p* < 0.001). To contextualize this effect, agents in the social learning condition generate a market in which some groups receive fewer than 1% of total hires, while over 50% of hires concentrate in a small subset of groups. In contrast, agents in the asocial learning condition create a more equal market, with each group receiving approximately 10% of hires and no single group exceeding 20% of total selections (Figs. 3A, 3D). In other words, rational Bayesian learners that learn from each other reach a misleading consensus on which group is optimal where none objectively exists.

Generative Agents. The explicitly defined belief updating and decision-making process in Bayesian learners improves interpretability, but it lacks realism. There is a burgeoning interest in using large language model based generative agents to simulate social dynamics (*40*), and we document what would happen in such a case. Moreover, newer models are trained to align with egalitarian values to be fair and unbiased (*45*) more than Bayesian learners, providing an interesting contrast. In the second set of experiments, we substitute Bayesian learners with OpenAI's GPT-4o, the newest model at the time of this experiment (*46*). We build a multi-agent system in which a central moderator interacts with each generative agent through structured language prompts. Each interaction begins with a system message establishing the task context and decision rules. Next, an assistant message reminds the agent of its previous choice and outcome. Finally, a user message presents feedback, either from the agent's own past decisions (asocial learning) or from the entire network (social learning), and prompts the agent to make a new decision. Because GPT-4o is not designed for numerical inference, the system translates observed results into preprocessed belief summaries over candidate groups (see full prompt designs in SM). Each network is configured with ten generative agents hiring from ten candidate groups over 200 rounds of decisions (SM). Overall, the patterns we observe are consistent with and even more pronounced than those for Bayesian learners. In the unequal-productivity condition, learning from other agents helps the group of agent managers to find the optimal candidate much faster, yielding 4.75% (95% *CI* [3.11%, 5.90%], *p* < 0.001) higher cumulative rewards as compared to asocial learning condition. Social learning improves efficiency (Fig. 2B). However, when the average productivities of the candidate groups are identical, social learning creates inequality (Figs. 2E, 2H): agent managers create a labor market that shows 58.4% lower entropy than those who learn in isolation (*b* = –1.713, 95% *CI* [–1.876, –1.550], *p* < 0.001). To contextualize this effect, manager agents in the social learning condition hire about 80% of the same group despite all candidates being equally productive. In contrast, manager agents in the asocial learning condition hired more evenly with each group receiving 5–15% of total hires (Figs. 3B, 3E). In addition to demonstrating how social learning can limit the system from exploring further, this study suggests that emergent collective biases can be orders of magnitude more severe in value-aligned generative agents as compared to Bayesian learners, raising fairness concerns for future adoption of AI systems in dynamic environments such as hiring.

Human Participants. In the final set of experiments, we simulate a collaborative hiring market with online human participants (*N* = 2,000). We design an incentive-compatible, multi-player multi-round game in which participants role-play as part of a hiring committee. Their goal is to make good hiring decisions, and the hiring outcomes are translated into their actual bonus payments (SM). Each network is configured with ten hiring managers, and they need to choose from ten groups of job candidates whose group identity is defined by arbitrary color of their icon in a total of 50 rounds of decisions (SM). Participants are randomly assigned to one of the 2-by-2 conditions: equal versus unequal productivity and asocial versus social learning. The sample has a mean age of 37.8 years, is 53.8% female, and is racially representative of the larger American online population (70.6% White, 9.7% Black, 6.3% Latinx, 6.2% Asian). These demographic variables are balanced across conditions, thus do not contribute to the observed effects (SM). Consistent with our previous observations, social learning creates inequality in the simulated labor market when all candidates are equally productive. Specifically, when there is one optimal group, learning from other participants indeed helps the group to identify the optimal candidates much faster, resulting in earning 17.01% higher cumulative rewards than those in the asocial learning condition (95% *CI* [11.78%, 21.88%], $p < 0.001$; Fig. 2C). However, when there is no single optimal group, learning from other participants does not help (Figs. 2F, 2I). On the metric of inequality, entropy at the final round is 48.5% lower in the social learning condition than in the asocial learning condition ($b = -1.496$, 95% *CI* [–1.636, –1.356], $p < 0.001$), indicating that participants concentrate heavily on limited number of groups of candidates in their hiring decisions rather than exploring all groups equally. In over 78% of trials, a single group received more than 50% of final-round hires (Figs. 3C, 3F). Note that this experiment runs on a relatively short horizon of 50 rounds, nonetheless, unequal allocation of hiring decisions emerges early. Participants' reflections in the social learning condition reveal how pooled information guided their decisions (SM). Several note adapting to group-level trends: "…watched which one the other people were voting for and followed their lead…", "…one option quickly emerged as the most productive…", and "…virtually every other person was selecting that person helped to solidify my choices." By comparison, participants in the asocial learning condition emphasize independent learning from direct outcomes. Comments highlight exploration, tracking personal success rates, and forming individual beliefs: "…early on it was trial-and-error to test out several groups…", "…I tried to diversify my selections during the first several rounds…", and "…colors that gave immediate bad feedback made me not want to select that color again."

In short, across all three simulations, unequal group shares in the hiring market emerge endogenously. When there is one optimal group whose candidates are most productive, learning from others enables agents to identify it more quickly, increasing efficiency. However, when all candidate groups are equally productive, the same process causes stagnation: Bayesian learners, generative agents, and human participants coordinate in a fully connected network have created inequality, not despite sharing information, but because of it. To understand boundary conditions, we systematically varied sampling strategies, the number of agents and groups, the agents' priors, the baseline productivity of the candidate groups, and the initial states (*SM*). We found that specific parameterizations may affect the magnitude of inequality, but not its direction, and social learning consistently creates collective bias. When information is shared, even small early differences in rewards across arms become group-wide signals, pulling all agents toward whichever arm appears slightly better. This synchronization reduces exploration, concentrates choices, and drives down entropy. At the same time, adopting exploratory sampling strategies, introducing optimistic priors, and increasing the number of groups all helped to increase exploration and thus reduce, though not eliminate, collective bias.

Discussion. Our studies are clearly unlike real labor markets in several ways. For example, we expect other biases – individual preferences, decision heuristics, structural barriers, and loss of information in transmission – all to play important roles, amplifying the effects we observe. We also suspect that markets vary in their interconnectedness, whereas our data capture only the two extremes. Although these differences limit the immediate relevance of our studies to real-world labor markets, our findings nevertheless suggest that labor market inequality, one consequential form of collective bias, can emerge naturally from the simple opportunity to learn from others. This pattern aligns with, though does not prove, observations in Mexican migrant networks (*47*), where jobseekers begin with diverse occupational choices, but later, through increasing peer-based job seeking, cluster into a limited set of agricultural and unskilled jobs. Admittedly, any of these biases could contribute to such stratification, but our data suggest that even in the absence of existing bias, rationally integrating the experiences of others is sufficient.

This mechanism provides novel policy insights. When inequality is framed as an amplification of existing bias, policies typically aim to revise those biases (*48*). At the individual level, skill training programs aim to equip workers with desired skills, and decision-making workshops aim to change cognitive biases in managers' minds. At the structural level, promoting minority candidates serves to remedy an already-tainted market composition. At the interactive level, making private information publicly available aims to reduce the side effects of information cascades (*49*). Our work offers a different takeaway: even when job candidates are equally qualified, the market has balanced representations and is fully connected, and agents are Bayesian-rational who share what they know with full transparency, inequality can still emerge. Pooling information compresses experiences, reduces exploration, and amplifies early random fluctuations. The challenge, then, is not simply to correct known biases, but to design networks that compress information efficiently without crowding out exploration.

We have emphasized consistency across three types of agents, but their differences are equally revealing. First, neither generative agents nor human participants match the rational behavior of Bayesian learners. While it is qualitatively true that there are main treatment effects across three agents, this gap indicates that in our setup, neither generative agents nor human participants behave in a perfectly Bayesian way. Future work can study mechanisms to close this gap. Second, generative agents in our study produce the greatest market inequality with the least variation, outperforming human participants and Bayesian learners in this regard. Their heightened sensitivity to early feedback and consistent reluctance to explore despite value alignment suggest that artificial intelligence agent systems, especially those in socially networked dynamic environments, may be particularly prone to self-reinforcing patterns. Beyond existing de-biasing strategies focused on updating pretraining data, adding finetuning data, or adjusting decision thresholds (*45*), our findings highlight the need to preemptively design network structures that can maintain efficient yet equitable information flow, which will become increasingly important when these agents interact dynamically with the world.

References and notes

1. D. M. Wegner, "Transactive memory: A contemporary analysis of the group mind" in *Theories of Group Behavior*, B. Mullen, G. R. Goethals, Eds. (Springer, 1987), pp. 185–208.
2. G. D. Greeley, V. Chan, H. Y. Choi, S. Rajaram, Collaborative recall and the construction of collective memory organization: The impact of group structure. *Top. Cogn. Sci.* **16**, 282–301 (2024).
3. F. Galton, Vox populi. *Nature* **75**, 450–451 (1907).
4. A. Almaatouq, A. Alotaibi, M. Radaelli, E. Pentland, A. G. Pentland, Adaptive social networks promote the wisdom of crowds. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 11379–11386 (2020).
5. M. D. Hardy, B. D. Thompson, P. M. Krafft, T. L. Griffiths, Resampling reduces bias amplification in experimental social networks. *Nat. Hum. Behav.* **7**, 2084–2098 (2023).
6. B. Thompson, B. van Opheusden, T. Sumers, T. L. Griffiths, Complex cognitive algorithms preserved by selective social learning in experimental populations. *Science* **376**, 95–98 (2022).

7. J. Henrich, F. J. Gil-White, The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evol. Hum. Behav.* **22**, 165–196 (2001).
8. A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, T. W. Malone, Evidence for a collective intelligence factor in the performance of human groups. *Science* **330**, 686–688 (2010).
9. R. L. Goldstone, T. M. Gureckis, Collective behavior. *Top. Cogn. Sci.* **1**, 412–438 (2009).
10. Y. Xiang, N. Vélez, S. J. Gershman, Collaborative decision making is grounded in representations of other people's competence and effort. *J. Exp. Psychol. Gen.* **152**, 1565–1579 (2023).
11. S. E. Asch, "Effects of group pressure upon the modification and distortion of judgments" in *Groups, Leadership and Men: Research in Human Relations*, H. Guetzkow, Ed. (Carnegie Press, 1951), pp. 177–190.
12. R. M. Krauss, M. Deutsch, Communication in interpersonal bargaining. *J. Pers. Soc. Psychol.* **4**, 572–577 (1966).
13. G. Pennycook, D. G. Rand, The psychology of fake news. *Trends Cogn. Sci.* **25**, 388–402 (2021).
14. M. J. Salganik, P. S. Dodds, D. J. Watts, Experimental study of inequality and unpredictability in an artificial cultural market. *Science* **311**, 854–856 (2006).
15. S. Bikhchandani, D. Hirshleifer, I. Welch, A theory of fads, fashion, custom, and cultural change as informational cascades. *J. Polit. Econ.* **100**, 992–1026 (1992).
16. D. T. Schultner, B. R. Lindström, M. Cikara, D. M. Amodio, Transmission of social bias through observational learning. *Sci. Adv.* **10**, eadk2030 (2024).
17. X. Bai, T. L. Griffiths, S. T. Fiske, Costly exploration produces stereotypes with dimensions of warmth and competence. *J. Exp. Psychol. Gen.* (2024).
18. D. Easley, J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World* (Cambridge University Press, 2010).
19. F. C. Bartlett, *Remembering: A Study in Experimental and Social Psychology* (Cambridge University Press, 1932).
20. T. L. Griffiths, M. L. Kalish, S. Lewandowsky, Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 3503–3514 (2008).
21. A. Lyons, Y. Kashima, How are stereotypes maintained through communication? The influence of stereotype sharedness. *J. Pers. Soc. Psychol.* **85**, 989–1005 (2003).
22. Z. Kunda, The case for motivated reasoning. *Psychol. Bull.* **108**, 480–498 (1990).
23. M. McPherson, L. Smith-Lovin, J. M. Cook, Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.* **27**, 415–444 (2001).
24. E. S. Phelps, The statistical theory of racism and sexism. *Am. Econ. Rev.* **62**, 659–661 (1972).
25. K. J. Arrow, "The theory of discrimination" in *Discrimination in Labor Markets*, O. Ashenfelter, A. Rees, Eds. (Princeton University Press, 1973), pp. 3–33.
26. S. Coate, G. C. Loury, Will affirmative-action policies eliminate negative stereotypes? *Am. Econ. Rev.* **83**, 1220–1240 (1993).
27. G. W. Allport, *The Nature of Prejudice* (Addison-Wesley, 1954).
28. G. S. Becker, *The Economics of Discrimination* (University of Chicago Press, 1957).
29. J. A. Bohren, A. Imas, M. Rosenberg, The dynamics of discrimination: Theory and evidence. *Am. Econ. Rev.* **109**, 3395–3436 (2019).
30. A. H. Eagly, V. J. Steffen, Gender stereotypes stem from the distribution of women and men into social roles. *J. Pers. Soc. Psychol.* **46**, 735–754 (1984).

31. D. Pager, H. Shepherd, The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annu. Rev. Sociol.* **34**, 181–209 (2008).
32. E. Bonacich, A theory of ethnic antagonism: The split labor market. *Am. Sociol. Rev.* **37**, 547–559 (1972).
33. J. L. Doleac, B. Hansen, The unintended consequences of "ban the box": Statistical discrimination and employment outcomes when criminal histories are hidden. *J. Labor Econ.* **38**, 321–374 (2020).
34. B. Rubineau, R. M. Fernandez, Missing links: Referrer behavior and job segregation. *Manage. Sci.* **59**, 2470–2489 (2013).
35. T. C. Schelling, Dynamic models of segregation. *J. Math. Sociol.* **1**, 143–186 (1971).
36. D. Centola, J. Becker, D. Brackbill, A. Baronchelli*, Experimental evidence for tipping points in social convention. Science* **360**, 1116–1119 (2018).
37. A. F. Ashery, L. M. Aiello, A. Baronchelli, *Emergent social conventions and collective bias in LLM populations. Sci. Adv.* **11**, eadu9368 (2025).
38. X. Bai, S. T. Fiske, T. L. Griffiths, Globally inaccurate stereotypes can result from locally adaptive exploration. *Psychol. Sci.* **33**, 671–684 (2022).
39. D. Li, L. Raymond, P. Bergman, Hiring as exploration. *Rev. Econ. Stud.* **92**, 1–41 (2025).
40. J. Komiyama, S. Noda, On statistical discrimination as a failure of social learning: A multi-armed bandit approach. *Manage. Sci.* **70**, 1–15 (2024).
41. J. S. Park, C. Q. Zou, A. Shaw, B. M. Hill, C. Cai, M. R. Morris, R. Willer, P. Liang, M. S. Bernstein, Generative agent simulations of 1,000 people. *arXiv:2411.10109* (2024).
42. P. Landgren, V. Srivastava, N. E. Leonard, Distributed cooperative decision making in multi-agent multi-armed bandits. *Automatica* **125**, 109445 (2021).
43. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, ed. 1, 1998).
44. W. R. Thompson, On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25**, 285–294 (1933).
45. Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, et al., Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv:2204.05862 (2022).
46. A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. J. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al., GPT-4o system card. arXiv:2410.21276 (2024).
47. K. Munshi, Networks in the modern economy: Mexican migrants in the US labor market. *Q. J. Econ.* **118**, 549–599 (2003).
48. H. Fang, A. Moro, Theories of statistical discrimination and affirmative action: A survey. Handb. Soc. Econ. **1**, 133–200 (2011).
49. D. Kübler, G. Weizsäcker, Information cascades in the labor market. J. Econ. **80**, 211–229 (2003).

Acknowledgements

Funding: This work was supported by the Becker Friedman Institute and start-up funds at the University of Chicago to X.B. **Author contributions:** Conceptualization: B.G., X.B., Methodology: B.G., X.B., Investigation: B.G., X.B., Visualization: B.G., X.B., Writing: B.G., X.B. **Competing interests:** None declared. **Data and materials availability:** Data, materials, and study registrations are available through the Open Science Framework at https://osf.io/58nbt/. All other data needed to evaluate the conclusions in the paper are present in the paper or the supplementary materials.

Supplementary Materials

Materials and Methods
Supplementary Text
Figs. S1 to S14.
Tables S1 to S3.

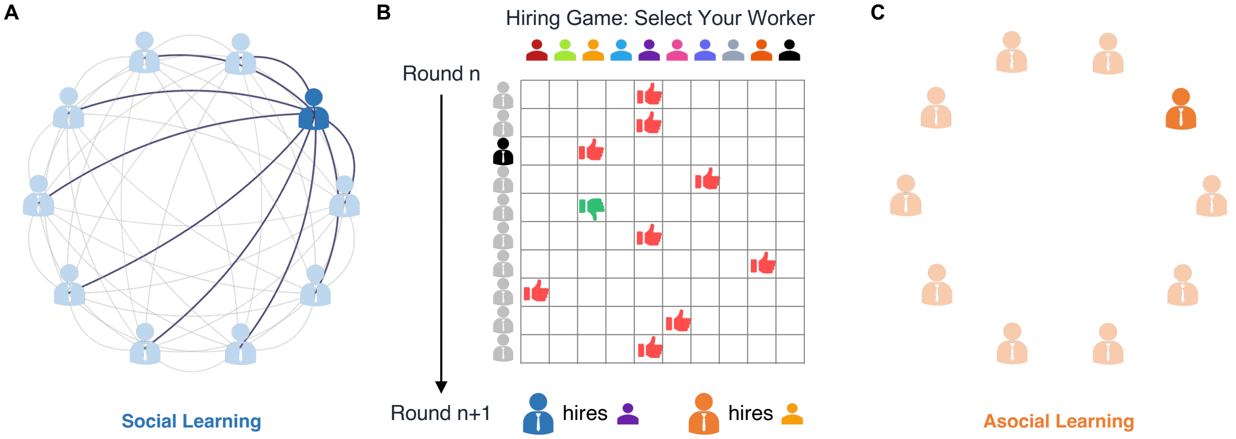


Fig. 1. Collective hiring as a multi-agent multi-armed bandit: The task is modeled as a multi-agent multi-armed bandit problem, where each hiring manager (agent, denoted as V in **A** and **C**) selects from K candidate groups (arm, denoted as colored human icon in **B**) with unknown success rates μ_α, modeled as i.i.d. Bernoulli variables, aiming to maximize cumulative reward in sequential decisions. **(A)** Social learning. Managers observe choices and outcomes of everyone in their group in a fully connected graph connecting agents V via edges E, denoted $G = (V, E)$, to update their beliefs on each arms' successes and failures, denoted $S = (\alpha_a, \beta_a)$. **(C)** Asocial learning. Managers observe choices and outcomes of their own in a disconnected graph, denoted $G = (V, \emptyset)$, to update their beliefs. **(B)** Hiring task. Managers repeatedly make hiring decisions by selecting one of ten candidates. Each round, they choose one candidate from a group and receive binary feedback: success shown as thumbs up or failure shown as thumbs-down. Managers are Bayesian learners, generative agents, and human participants, respectively.

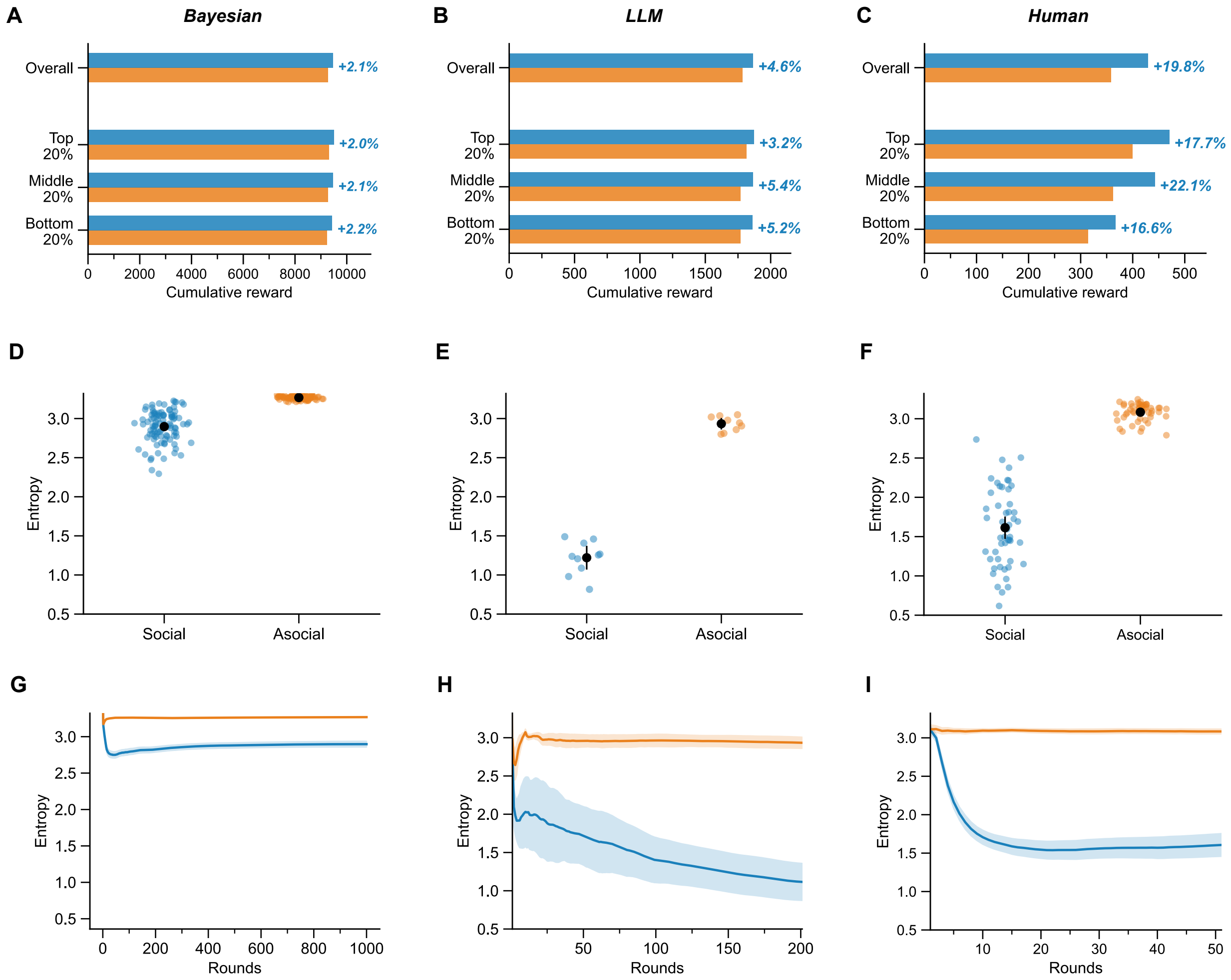


Fig. 2. Experimental results from Bayesian learners, generative agents, and human participants: (A–C) When there is one optimal candidate group, social learning improves efficiency in all environments. We plot relative performance gains between the two conditions with overall, top 20%, middle 20%, and bottom 20% stratified runs. (A) Bayesian learners earn about 2% more cumulative rewards in all quantiles ($p < 0.001$), indicating stable improvements. (B) GPT-4o agents show larger gains, averaging 4–5% ($p < 0.001$), with slightly stronger improvements in the top and bottom subsets. (C) Human participants achieve the largest increases, about 20% more rewards ($p < 0.001$), with consistently strong gains across quantiles. **(D–F)** When there is no single optimal group, social learning consistently increases inequality. (D) Bayesian learners see a decrease in entropy of 0.371 (11.35%; $b = –0.371$, 95% CI = [–0.411, –0.331], $p < 0.001$). (E) GPT-4o agents exhibit a sharper drop of 1.713 (58.39%; $b = –1.713$, 95% CI = [–1.876, –1.550], $p < 0.001$). (F) Human participants show a reduction of 1.496 (48.52%; $b = –1.496$, 95% CI = [–1.636, –1.356], $p < 0.001$), with over 78% of trials ending in majority hiring of a single group. **(G–I)** Entropy dynamics over time. (G) Bayesian learners show an initial rise, then decline 5% after round 20, stabilizing at 11.35% lower in the social than the asocial learning. (H) GPT-4o agents decline monotonically, ending with entropy 2.35× lower than Bayesian learners. (I) Human participants show a similar pattern, ending with entropy 2.12× lower in the social learning than in the asocial learning condition.

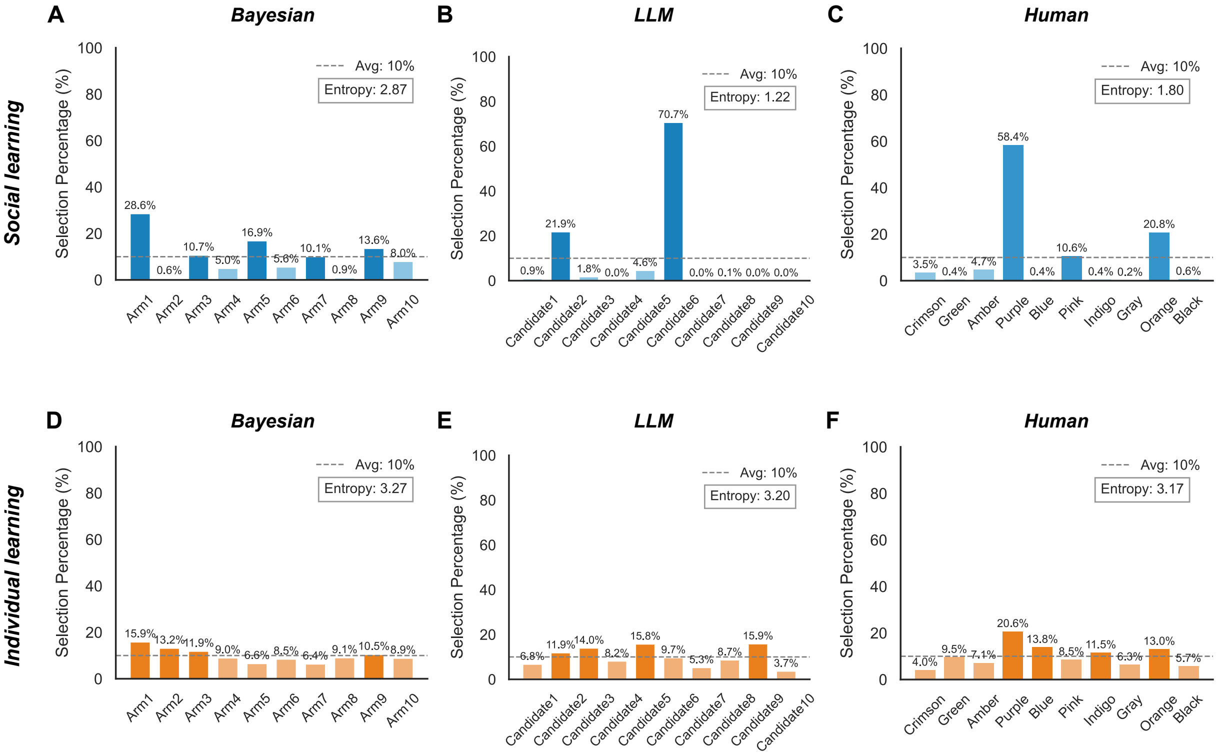


Fig. 3. Example final-round entropy from representative simulation runs for each agent type under social learning **(A–C)** and asocial learning **(D–F)** conditions. Entropy reflects the distributional balance of hiring decisions across candidate groups, with lower values indicating greater concentration. Under asocial learning **(D–F)**, markets remain relatively balanced. Under social learning **(A–C)**, however, inequality emerges in all agents. **(A)** Bayesian learners retain more exploratory behavior, yet still produce concentrated outcomes. **(B)** GPT-4o agents and **(C)** human participants show rapid convergence on one or two groups, reflecting stark reductions in diversity and early lock-in to dominant options.

Supplementary Materials for
Collective Bias Emerges Even from Rational Social Learning

Bufan Gao[1], Xuechunzi Bai[1]*

[1]Department of Psychology, University of Chicago, Chicago, IL 60637, USA.
*Corresponding author. Email: baix@uchicago.edu

This PDF file includes:

Materials and Methods
Supplementary Text
Figures S1 to S14
Tables S1 to S3
References

Other Supplementary Materials for this manuscript:

MDAR Reproducibility Checklist

Abstract:

This file contains supplementary materials for “Collective bias emerges even from rational social learning.” It is intended as a reference for readers seeking information on specific topics and not to be read from beginning to end. The materials cover mathematical, implementational, and empirical details related to the general formalism, key metrics, Bayesian learners, generative agents, and human participants discussed in the main text.

Contents

1 Formalism

1.1 Multi-Armed Bandit (MAB)

We begin with the classical stochastic multi-armed bandit (MAB) problem, which formalizes sequential decision-making under uncertainty [9]. Imagine a decision-maker repeatedly choosing among several options, often called "arms," each with an unknown payoff. With every choice, the agent receives some reward and gradually learns about which arms are better. The central challenge is balancing exploration (trying new or uncertain arms to gain information) with exploitation (sticking to the best-know arm to maximize rewards).

Single-agent structure. Formally, consider a single agent interacting with a set of N arms, indexed by $k \in \{1, \ldots, N\}$, over a finite horizon of T rounds, indexed by $t \in \{1, \ldots, T\}$. At each round t, the agent selects one arm $a_t \in \{1, \ldots, N\}$ and observes a binary reward

$$r_t \sim \text{Bernoulli}(\mu_{a_t}),$$

where the reward distribution of arm k is a bounded random variable in $[0, 1]$ with unknown mean μ_k. The agent's objective is to maximize its cumulative reward

$$\sum_{t=1}^{T} r_t,$$

equivalently to minimize the expected cumulative regret

$$R(T) = T\mu^* - \sum_{t=1}^{T} \mathbb{E}[r_t],$$

where $\mu^* = \max_k \mu_k$ is the mean reward of the optimal arm.

Bayesian inference. In the main text, we focus on how Bayesian rational learners approach this problem, building on prior research [2]. In the Bayesian formulation, the agent maintains for each arm k a posterior distribution over the success probability μ_k. This posterior is represented as

$$\mu_k \sim \text{Beta}(\alpha_k^t, \beta_k^t),$$

where α_k^t and β_k^t denote the cumulative number of observed successes and failures of arm k up to round t. Parameters are initialized with an uninformative prior $\alpha_k^0 = \beta_k^0 = 1$. After observing the reward r_t from the selected arm a_t, the posterior parameters are updated as

$$\alpha_{a_t}^{t+1} = \alpha_{a_t}^t + \mathbf{1}\{r_t = 1\}, \qquad \beta_{a_t}^{t+1} = \beta_{a_t}^t + \mathbf{1}\{r_t = 0\}.$$

1.2 Multi-Agent Multi-Armed bandit (MAMAB)

Core to this research, we extend the single-agent MAB framework to a setting with many decision-makers, also known as distributed decision-making [7]. Instead of one individual agent in isolation, we now consider a system of M agents, indexed by $i \in \{1, \ldots, M\}$, who all interact with the same set of N arms. Each agent repeatedly faces the same basic tradeoff of whether to explore or exploit, but their choices and outcomes can be influenced by the presence of others.

Multi-agent structure. Formally, at each round t, agent i selects an arm $a_{i,t}$ and receives a binary reward

$$r_{i,t} \sim \text{Bernoulli}(\mu_{a_{i,t}}).$$

Distinctive in this setup is the flow of information among agents. We represent this by an undirected graph $G = (V, E)$, where nodes $V = \{1, \ldots, M\}$ correspond to agents and edges E denote bidirectional information flow. That is, if $(i, j) \in E$, then agent i observes not only its own action and reward, but also those of agent j, and vice versa, see Figure S1. Thus, the observation set available to agent i at round t is

$$O_{i,t} = \{(a_{i,t}, r_{i,t})\} \cup \{(a_{j,t}, r_{j,t}) : j \in \mathcal{N}(i)\},$$

where $\mathcal{N}(i)$ denotes the set of agents directly connected to agent i in the graph G.

$$\alpha_{i,k}^{t+1} = \alpha_{i,k}^{t} + \sum_{(a,r)\in O_{i,t}} \mathbf{1}\{a = k, r = 1\}, \qquad \beta_{i,k}^{t+1} = \beta_{i,k}^{t} + \sum_{(a,r)\in O_{i,t}} \mathbf{1}\{a = k, r = 0\}.$$

Here, the summations run over all observations in $O_{i,t}$. The indicator function $\mathbf{1}\cdot$ adds one whenever an observed action corresponds to arm k and returns the specified reward (success for α and failure for β). In other words, each agent counts how many successes and failures it has seen for each arm, combining its own data with that of its linked neighbors.

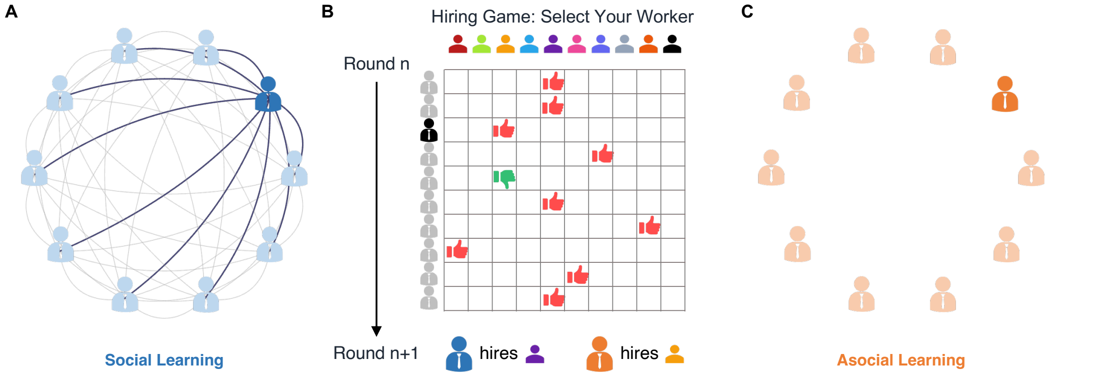


Figure S1: Adapted from the main text where we construct collective hiring decisions as a multi-agent multi-armed bandit problem. A. and C. are the general structures of the agents (multi-agent), and B is the sequential decision process (multi-armed bandit).

2 Metrics

2.1 Efficiency

We evaluate group performance in terms of efficiency, defined as the cumulative reward obtained by all agents over the horizon T. Formally,

$$\mathrm{Eff}(T) = \sum_{i=1}^{M} \sum_{t=1}^{T} r_{i,t}.$$

Equivalently, efficiency can be expressed as the expected cumulative reward,

$$\mathbb{E}[\mathrm{Eff}(T)] = \sum_{i=1}^{M} \sum_{t=1}^{T} \mu_{a_{i,t}},$$

which captures the system's ability to maximize long-term gains. High efficiency indicates that agents are collectively identifying and exploiting productive options, whereas low efficiency reflects wasted opportunities or persistent exploration of suboptimal arms.

Relative efficiency. To isolate the effect of information sharing, we compare efficiency under social learning to efficiency under asocial learning. The relative efficiency gain is defined as

$$\Delta\mathrm{Eff}(T) = \mathrm{Eff}^{\mathrm{social}}(T) - \mathrm{Eff}^{\mathrm{asocial}}(T),$$

or, in normalized form,

$$\frac{\mathrm{Eff}^{\mathrm{social}}(T)}{\mathrm{Eff}^{\mathrm{asocial}}(T)} - 1.$$

This quantity captures the improvement in group performance that arises from social learning. In other words, a positive value indicates that the group obtains more rewards when the agents learn from each other than when they rely only on individual experiences, whereas a negative value indicates that information sharing reduces performance, and a value near zero suggests little to no measurable benefit from social learning.

2.2 Inequality

We evaluate bias by quantifying inequalities in allocation across arms, operationalized as the empirical distribution of cumulative arm selections, i.e., the relative frequencies with which each arm is chosen over horizon T. Let

$$n_k(T) = \sum_{i=1}^{M} \sum_{t=1}^{T} \mathbf{1}\{a_{i,t} = k\}$$

be the total number of times arm k is selected by the group up to time T. Using this, we define the empirical allocation distribution over arms as

$$p_k(T) = \frac{n_k(T)}{\sum_{j=1}^{N} n_j(T)} = \frac{n_k(T)}{MT}, \qquad k = 1, \dots, N.$$

where $p_k(T)$ represents the relative frequency with which arm k is chosen. We measure inequality via the Shannon entropy of this distribution

$$H_{\text{arm}}(T) = -\sum_{k=1}^{N} p_k(T) \log p_k(T).$$

Entropy attains its maximum $\log N$ when selections are perfectly balanced across arms and decreases as selections become concentrated. Lower $H_{\text{arm}}(T)$ therefore indicates greater concentration on a subset of arms.

Relative inequality. To isolate the effect of social learning, we compare entropy under social learning to asocial learning. We report the difference scores between the two conditions

$$\Delta H_{\text{arm}}(T) = H_{\text{arm}}^{\text{social}}(T) - H_{\text{arm}}^{\text{asocial}}(T),$$

and the normalized ratio

$$\frac{H_{\text{arm}}^{\text{social}}(T)}{H_{\text{arm}}^{\text{asocial}}(T)} - 1.$$

Hence, a negative $\Delta H_{\text{arm}}(T)$ indicates that social learning increases concentration relative to the asocial case, thus more likely to generate collective-level bias.

3 Bayesian Learners

We conducted a series of ablation studies to examine how different parameters in the Bayesian simulations influence the outcomes of our design. In particular, we varied sampling strategies (Section 3.1), scaled the number of agents and arms (Section 3.2), tested different agent priors (Section 3.3), manipulated levels of ground truth productivity (Section 3.4), and sampled across all initial states of the network (Section 3.5). In a nutshell, while these specifications affected the *magnitude* of the core phenomenon, they did not alter its *direction*: across a wide range of modeling assumptions, social learning consistently generated lower entropy, that is, greater collective bias, than asocial learning.

3.1 Varying sampling strategies

Overview. There are different sampling algorithms to solve the multi-armed bandit problem. We studied three commonly used strategies – Thompson sampling (TS) [8], ε-greedy [9], and upper confidence bound (UCB) [6] – to understand how they influence the behaviors of the network.

3.1.1 Thompson sampling

In Thompson sampling, each agent maintains a posterior distribution over the success probability of each arm k based on past observations. At each round, the agent samples a parameter value

$$\theta_{i,k}^t \sim \mathrm{Beta}(\alpha_{i,k}^t, \beta_{i,k}^t)$$

for every arm k and selects the arm with the largest sampled value, breaking ties uniformly at random. The full procedure is summarized in Algorithm 1. The agent then updates their posterior distributions following the observation and updating rules defined in Section 1.

Algorithm 1 Thompson sampling

```
Initialize priors $\{\alpha_k^0, \beta_k^0\}_{k=1}^N$.
for $t = 1$ to $T$ do
    for each arm $k$ do
        Sample $\theta_k^t \sim \mathrm{Beta}(\alpha_k^t, \beta_k^t)$.
    end for
    Select $a_t \in \arg\max_k \theta_k^t$                ▷ break ties uniformly
    Proceed with observation and belief update.
end for
```

Given a Beta-Bernoulli model, the reward probability θ is assumed to follow $\mathrm{Beta}(\alpha, \beta)$ prior. The mean of this distribution is

$$\mathbb{E}[\theta] = \mu = \frac{\alpha}{\alpha + \beta},$$

and the variance is

$$\mathrm{Var}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

For large $\alpha + \beta$, the variance can be approximated as

$$\mathrm{Var}(\theta) \approx \frac{\mu(1-\mu)}{n}, \quad \mu = \frac{\alpha}{\alpha+\beta}, \; n = \alpha + \beta$$

Intuitively, this mean reflects the current best estimate of the reward probability, while the variance captures uncertainty, shrinking as the effective sample size n grows.

3.1.2 ε-greedy

In ε-greedy, each agent selects the arm with the highest posterior mean with probability $1 - \varepsilon$, and explores uniformly at random among all arms with probability ε. The posterior mean of arm k at round t is given by

$$\mu_{i,k}^t = \frac{\alpha_{i,k}^t}{\alpha_{i,k}^t + \beta_{i,k}^t}.$$

The procedure is summarized in Algorithm 2, followed by the updating rules defined in Section 1.

Algorithm 2 ε-greedy

```
Initialize priors $\{\alpha_k^0, \beta_k^0\}_{k=1}^N$.
for $t = 1$ to $T$ do
    for each arm $k$ do
        Compute posterior mean $\mu_k^t = \frac{\alpha_k^t}{\alpha_k^t + \beta_k^t}$.
    end for
    if Uniform$(0, 1) < \varepsilon$ then
        Select $a_t$ uniformly at random from $\{1, \ldots, N\}$.          ▹ Explore
    else
        Select $a_t \in \arg\max_k \mu_k^t$          ▹ Exploit
    end if
    Proceed with observation and belief update.
end for
```

Intuitively, this agent chooses an option it currently believes is best, but with a small probability ε, it instead picks a random option. A bigger ε means more exploration.

3.1.3 Upper confidence bound

In upper confidence bound, each agent selects the arm with the largest index formed by its posterior mean and an exploration bonus. For arm k at round t, the posterior mean is

$$\mu_k^t = \frac{\alpha_k^t}{\alpha_k^t + \beta_k^t}.$$

We define the effective number of pulls for arm k directly from the Beta–Bernoulli sufficient statistics,

$$n_k(t) = \alpha_k^t + \beta_k^t,$$

so that the pull count is consistent with the belief-update process. The exploration bonus is then

$$c_k(t) = \sqrt{\frac{2 \ln t}{\max\{1, n_k(t)\}}},$$

and the sampling index is

$$\mathrm{UCB}_k^t = \mu_k^t + c_k(t).$$

At each round, the agent selects the arm with the largest index, breaking ties uniformly at random. The procedure is summarized in Algorithm 3, followed by the updating rules defined in Section 1.

Algorithm 3 Upper confidence bound

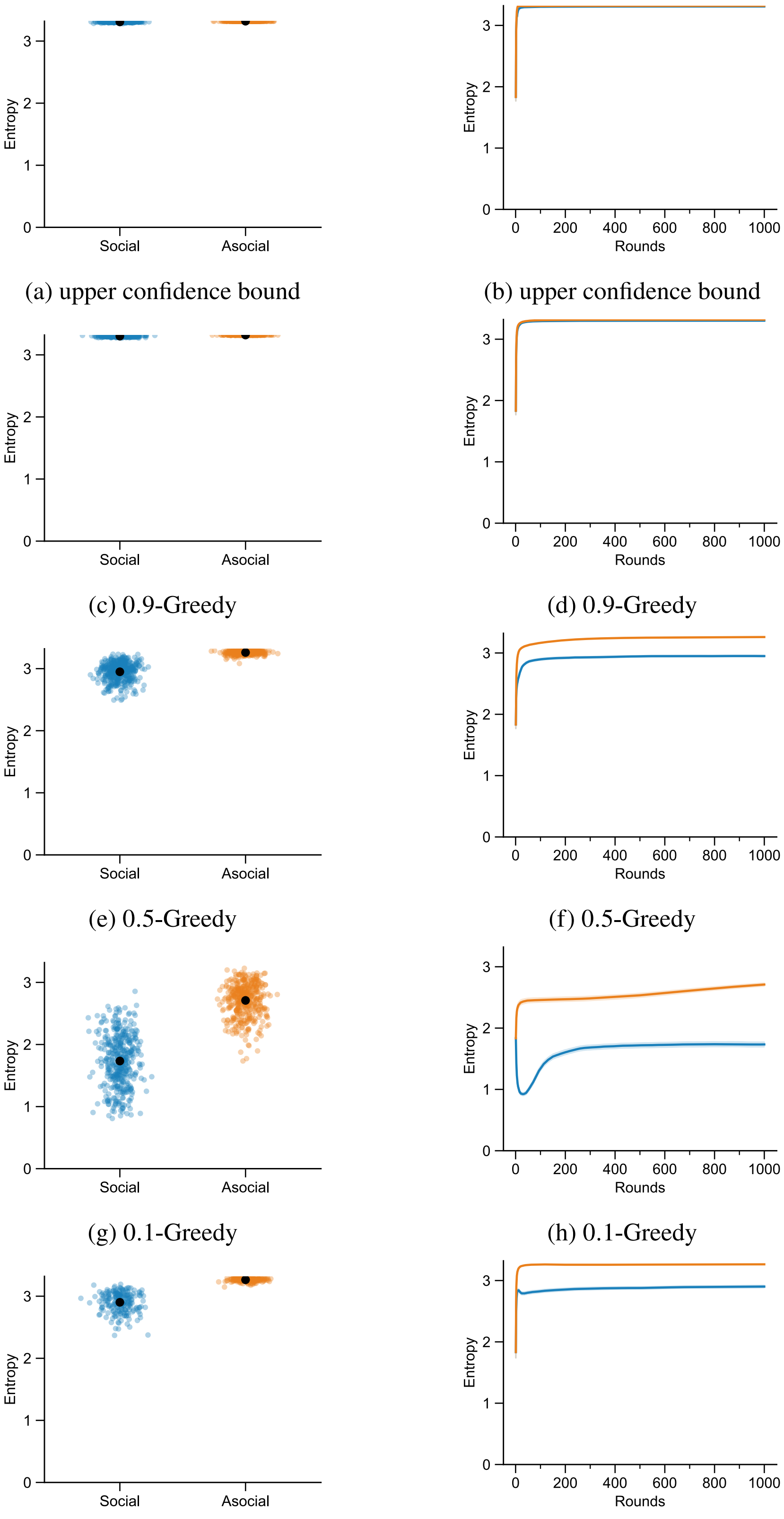
```
Initialize priors $\{\alpha_k^0, \beta_k^0\}_{k=1}^N$.
for $t = 1$ to $T$ do
    for each arm $k$ do
        $\mu_k^t = \frac{\alpha_k^t}{\alpha_k^t + \beta_k^t}$.                      ▷ Compute posterior mean
        $n_k(t) = \alpha_k^t + \beta_k^t$.                                          ▷ Compute effective pulls
        $c_k(t) = \sqrt{\frac{2 \ln t}{\max\{1, n_k(t)\}}}$.                        ▷ Compute exploration bonus
        $\mathrm{UCB}_k^t = \mu_k^t + c_k(t)$.                                      ▷ Compute index
    end for
    Select $a_t \in \arg\max_k \mathrm{UCB}_k^t$
    Proceed with observation and belief update.
end for
```

Intuitively, this strategy encourages exploration. The index of each arm consists of two components: the posterior mean and the uncertainty, which decreases as the arm is sampled more frequently. It encourages the agent to revisit the arms with limited sampling.

Setting. We compared the above three sampling strategies in a benchmark environment with $M = N = 10$ agents and arms, equal-productivity condition $\mu = 0.9$, and $T = 1000$ rounds with 100 simulation runs. We also systematically varied the initial states of the network, see below. We compared three sampling strategies: Thompson sampling with an uninformative prior $\mathrm{Beta}(\alpha, \beta) = (1, 1)$, ε-Greedy with $\varepsilon \in \{0.1, 0.5, 0.9\}$, and upper confidence bound. For each strategy, we compared final-round entropy between social and asocial learning conditions. The lower the value is, the more inequality in a given network.

Results. For each sampling strategy, we ran Ordinary Least Squares regression with learning condition as the independent variable and entropy, as defined in Section **??**, as the outcome variable. First, Thompson sampling showed a lower entropy in the social than the asocial learning condition ($b = -0.364$, 95% CI $[-0.369, -0.358]$, $p < 0.001$), resulting in a 11.13% reduction in entropy. Next, ε-greedy suggested entropy depends on ε: with $\varepsilon = 0.1$ inequality was largest ($b = -0.976$, 95% CI $[-1.023, -0.929]$, $p < 0.001$; 36.03% reduction), with $\varepsilon = 0.5$ the effect was moderate ($b = -0.311$, 95% CI $[-0.325, -0.297]$, $p < 0.001$; 9.53% reduction), and with $\varepsilon = 0.9$ showed minimal differences between the two conditions ($b = -0.0170$, 95% CI $[-0.0179, -0.0161]$, $p < 0.001$; 0.51% reduction). Lastly, upper confidence bound produced the least inequality: the entropy was very close to zero, although still statistically significantly different from zero given the large sample size ($b = -0.011$, 95% CI $[-0.012, -0.011]$, $p < 0.001$; entropy reduction by 0.34%). Figure S2 shows entropy on the left columns and entropy trajectories over time on the right columns. For highly exploratory strategies such as upper confidence bound and greedy search with $\varepsilon = 0.9$, the trajectories under social and asocial learning were nearly indistinguishable, indicating that exploration can reduce inequality. However, strategies with moderate or lower exploration such as greedy search with $\varepsilon = 0.1$ or $\varepsilon = 0.5$, or Thompson sampling, entropy first declined sharply as agents concentrated on a small set of arms before rising again, reflecting initial lock-in effects followed by partial recovery.

Implication. These results demonstrate that exploration plays a central role in reducing inequality. Conversely, they also suggest that strategies with rational updating – without strong built-in incentives to explore – can easily recreate inequality when agents are connected in a network. For methodological clarity, we use Thompson sampling for the main text, as it strikes a balance between exploration and exploitation and has been shown to resemble human exploratory behavior [2].



(a) upper confidence bound

(b) upper confidence bound

(c) 0.9-Greedy

(d) 0.9-Greedy

(e) 0.5-Greedy

(f) 0.5-Greedy

(g) 0.1-Greedy

(h) 0.1-Greedy

(i) TS

(j) TS

Figure S2: Entropy under alternative sampling strategies. Each pair corresponds to one strategy: the strip plot on the left shows final-round entropy, the trajectory plot on the right shows temporal changes.

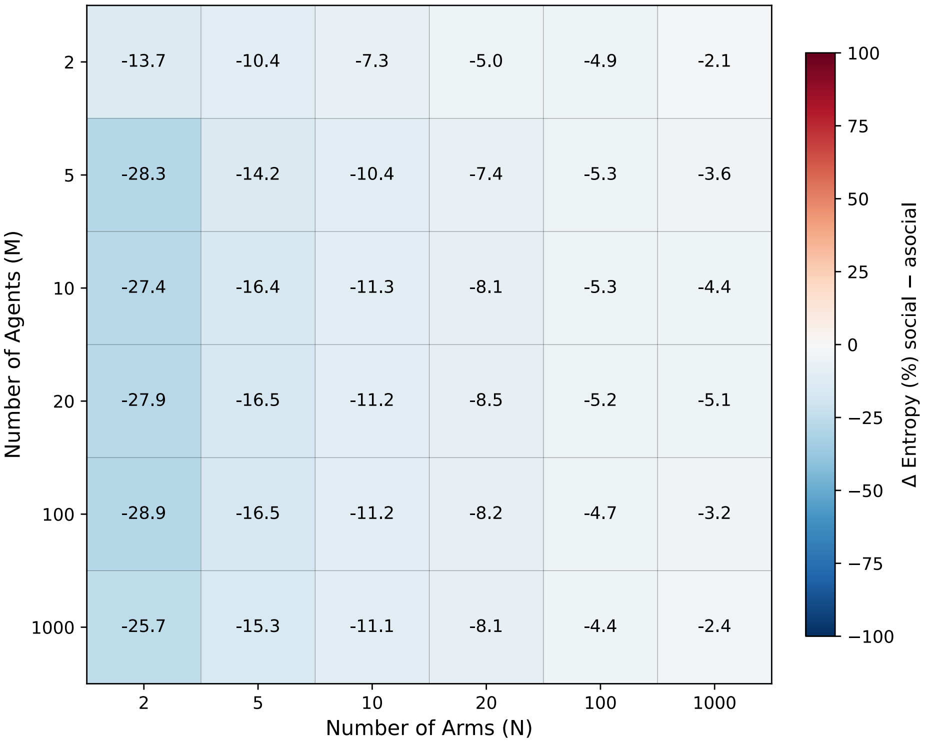
3.2 Scaling the number of agents and arms

Overview. Within the multi-agent multi-armed bandit framework, the scale of the environment is defined by two parameters: the number of agents M (decision makers) and the number of arms N (available options). We systematically varied both M and N to assess how the number of agents and options of the decision space influence the dynamics of inequality under social learning.

Setting. We varied the scale of the environment by drawing $M, N \in \{2, 5, 10, 20, 100, 1000\}$ and tested all pairwise combinations. Each configuration was run for $T = 1000$ rounds with 100 independent replicates under both social and asocial learning. Agents used Thompson sampling with an uninformative prior $\text{Beta}(1, 1)$ in the equal–productivity setting $\mu_k = \mu = 0.9$ for all k. To control for sensitivity to initial conditions, for each (M, N) we tested every distinct distribution of agents across arms (i.e., each unique unlabeled initial allocation) when the total number of such allocations was no more than 100. For very large systems where the number of possible allocations exceeded 100 (for example, when both M and N are in the hundreds or thousands), we limited the analysis to a randomly selected sample of 100 allocations drawn from the full set of possibilities. For each (M, N) pair, we compared the final-round entropy between social and asocial learning conditions.

Results. Regression analyses across the full (M, N) grid showed that social learning consistently reduced entropy relative to asocial learning ($p < 0.001$ for all comparisons), detailed in Figure S3. The magnitude of this reduction, however, varied systematically with scale. On the one hand, increasing the number of arms weakened the concentration effect of social learning: the more independent options in the network, the slower the agents converged on limited options, thus explored longer. On the other hand, increasing the number of agents initially amplified the effect, as more agents reinforced emerging preferences, but this effect plateaued once the number of agents matched or exceeded the number of arms. Beyond this critical point, additional agents did not further reduce entropy, because pooling already ensured collapse onto a narrow subset of arms. Quantitatively, the marginal effect of adding arms was stronger than the marginal effect of adding agents. If we think of the number of arms as the number of social groups in a society, this analysis indicates that the fewer groups a society has, the more unequal treatment due to social learning. We should expect to see minimal inequality when each person is treated individually without being categorized into any higher-level groups.

Implication. These findings demonstrate that inequality or collective bias is more sensitive to the number of options than to the number of agents. Adding agents amplifies convergence pressures but does not fundamentally change the trajectory once the population is sufficiently large. By contrast, increasing the number of arms consistently sustains exploration and mitigates stratification. Mechanistically, inequality under social learning arises because pooling information magnifies early stochastic successes into group-wide advantages: once an arm happens to pull ahead, all agents are drawn toward it, and the concentration persists even if the leader shifts over time. In applied domains, this implies that interventions that expand the set of available options may be more effective at preserving diversity than simply increasing the number of participants. To balance computational tractability with experimental costs, we selected $M = N = 10$ as the baseline configuration for the main study.



Number of Agents (M) \ Number of Arms (N)	2	5	10	20	100	1000
2	-13.7	-10.4	-7.3	-5.0	-4.9	-2.1
5	-28.3	-14.2	-10.4	-7.4	-5.3	-3.6
10	-27.4	-16.4	-11.3	-8.1	-5.3	-4.4
20	-27.9	-16.5	-11.2	-8.5	-5.2	-5.1
100	-28.9	-16.5	-11.2	-8.2	-4.7	-3.2
1000	-25.7	-15.3	-11.1	-8.1	-4.4	-2.4

Figure S3: Scaling analysis: with few arms, entropy under social learning dropped rapidly, producing sharp concentration. With many arms, the decline was slower and the gap between social and asocial conditions narrowed. In contrast, adding more agents steepened the early drop in entropy, but beyond moderate group sizes (roughly $M \geq N$) the curves converged, suggesting diminishing returns to increasing the number of agents.

3.3 Testing different priors

Overview. Thompson sampling (TS) behavior depends critically on the choice of prior. We therefore examined how alternative prior specifications—optimistic, pessimistic, and uninformative—shape the emergence of inequality under social learning. This analysis evaluates whether prior beliefs bias agents toward concentration or exploration, thereby altering collective outcomes.

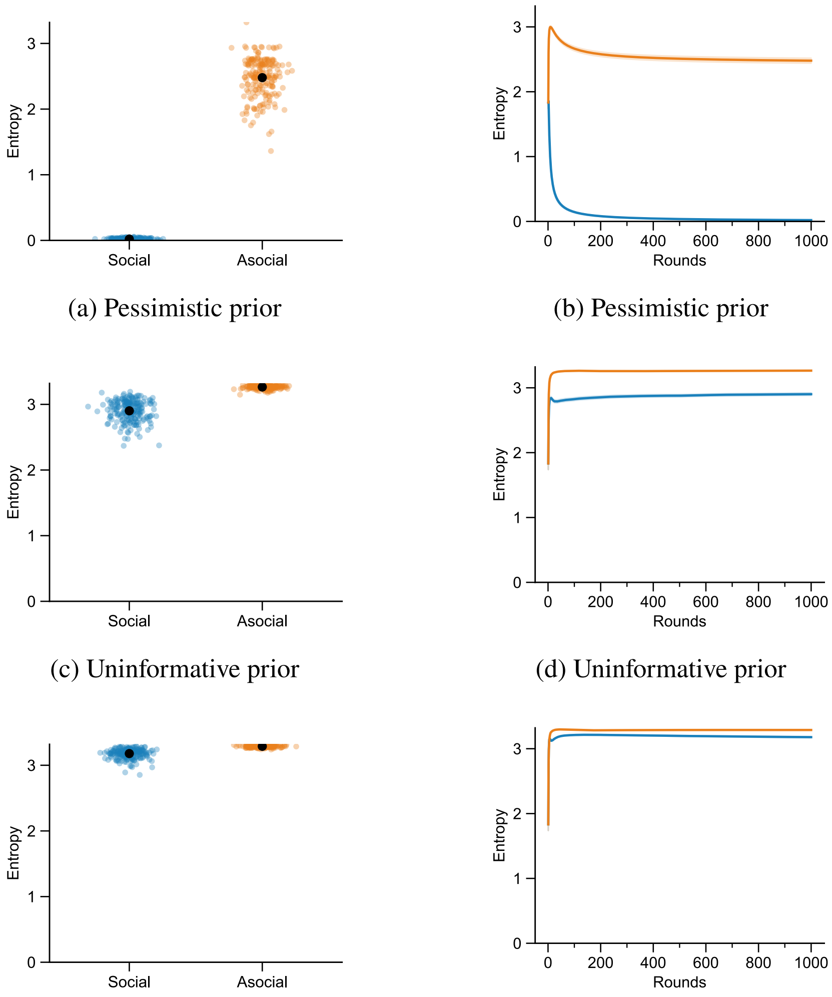
Setting. We simulated $M = N = 10$ agents and arms under the equal-reward setting with $\mu = 0.9$. Each configuration was run for $T = 1000$ rounds with 100 replicates, systematically varying the prior specification across 42 initial states. We examined three prior belief distributions:

- *Uninformative prior* $\text{Beta}(1, 1)$, corresponding to no prior counts (no bias).
- *Optimistic prior* $\text{Beta}(9, 1)$, equivalent to assuming many prior successes (optimistic belief).
- *Pessimistic prior* $\text{Beta}(1, 9)$, equivalent to assuming many prior failures (pessimistic belief).

We measured the social–asocial effect on final-round entropy, fitting separate models for each prior specification.

Results. Regression analyses revealed strong differences across prior types. With the optimistic prior Beta(9, 1), entropy reduction was 0.111 ($b = -0.111$, 95% CI $[-0.120, -0.101]$, $p < 0.001$), corresponding to a 3.36% relative decline. With the uninformative prior $\text{Beta}(1, 1)$, entropy reduction was 0.364 ($b = -0.364$, 95% CI $[-0.369, -0.358]$, $p < 0.001$), an 11.13% decline. By contrast, the pessimistic prior $\text{Beta}(1, 9)$ generated an extreme effect: entropy reduction of 2.459 ($b = -2.459$, 95% CI $[-2.499, -2.419]$, $p < 0.001$), equivalent to a 99.19% collapse. Figure S4 illustrates the trajectory of entropy over time. With uninformative or optimistic priors, asocial learning produced a rapid early increase in entropy that then stabilized, while social learning maintained consistently lower entropy. Under pessimistic priors, by contrast, entropy declined immediately and stabilized at near-zero levels, indicating rapid convergence of all agents onto a single arm. These patterns confirm that prior specification strongly shapes the degree and speed of inequality, with optimistic priors weakening inequalities, uninformative priors producing moderate disparities, and pessimistic priors amplifying convergence, leading to the most unequal outcome.

Implications. These findings underscore the sensitivity of multi-agent dynamics to initial beliefs. Two conclusions follow. First, across all priors, social learning produced more unequal outcomes than asocial learning, reinforcing our main claim that information sharing systematically drives concentration. Second, the direction of bias matters: optimistic priors can partially alleviate inequality, while pessimistic priors magnify it dramatically, collapsing the system onto a single option. In real-world terms, this suggests that overly negative preconceptions about groups can rapidly escalate into extreme segregation once information is shared, whereas positive priors may mitigate but not eliminate inequality. This is consistent with real-world observations that a breakthrough environment (better outcome than expectation) curates spiraling discrimination, whereas a breakdown environment (worse outcome than expectation) creates self-correction [3] To find a middle ground, in the main text, we adopted the uninformative prior $\text{Beta}(1, 1)$, as it represents unbiased initial beliefs and avoids seeding artificial optimism or pessimism, thereby isolating the effect of social learning.



(a) Pessimistic prior

(b) Pessimistic prior

(c) Uninformative prior

(d) Uninformative prior

(e) Optimistic prior

(f) Optimistic prior

Figure S4: Entropy outcomes under different prior specifications. Each row corresponds to one prior type (pessimistic, uninformative, optimistic). Left column: final-round entropy distributions (strip plots). Right column: entropy trajectories over 1000 rounds.

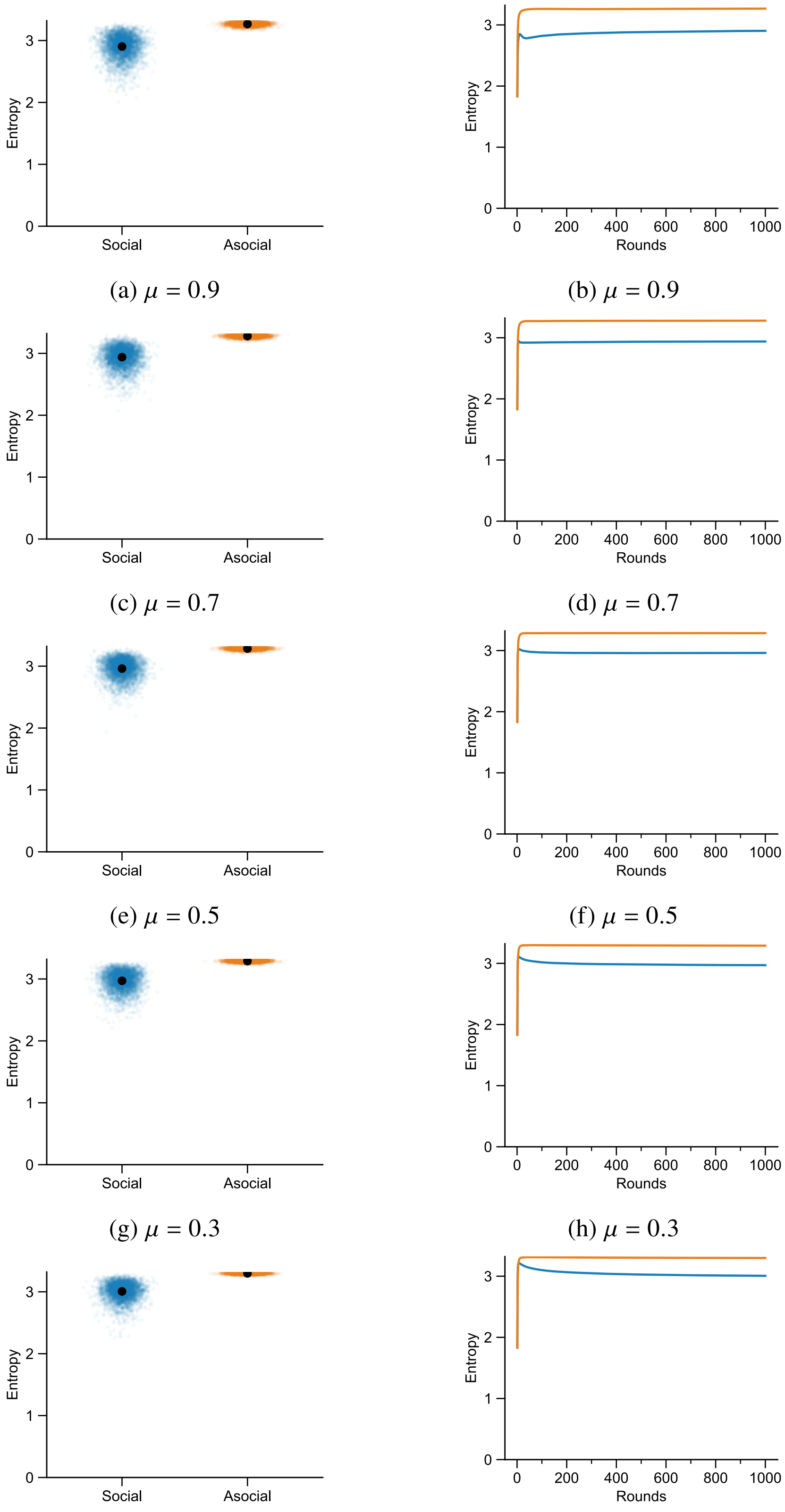
3.4 Manipulating ground truth

Overview. Beyond initial conditions, the strength of the underlying productivity signal may also shape inequality dynamics. We therefore tested whether varying the absolute success probability of all options alters the extent to which social learning drives concentration. This analysis evaluates whether stronger or weaker performance signals amplify or attenuate the emergence of inequality.

Setting. In the equal-productivity setting with $M = N = 10$, we varied the Bernoulli success probability across $\mu \in \{0.9, 0.7, 0.5, 0.3, 0.1\}$. Each configuration was run for $T = 1000$ rounds with 100 independent replicates under both social and asocial learning. All initial states were included, and agents followed Thompson sampling with an uninformative prior $\text{Beta}(1, 1)$. For each value of μ, we compared the final-round entropy between social and asocial learning conditions.

Results. Across all productivity levels, social learning consistently reduced entropy compared to asocial learning ($p < 0.001$ for all contrasts), confirming that information sharing reliably produced more concentrations. The magnitude of this reduction, however, depended on μ. At $\mu = 0.9$, entropy reduction was strongest ($b = -0.364$, 95% CI $[-0.369, -0.358]$, $p < 0.001$; 11.13% reduction). As μ decreased, the effect gradually weakened: $b = -0.341$ at $\mu = 0.7$ (10.38% reduction), $b = -0.324$ at $\mu = 0.5$ (9.85% reduction), $b = -0.319$ at $\mu = 0.3$ (9.69% reduction), and $b = -0.292$ at $\mu = 0.1$ (8.85% reduction). The decline in effect size flattened at lower μ, indicating diminishing marginal impact once productivity signals became weak. Figure S5 illustrates these patterns. At high productivity levels, asocial learning maintained relatively high entropy while social learning converged rapidly. At lower productivity levels, both conditions exhibited slower convergence, but the social–asocial gap remained robust regardless the ground truth.

Implication. These findings demonstrate that absolute productivity levels modulate but do not eliminate the inequality-generating effect of social learning. When signals are strong, social learning rapidly amplifies early stochastic successes into group-wide advantages, producing sharp concentration. When signals are weak, convergence is slower and the social–asocial gap is somewhat narrower, but inequality still reliably emerges. Mechanistically, this occurs because pooling information reduces exploration and synchronizes choices: once any arm gains even a temporary advantage, it is collectively reinforced, regardless of the baseline productivity level. This analysis implies that interventions targeting baseline productivity (e.g., boosting or lowering group quality) cannot fundamentally prevent stratification. The underlying driver is social learning itself, not the absolute quality of options. In the main text, we selected $\mu = 0.9$ as the baseline, since this parameter ensures all arms are equally high-quality and allows a more ironic effect if inequality still emerges as a consequence of social learning.



(a) $\mu = 0.9$

(b) $\mu = 0.9$

(c) $\mu = 0.7$

(d) $\mu = 0.7$

(e) $\mu = 0.5$

(f) $\mu = 0.5$

(g) $\mu = 0.3$

(h) $\mu = 0.3$

(i) $\mu = 0.1$

(j) $\mu = 0.1$

Figure S5: Effect of ground-truth means on entropy under social and asocial learning. Each pair of panels corresponds to one μ value: left shows final-round entropy (strip plot), right shows entropy trajectories over 1000 rounds.

3.5 Sampling initial states

Overview. In addition to priors and sampling strategies, outcomes in the multi-agent system can be influenced by the initial distribution of agents across options. We therefore tested whether different starting states –ranging from concentrated allocations to evenly distributed ones – affect the extent to which social learning generates inequality. This analysis evaluates the robustness of our findings to exogenous differences in initial conditions.

Setting. We fixed the environment to $M = N = 10$ agents and arms in the equal–productivity setting $\mu = 0.9$. Agents followed Thompson sampling with an uninformative prior $\text{Beta}(1, 1)$ for $T = 1000$ rounds with 100 independent replicates. We examined 42 distinct initial states, defined as the unlabeled occupancy pattern of agents across arms—ranging from fully concentrated allocations (all 10 agents on a single arm) to maximally even allocations (one agent per arm), with intermediate cases such as (4,4,2) or (7,2,1) (see Table S1). For each initial state, we compared the final-round entropy between social and asocial learning conditions.

Results. Across all 42 initial states, entropy consistently declined more under social learning than under asocial learning. Concentrated starting states produced larger early disparities and faster convergence, whereas evenly distributed states delayed the onset of concentration. Despite these differences in early dynamics, all trajectories converged to the same qualitative outcome by the end of 1000 rounds: social learning yielded more concentrated outcomes than asocial learning. The magnitude of the social–asocial gap varied somewhat across initial states, but the direction of the effect was invariant. Statistical analysis confirmed that entropy reduction showed no systematic correlation with the type of initial state entropy ($p > 0.1$). Results are summarized in Figure S6, which displays the distribution of entropy reductions across the 42 configurations.

Implication. These findings demonstrate that the emergence of inequality under social learning is robust to initial fluctuations in allocation. Even when the system began without bias (one agent per arm), social learning reliably amplified small stochastic differences into persistent disparities. To ensure comparability across experiments and to avoid seeding artificial inequality, we therefore standardized the initial condition in the main analyses to the maximally even allocation. This design choice ensures that observed disparities can be attributed to the dynamics of social learning itself rather than to exogenous asymmetries in starting conditions.

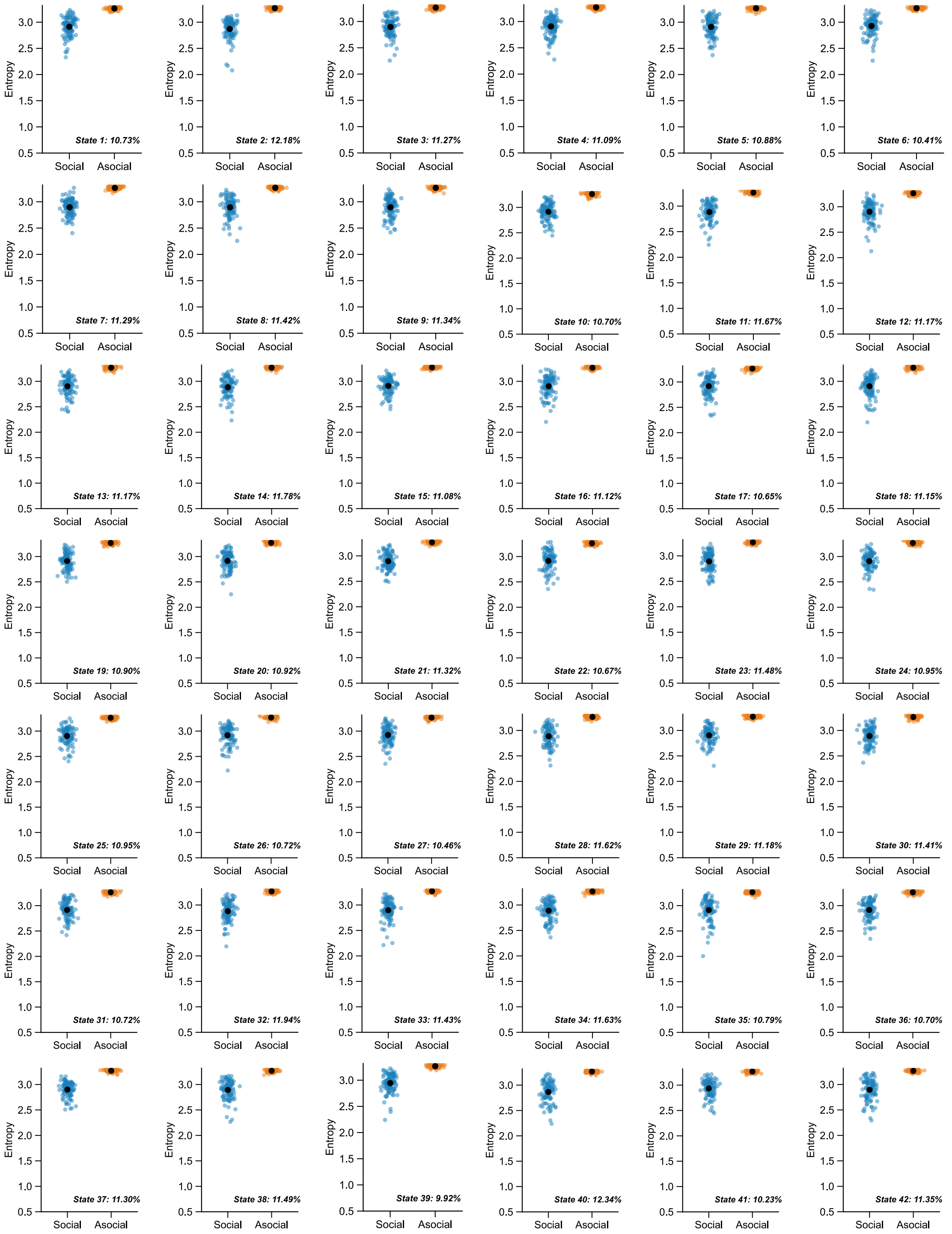


Figure S6: Final-round entropy outcomes for all 42 distinct initial states in the $M = N = 10$ equal-reward setting ($\mu = 0.9$). Each panel shows the entropy distribution for social vs asocial learning, with ΔH (%) reported in the legend. The initial states correspond to all unlabeled occupancy patterns of 10 agents across 10 arms (see Table S1).

Table S1: All 42 distinct unlabeled occupancy patterns (initial states) for $M = N = 10$. Each tuple indicates the number of agents assigned to each arm at $t = 0$, sorted in nonincreasing order. The last column reports the initial entropy $H(0)$.

State	Pattern	$H(0)$	State	Pattern	$H(0)$
1	(10,0,0,0,0,0,0,0,0,0)	0.00	22	(4,3,3,0,0,0,0,0,0,0)	1.59
2	(9,1,0,0,0,0,0,0,0,0)	0.47	23	(4,3,2,1,0,0,0,0,0,0)	1.92
3	(8,2,0,0,0,0,0,0,0,0)	0.72	24	(4,3,1,1,1,0,0,0,0,0)	2.12
4	(8,1,1,0,0,0,0,0,0,0)	0.92	25	(4,2,2,2,0,0,0,0,0,0)	2.00
5	(7,3,0,0,0,0,0,0,0,0)	0.88	26	(4,2,2,1,1,0,0,0,0,0)	2.12
6	(7,2,1,0,0,0,0,0,0,0)	1.16	27	(4,2,1,1,1,1,0,0,0,0)	2.32
7	(7,1,1,1,0,0,0,0,0,0)	1.36	28	(4,1,1,1,1,1,1,0,0,0)	2.59
8	(6,4,0,0,0,0,0,0,0,0)	0.97	29	(3,3,3,1,0,0,0,0,0,0)	1.79
9	(6,3,1,0,0,0,0,0,0,0)	1.30	30	(3,3,2,2,0,0,0,0,0,0)	1.99
10	(6,2,2,0,0,0,0,0,0,0)	1.37	31	(3,3,2,1,1,0,0,0,0,0)	2.12
11	(6,2,1,1,0,0,0,0,0,0)	1.57	32	(3,3,1,1,1,1,0,0,0,0)	2.32
12	(6,1,1,1,1,0,0,0,0,0)	1.92	33	(3,2,2,2,1,0,0,0,0,0)	2.32
13	(5,5,0,0,0,0,0,0,0,0)	1.00	34	(3,2,2,1,1,1,0,0,0,0)	2.46
14	(5,4,1,0,0,0,0,0,0,0)	1.37	35	(3,2,1,1,1,1,1,0,0,0)	2.59
15	(5,3,2,0,0,0,0,0,0,0)	1.49	36	(3,1,1,1,1,1,1,1,0,0)	2.79
16	(5,3,1,1,0,0,0,0,0,0)	1.72	37	(2,2,2,2,2,0,0,0,0,0)	2.32
17	(5,2,2,1,0,0,0,0,0,0)	1.85	38	(2,2,2,2,1,1,0,0,0,0)	2.59
18	(5,2,1,1,1,0,0,0,0,0)	2.05	39	(2,2,2,1,1,1,1,0,0,0)	2.79
19	(5,1,1,1,1,1,0,0,0,0)	2.32	40	(2,2,1,1,1,1,1,1,0,0)	2.99
20	(4,4,2,0,0,0,0,0,0,0)	1.59	41	(2,1,1,1,1,1,1,1,1,0)	3.12
21	(4,4,1,1,0,0,0,0,0,0)	1.72	42	(1,1,1,1,1,1,1,1,1,1)	3.32

3.6 Summary of bayesian learner experiments

Across five experiments, we systematically evaluated alternative sampling strategies (Section 3.1), scaling of agents and arms (Section 3.2), prior specifications (Section 3.3), baseline productivity (Section 3.4), and variation in initial states (Section 3.5). Together, these analyses suggest that the central phenomenon – social learning amplifies early fluctuations into persistent inequality – holds robustly across a wide range of modeling assumptions.

The take-away is twofold. First, the precise decision rule and parameterization affect the *magnitude* of inequality, but not its direction: social learning consistently reduces entropy relative to asocial learning. Second, robustness checks highlight the levers that modulate this effect. Exploration strategies (Section 3.1) and optimistic priors (Section 3.3) attenuate inequality, whereas pessimistic priors amplify it to the extreme. Increasing the number of arms sustains diversity more effectively than increasing the number of agents (Section 3.2). Initial fluctuations alter early trajectories but not long-run outcomes (Section 3.5). Finally, baseline productivity levels scale the strength but not the presence of the effect (Section 3.4).

For clarity and comparability, we therefore adopted a canonical setup in the main text: Thompson sampling with uninformative priors, $M = N = 10$, maximally even initial allocation, and equal-probability baseline $\mu = 0.9$. This configuration provides an unbiased and tractable benchmark. With the support of the above supplementary analyses, we hope to demonstrate the full spectrum that could be considered in future studies.

4 Generative Agents

The explicitly defined belief updating and decision making process in Bayesian learners improves interpretability, but it lacks realism. There is a burgeoning interest in using large language models (LLMs) based generative agents to simulate social dynamics, and in this section, we provide a detailed analysis of what would these agents do in our multi-agent multi-armed bandit environment. Moreover, newer models are trained to align with egalitarian values to be fair and unbiased more than Bayesian learners. These analyses also provide an interesting contrast to what would happen with fairness-minded generative agents: Will they explore in the network more than rational Bayesian learners?

Background. In a series of pilot tests, we framed the task like a hiring game but otherwise minimal instructions, and letting LLMs complete this task only through text. We did not introduce the underlying mechanism of what it means to make sequential decisions in a multi-armed bandit structure, but instead relied on the agents to infer the need to explore and exploit from context. However, early trials showed two failure patterns that cast doubt on whether LLMs fully understand this task. In particular, some agents selected options at random or in simple sequences and offered plausible but inconsistent explanations. Other agents locked into one option and ignored available information about alternatives. The parameter of all arms have equal expected productivity further obscured interpretation, since random switching can look like deliberate exploration, or simply not understanding. These observations motivated systematic prompt design to guide models toward making meaningful decisions. Below, we summarize our in-depth analysis of LLM behaviors under various prompting strategies in Section 4.1, and the finalized design of the LLM multi-agent multi-armed bandit experiment in Section 4.2.

4.1 Analyzing LLMs for bandit tasks

Challenges. With naive prompting, early trials revealed systematic challenges that cast doubt on whether the LLM agents truly understood the decision task. First, many agents produced random or sequential choices (e.g., cycling through options in order). While their textual justifications often sounded plausible – framing choices as "trying out a new candidate" or "continuing exploration" – their reasoning contained logical inconsistencies, suggesting post-hoc narratives rather than genuine strategy. Second, other agents fell into rigid exploitation, repeatedly selecting the same option even when shared outcome information clearly indicated that another option yielded higher rewards. In these cases, they effectively ignored exploration altogether. In these cases, agents appeared to neglect exploration entirely. These behavioral patterns mirror a key difficulty also present in human experiments: distinguishing between meaningful task comprehension and superficial behavior. In human studies, researchers can administer comprehension checks or debriefing surveys to directly assess whether participants have understood the instructions. For LLMs, by contrast, no such verification is possible. Their responses are limited to generated text, which may sound coherent without reflecting genuine reasoning [4]. The parameter of all arms have equal expected productivity further obscured interpretation, as it was unclear whether observed switching reflected deliberate exploration or random fluctuation, and whether exploitation represented rational preference or misunderstanding of the broader task structure.

Related work. A closely related study [5] examined LLM decision making in single agent bandit settings. The authors found most configurations failed to balance exploration and exploitation and often converged to either greedy over exploitation or random under exploration. Only one configuration achieved stable performance. That configuration combined GPT 4 at zero temperature with summarized interaction history, reinforced chain of thought reasoning, and suggestive framing. These findings highlight the importance of structured prompts, compact summaries of past outcomes, and deterministic decoding. Our study builds on these insights and extends them to multi-agent interaction and to an applied hiring context where shared information introduce additional complexities.

Prompt strategies. The observed shortcomings motivated an explicit program of prompt engineering with the objective of achieving both task comprehension and effective performance. Given the large design space of possible prompts, we adopted a procedure that provided deterministic guidance for modifications based on behavioral diagnostics. We systematically refined prompts using a consistent set of evaluation metrics to identify failure modes and then made targeted edits that addressed each failure without altering the environment. The workflow proceeded iteratively.

We began with baseline prompts that provided minimal guidance. We analyzed behavior after each test using the predefined diagnostics, see metrics below. We then introduced targeted modifications to address specific failures. Next, we tested each modification in isolation to disambiguate which modification lead to which type of failures. We repeated this cycle until agents demonstrated stable understanding and performance in the multi-agent multi-armed bandit setting.

To avoid confounds, we decomposed the overall problem into a sequence of tasks with progressively increasing complexity. Agents were required to master simpler single agent problems before advancing to multi agent settings. When learning signals were too weak, we replaced equal expected productivity with unequal productivity to first make sure our prompts could guide agents to properly explore and exploit when navigating in an environment with a ground truth. This staged design enabled incremental acquisition of the behaviors required for the full problem and ensured a systematic improvement of LLM behavior without introducing confounding changes to the task or environment.

4.1.1 Variations and metrics

Variations. We systematically tested a range of prompt and model configurations to evaluate how LLMs perform in bandit-style decision tasks. The variations included:

1. **Summarized History (Means vs. Counts):** Instead of providing raw sequences of past outcomes, we supplied either per-arm averages of rewards (means) or success/failure counts. Summarized statistics reduced arithmetic errors, directed the model to attend to relevant information.
2. **Immediate Feedback Memory:** Prompts included the agent's most recent choice and reward, together with a minimal carry-over memory. This step ensured that the LLM could anchor its reasoning on immediate feedback while retaining short-term context.
3. **Reinforced Chain-of-Thought (CoT) Reasoning:** Prompts encouraged step-by-step reasoning, guiding the LLM to weigh exploration–exploitation trade-offs more explicitly.
4. **Suggestive Framing:** Prompts were framed in ways that subtly emphasized the importance of balancing exploration and exploitation, giving a slice of heuristics to assist model decisions.
5. **Model Temperature:** We compared deterministic outputs (temperature $T = 0.0$) with more stochastic ones ($T = 0.7$). Zero temperature reduced randomness, yielding more consistent behaviors, while higher temperature sometimes destabilized behavior.
6. **Model Scale:** We compared GPT-4o and GPT-4o-mini to assess sensitivity to model size.

Metrics. Following [5], we employed surrogate statistics to diagnose whether LLMs successfully balance exploration and exploitation over repeated trials:

- **Suffix Failure Frequency (SuffFailFreq):** The fraction of replicates where the best arm is never chosen after a certain round t. Persistent suffix failures indicate long-term failure to explore.
- **MinFrac:** The minimum fraction of pulls received by any arm across rounds. When scaled by K (the number of arms), $K \cdot$ MinFrac close to 1 indicates uniform-like failure (treating all arms equally, without eliminating suboptimal options).
- **GreedyFrac:** The proportion of rounds in which the model selects the arm with the highest empirical mean reward, reflecting over-exploitation tendencies.
- **Median Reward:** The median of time-averaged rewards across replicates, expected to stabilize near 0.5 if exploration and exploitation are balanced.

These metrics provide more diagnostic power than raw cumulative rewards, which are often too noisy at moderate horizons. They allow us to distinguish between specific failure modes such as suffix failures and uniform-like failures.

Failures and modifications. Using these metrics, we identified several recurring failure modes (see evaluation results in Section 4.1.2 and corresponding modifications in Section 4.2). First, nonzero Suffix Failure Frequency indicated inadequate summarization of past outcomes, which we addressed by refining how history was presented. Second, MinFrac values of zero revealed that some arms were entirely ignored, for which we refined prompts to encourage exploration of underutilized options. Third, extreme GreedyFrac values suggested imbalances between exploration and exploitation, and we clarified the framing of this taks. Finally, unstable Median Rewards reflected overly random behavior, which we mitigated by lowering temperature and by providing additional contextual scaffolding.

4.1.2 Experiments and results

To evaluate LLM performance systematically, we adopted a phased approach that gradually increased the complexity of the task. We began with a simple single-agent bandit setting (Phase 1), then introduced domain framing in the hiring context (Phase 2). Next, we scaled to a multi-agent binary bandit (Phase 3), and finally to a multi-agent hiring scenario with feature-based candidates (Phase 4). This progression from simple to complex ensured that LLMs mastered basic explore–exploit trade-offs before advancing to collective decision making under social learning. Each phase provided insights into failure modes and guided prompt refinements, yielding the finalized design for the main experiment.

Phase 1: Single-Agent Binary Bandit (Asocial learning). A single LLM agent interacted with a four-arm bandit over 100 rounds (10 turns). We tested GPT-4o and GPT-4o-mini at two temperature settings (0.0 and 0.7). To probe the role of prompt design, we compared seven variants: (1) summarized history with per-arm means, reinforced CoT, suggestive framing, and memory-in-context; (2) same as (1) but without per-arm means; (3–4) versions with reduced framing; (5–6) versions without CoT or history; and (7) without memory. Only GPT-4o with temperature 0.0 under prompts (1) and (2) consistently balanced exploration and exploitation, reflected in stable Median Rewards and nonzero MinFrac. All other conditions exhibited failures: random or repetitive choices (Failure Type #2), unstable Median Rewards (Failure Type #4), or inability to converge (Failure Type #1). We adopted prompt (1) with temperature 0.0 as the baseline, with summarized history, CoT, and suggestive framing.

Phase 2: Single-Agent Hiring Bandit (Asocial learning). We adapted the bandit into a hiring task: the LLM acted as a hiring manager choosing among four anonymized candidates. Rewards followed Bernoulli draws with fixed but hidden probabilities. We retained GPT-4o ($T = 0$) and tested two prompt versions: (1) directly adapted from Phase 1, using a hiring cover story; (2) same but without explicit mention of the probabilistic reward distribution. Version (1) successfully balanced exploration and exploitation: agents explored candidates and exploited successful ones. Version (2) often failed to exploit prior successes, producing higher Suffix Failure and unstable Median Rewards. Accordingly, we decided to use Version (1), given that explicit probabilistic framing seems to be essential for LLM comprehension of explore–exploit trade-offs in the hiring contexts.

Phase 3: Multi-Agent Binary Bandit (Social learning). Four LLM agents operated in a binary bandit environment with four arms for 40 rounds (10 turns), under social learning. Each agent saw pooled outcomes across agents. Prompts built on Phase 1 but were adapted to include shared history. LLM agents demonstrated partial coordination: GreedyFrac and Median Rewards stabilized under prompt (1), but failures emerged when history was poorly summarized, leading to ignored arms (MinFrac $\approx$ 0, Failure Type #2). We refined prompts to highlight both individual and collective outcomes, ensuring agents incorporated shared evidence. This change improved coordination among agents, and stabilized exploration–exploitation dynamics in the multi-agent setting.

Phase 4: Multi-Agent Hiring Bandit (Social learning). Extending Phase 3, four LLM agents acted as hiring managers choosing among four candidates with two categorical attributes. Candidate success probabilities varied by interaction with firm profiles. The experiment ran for 40 rounds (10 turns) under GPT-4o ($T = 0$). Prompts summarized outcomes by candidate feature and firm. Under social learning, agents updated strategies from both personal and shared outcomes. However, some agents ignored unexplored candidates or miscalculated probabilities, reflected in low MinFrac and unstable Median Rewards. We revised prompts to explicitly include per-firm probability estimates and to mark unobserved options as “unknown” (prior 0.5). This change encouraged systematic exploration of under-sampled options and improved convergence toward optimal hiring strategies.

Table S2: LLM prior experiment results.

Phase	Model	Temp.	Hist.[1]	Mem.[2]	CoT[3]	Frame[4]	SFF[5]	MinFrac[6]	GreedyFrac[7]	Median R.[8]	Success
1	GPT-4o	0.0	Means	✗	✓	✓	0.0	0.13	0.91	0.54	✓
	GPT-4o-mini	0.0	Means	✗	✓	✓	0.2	0.13	0.94	0.94	✗
	GPT-4o	0.0	Counts	✗	✓	✓	0.0	0.15	0.76	0.52	✓
	GPT-4o-mini	0.7	Means	✗	✓	✓	0.0	0.01	0.90	0.81	✗
	GPT-4o	0.7	Means	✗	✓	✓	0.0	0.24	0.84	0.67	✗
	GPT-4o-mini	0.0	Counts	✓	✓	✓	0.0	0.13	0.60	0.54	✗
	GPT-4o	0.0	Counts	✓	✓	✓	0.0	0.15	0.93	0.49	✓
	GPT-4o-mini	0.0	Counts	✓	✓	✗	0.0	0.14	0.81	0.64	✗
	GPT-4o	0.0	Counts	✓	✓	✗	0.0	0.17	0.91	1.0	✗
	GPT-4o-mini	0.0	✗	✓	✓	✗	0.0	0.51	0.48	0.39	✗
	GPT-4o	0.0	✗	✓	✓	✓	0.0	0.34	0.63	0.06	✗
	GPT-4o-mini	0.0	Counts	✗	✗	✓	0.5	0.0	0.76	0.54	✗
	GPT-4o	0.0	Counts	✗	✗	✓	0.0	0.09	0.91	0.67	✗
2	GPT-4o	0.0	Means	✓	✓	✓	0.0	0.14	0.92	0.7	✓
	GPT-4o	0.0	Means	✓	✓	✗	0.1	0.08	0.89	0.87	✗
3	GPT-4o	0.0	Means	✓	✓	✓	0.0	0.14	0.92	0.56	✓
4	GPT-4o	0.0	Means	✓	✓	✓	0.0	0.16	0.91	0.71	✓

[1] Hist. = History input. Either per–arm averages of rewards (Means) or raw success/failure counts (Counts).
[2] Mem. = Last–round reward and short memory of prior outcomes.
[3] CoT = Reinforced chain-of-thought reasoning.
[4] Frame = Suggestive framing emphasizing exploration–exploitation trade-offs.
[5] SFF = Suffix Failure Frequency: proportion of trials where the optimal arm is not chosen in later rounds.
[6] MinFrac = Minimum fraction of times any arm is selected; captures whether all options are explored.
[7] GreedyFrac = Fraction of choices of the immediate best arm; reflects exploitation tendency.
[8] Median R. = Median of time-averaged rewards across trials.
✓ = feature present, ✗ = feature absent. Metrics follow [5].

4.1.3 Summary

All empirical results are summarized in Table S2. Iterating through the four phases with modified prompt strategies, we converged on a successful LLM configuration, for which we use in the main experiment: GPT-4o with temperature fixed at 0.0, summarized history presented as per-arm means (without raw sequences or last-round memory), reinforced chain-of-thought reasoning, and suggestive framing. This combination consistently balanced exploration and exploitation, as reflected in stable median rewards, nonzero MinFrac values, and low suffix failure frequencies. Together, these results validate that prompt scaffolds are not optional but essential: only with carefully designed guidance can LLMs reliably reproduce the dynamics of exploration, exploitation, and information sharing that underpin inequality in multi-agent bandit settings. Next, we report the full details of prompt design choices and implementations in Section 4.2.

4.2 Simulating social learning in LLM bandit

The finalized experimental setup for our main study combined targeted prompt engineering with a moderator–agent interaction loop conducted entirely in natural language. In this session, we provide relevant details for a quick preview and will make the repo publicly available for future research.

There are three layers of this design: First, a centralized moderator summarizing the history of observed outcomes (conditioned on whether agents were in the social or asocial learning condition), issuing queries to the agents, collecting their text-based responses, and recording the resulting choices and rewards Section 4.2.1. Second, natural language instructions via prompts to guide model to make decisions. This step is inspired by our in-depth analysis above Section 4.1, including: summarized history, reinforced chain-of-thought reasoning, suggestive framing, a memory anchor for the most recent choice and outcome, and a fixed temperature of zero Section 4.2.2. Third, implementation details such as number of agents and number of arms, as well as the two experimental conditions: different versus identical expected productivity of arms, and social versus asocial learning Section 4.2.3. A schematic illustration of this experiment is shown in Figure S7, and example prompts in Section 4.2.4.

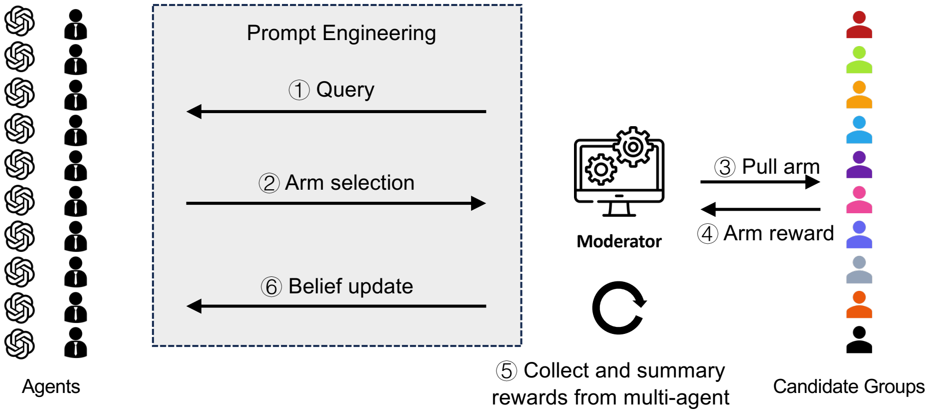


Figure S7: Schematic of the LLM-agent experiment loop. A moderator system summarizes past outcomes, queries each LLM agent for its next hiring decision, and records text-based responses under either social or asocial conditions.

4.2.1 Moderator and agent loop

Each experiment unfolded as a repeated interaction between a moderator and a set of agents. At the start of each run, M arms and N agents were initialized with uninformative $\text{Beta}(1, 1)$ priors. At each round t, the moderator provided a history of past outcomes: in the asocial learning condition, each agent observed only its own prior choices and rewards, while in the social learning condition, all agents also observed aggregated statistics from the group, including expected values per arm, number of pulls, and success rates. Each agent then received a natural-language prompt containing several elements: the last-round choice and outcome (memory scaffold), a summarized history table using per-arm means rather than raw sequences, framing language emphasizing the exploration–exploitation dilemma, an instruction to "think step by step" to encourage explicit reasoning, and an explicit output format `<Answer> CandidateX </Answer>` where $X \in \{1, \ldots, M\}$. Agents replied with free-text explanations followed by a tagged choice, which the moderator parsed from the `<Answer>` tag. Each chosen arm yielded a Bernoulli reward according to its latent success probability, and the process was repeated for T rounds.

4.2.2 Prompting strategies

The finalized prompt design included five key features. First, summarized history in the form of per-arm averages prevented arithmetic drift and kept inputs concise. Second, a memory scaffold specifying the last-round choice and reward anchored reasoning to immediate feedback. Third, prompts instructed the model to engage in chain-of-thought reasoning before committing to an answer. Fourth, natural-language cues highlighted the exploration–exploitation trade-off without prescribing a solution, providing suggestive framing. Finally, a fixed decoding temperature of $T = 0.0$ enforced deterministic outputs, eliminating spurious variability across runs.

4.2.3 Implementation details

To ensure comparability with the simulation experiments, we adopted the same baseline configuration: $M = N = 10$ (ten agents and ten arms) with equal ($\mu = 0.9$) and differing ($\mu = 0.9$ to 0.1) latent reward probabilities. Accordingly, all experiments used the GPT-4o model at zero temperature. Each run lasted $T = 200$ rounds and was repeated ten times. This design reflected two practical considerations. Deterministic decoding removed most sources of stochasticity, reducing the need for a large number of replicates. Moreover, prior simulation results indicated rapid convergence to a dominant arm, such that 200 rounds were sufficient to capture the emergence of inequality dynamics while controlling computational costs. To avoid unintended biases from semantic priors, candidate groups were represented symbolically as *Candidate1 – Candidate10*.

4.2.4 Prompt example

An example prompt used in the social learning condition is reproduced below.

```
/* System Prompts: Explain the game setup and rules clearly so the LLM
understands its objectives and can make rational decisions. */
{
   "role" : "system",
   "content" : "You are one of several hiring managers tasked with
   maximizing your firm’s rewards by hiring one of 10 candidates groups
   labeled Candidate1-Candidate10. Each candidate group is associated with
   a Bernoulli distribution that gives a reward feedback of either failure(0)
   or success(1), with unknown probabilities. Your goal is to maximize your
   firm’s total reward over a series of rounds by deciding which candidate
   group to hire next, based on the feedback provided after each round.
   At each time step, you will receive a summary of the results from both
   your own choices and the choices made by other firms. You must decide
   which candidate group (Candidate1-Candidate10) to pull, based on your
   understanding of the rewards so far. Let’s think step by step to make
   sure we make the best decision. After deciding, provide your final answer
   within the tags <Answer>CANDIDATE</Answer> where CANDIDATE is one of
   Candidate1-Candidate10."
},

/* Assistant Prompts: Choice made by the agent in last round. */
{  "role": "assistant", "content": "Candidate1."},

/* User Prompts: Provide information updates (a summary of all previous rounds).
Ask the LLM to make a choice for the next round. */
{  "role": "user",
   "content": "So far you have hired 22 times. Your previous choices and
   rewards, along with shared information from other firms, are summarized
   as follows:
              Candidate1: 69 times, average reward: 0.91
              Candidate2: 23 times, average reward: 0.86
              ...
              Candidate10: 16 times, average reward: 0.73
   Which candidate group will you choose next? Remember, You MUST provide your
   final answer within the tags <Answer>CANDIDATE</Answer>."}
```

Highlight legend: green = Suggestive framing cyan = Summarized history with means yellow = Chain-of-thought pink = Social learning hint orange = Symbolized representation

5 Human Participants

To study the empirical relevance of our hypothesis, that social learning creates more collective bias than asocial learning, we designed a multi-player online hiring experiment. We recruited groups of 10 human participants, role-playing as members of a hiring committee. They evaluated many job candidates and made hiring decisions based on their own (asocial) and everyone else's (social) hiring experiences. To alleviate existing bias, this study used artificial labels of job candidates, that is color of the icon to represent their group identity. The main dependent variables are the same as Bayesian learners and language agents, efficiency and inequality as defined in Section 2, and the main independent variables are the market structure (equal versus unequal productivity) and the structure of the agents (asocial versus social). To supplement the main results, in this section, we cover detailed design choices in Section 5.1, results from a pilot study in Section 5.2, descriptive analyses of the main study in Section 5.3, and additional and qualitative analyses of the main study in Section 5.4.

5.1 Study design and implementation

5.1.1 Pre-registration and approval

A pre-registration for the study is available online at the Open Science Framework (OSF) at https://osf.io/58nbt/. The research protocol was approved by the Institutional Review Board (Protocol No. 24-1184). Experimental framework (via Empirica [1]), data analysis scripts, and anonymized data are publicly available in the same repository.

5.1.2 Participants and recruitment

We recruited $N = 2000$ participants from Amazon Mechanical Turk via the CloudResearch platform. Eligibility required participants to be born in the United States, at least 18 years old, and proficient in English. Recruitment targeted a balanced gender distribution, and participants were screened using CloudResearch's quality filters (approval rating $\geq 90\%$). Repeat participation was disallowed. Data collection occurred across multiple recruitment sessions between April 8, 2025, and June 2, 2025. Demographic information (age and gender) was recorded and analyzed as part of the robustness checks. The average completion time was 20 minutes. Participants were compensated by a \$5 base payment plus a performance-contingent bonus of up to \$1, determined by their cumulative rewards in the game.

5.1.3 Experiment structure

The human experiment was implemented on Empirica, an open-source JavaScript framework for running multiplayer interactive experiments and games directly in the browser [1]. Empirica is designed for real-time, multi-player behavioral studies, allowing researchers to create complex interactive designs without custom server engineering while maintaining statistical rigor. This experiment follows a well-defined life cycle. After reading an introductory screen, participants enter a virtual lobby, where they wait until the entire group has completed the introduction. Each study is organized as a game composed of multiple rounds, and each round contains one or more stages. The game proceeds in a tightly synchronized sequence: all players must finish the current stage—for example, making a decision or reviewing shared outcomes—before the platform advances the whole group. When every participant completes the final stage of a round, the system moves to the next round; after the last round, it proceeds to the exit steps, where players complete tasks such as a brief survey at their own pace. This lock-step progression ensures that all participants remain perfectly coordinated and experience the experiment in real time together. By handling group matching, real-time updates, and automatic data logging, this platform provides the technological foundation for our multiplayer experiment.

5.1.4 Experiment procedure

Within the Empirica framework we built a custom interactive game, *Together Hire*, to parallel the multi-agent simulations described in earlier sections. In each session, a group of ten participants acted as members of a hiring committee making 50 sequential hiring decisions among 10 color-coded candidate groups, a design intended to avoid pre-existing social biases. Participants were randomly assigned, upon arrival and consent, into groups of 10 and placed in one of four experimental conditions: unequal productivity with asocial learning, unequal productivity with social learning, equal productivity with asocial learning, and productivity reward with social learning.

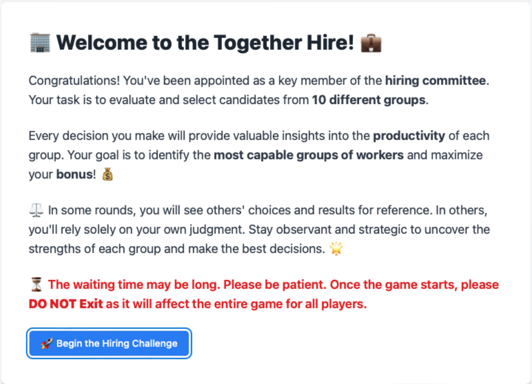
The user interface guided participants through a fixed sequence of stages: Introduction, Game Introduction, Tutorial, Main Choices, Group Allocation, and Exit Survey. It provided clear instructions while presenting a single hiring scenario that flexibly adapted to the four experimental conditions described above. The following subsection details the content and function of each stage.

Introduction and waiting room. Sessions began once all 10 participants had entered the "game room" and confirmed readiness. Participants were presented with written instructions explaining the hiring task, the binary success/failure outcomes, and how their bonus is calculated and rewarded. First, the Introduction screen (Figure S8a) welcomed participants to the "Hiring Boardroom" and provided an overview of the task. It emphasized their role as hiring managers, the presence of ten candidate groups, and the link between performance and monetary bonus. Next, the Game Introduction screen (Figure S8b) described the main mechanics of the task. Participants were told that each round they would make a hiring decision, receive binary success/failure feedback, and accumulate bonuses of one cent per success. The interface also specified whether the condition involved social learning (shared group-level outcomes) or asocial learning (private outcomes only).

Tutorial. Before the main hiring game began, participants completed a single practice round to become familiar with the interface. The Tutorial screen (Figure S8d) displayed ten brightly colored boxes, each representing a candidate group; below each box two counters showed the cumulative numbers of successes and failures. An animated arrow on the screen, accompanied by brief on-screen instructions (Figure S8d), guided participants to click on a pre-selected box—thereby making a hiring decision. This brief exercise demonstrated the mechanics of the hiring task and ensured that all participants entered the main game from an identical initial state, matching the starting conditions used in the Bayesian and generative-AI simulations.

Main game. Participants then played 50 consecutive rounds of the hiring task. Similar to the Tutorial screen, the Choices screen (Figure S8e) displayed ten colored boxes representing the candidate groups, with two counters beneath each box showing the cumulative numbers of successes and failures. In the social learning condition these counters reflected the pooled outcomes of all participants, whereas in the asocial learning condition they reflected only the individual's own outcomes. At every round participants selected one candidate group to hire, basing their decisions on the feedback and beliefs formed from previous outcomes. A countdown timer at the top of the screen allowed up to 30 seconds for each decision and displayed the participant's cumulative bonus; a warning banner appeared when only ten seconds remained. After all participants submitted their choices, the system automatically advanced the entire group to the next round. And the screen presented the updated hiring summary through the counters, providing the information participants relied on to guide their next decision.

Additional tasks and exit survey. After the main game, participants completed an additional group-allocation task to obtain an explicit measure of their beliefs about each candidate group's productivity beyond what could be inferred from their hiring decisions. On the Group Allocation screen (Figure S8c), participants were asked to allocate 100 hypothetical slots across the ten candidate groups according to their perceived relative productivity. The session concluded with an exit survey, which collected demographic information and invited participants to describe the strategies they used during the experiment. The detailed contents of this exit survey are provided below.



🏢 **Welcome to the Together Hire!** 💼

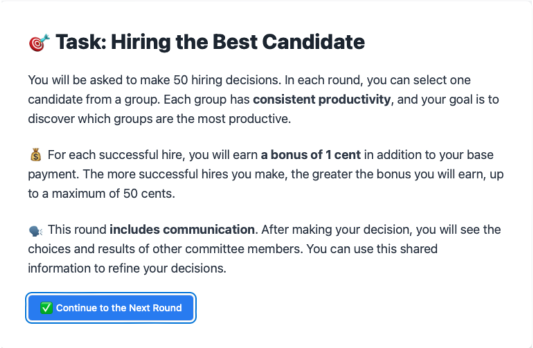
Congratulations! You've been appointed as a key member of the **hiring committee**. Your task is to evaluate and select candidates from **10 different groups**.

Every decision you make will provide valuable insights into the **productivity** of each group. Your goal is to identify the **most capable groups of workers** and maximize your **bonus**! 💰

⚖️ In some rounds, you will see others' choices and results for reference. In others, you'll rely solely on your own judgment. Stay observant and strategic to uncover the strengths of each group and make the best decisions. 🌟

⏳ **The waiting time may be long. Please be patient. Once the game starts, please DO NOT Exit as it will affect the entire game for all players.**

(a) Introduction screen.



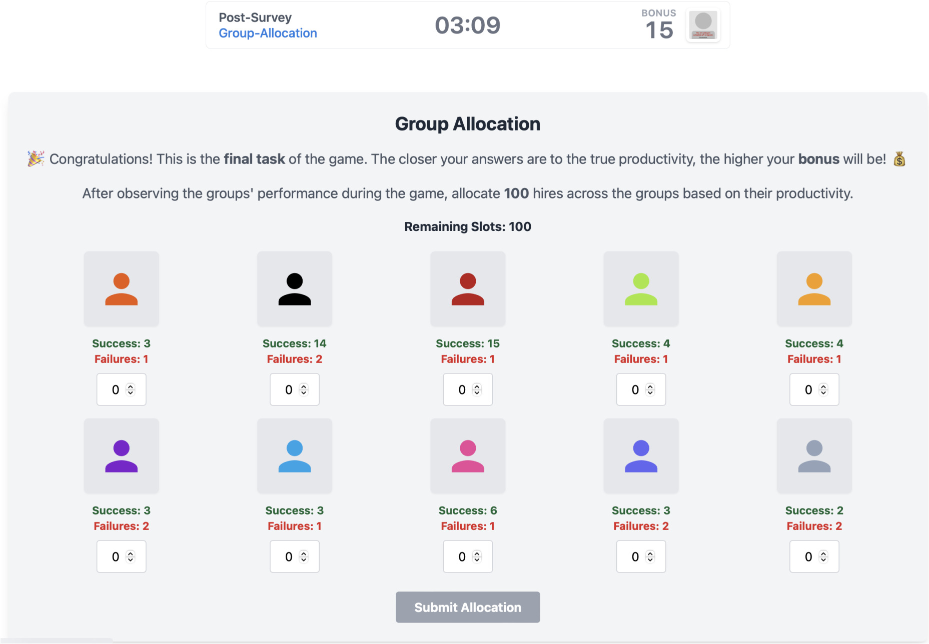
🎯 **Task: Hiring the Best Candidate**

You will be asked to make 50 hiring decisions. In each round, you can select one candidate from a group. Each group has **consistent productivity**, and your goal is to discover which groups are the most productive.

💰 For each successful hire, you will earn **a bonus of 1 cent** in addition to your base payment. The more successful hires you make, the greater the bonus you will earn, up to a maximum of 50 cents.

🗣️ This round **includes communication**. After making your decision, you will see the choices and results of other committee members. You can use this shared information to refine your decisions.

(b) Game introduction screen.



Group Allocation

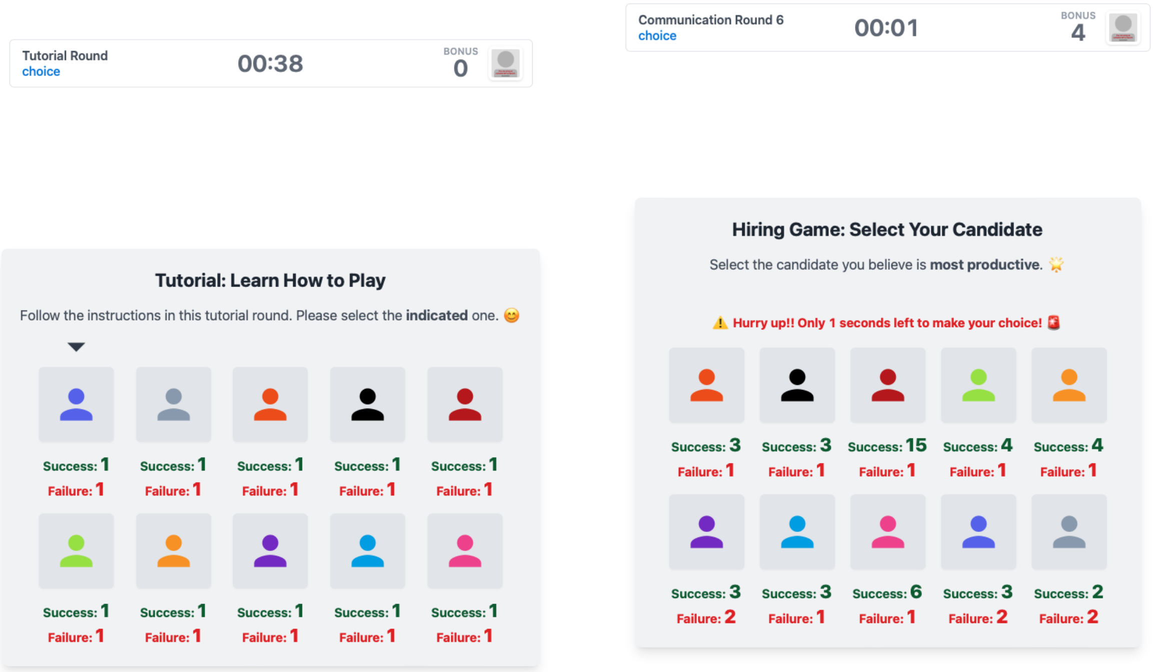
🎉 Congratulations! This is the **final task** of the game. The closer your answers are to the true productivity, the higher your **bonus** will be! 💰

After observing the groups' performance during the game, allocate **100** hires across the groups based on their productivity.

Remaining Slots: 100

Success: 3 Failures: 1 0	Success: 14 Failures: 2 0	Success: 15 Failures: 1 0	Success: 4 Failures: 1 0	Success: 4 Failures: 1 0
Success: 3 Failures: 2 0	Success: 3 Failures: 1 0	Success: 6 Failures: 1 0	Success: 3 Failures: 2 0	Success: 2 Failures: 2 0

(c) Group Allocation screen.



Tutorial Round
choice
00:38
BONUS
0

Tutorial: Learn How to Play

Follow the instructions in this tutorial round. Please select the **indicated** one. 😊

▼

Success: 1 Failure: 1	Success: 1 Failure: 1	Success: 1 Failure: 1	Success: 1 Failure: 1	Success: 1 Failure: 1
Success: 1 Failure: 1	Success: 1 Failure: 1	Success: 1 Failure: 1	Success: 1 Failure: 1	Success: 1 Failure: 1

(d) Tutorial screen.

Communication Round 6
choice
00:01
BONUS
4

Hiring Game: Select Your Candidate

Select the candidate you believe is **most productive**. 🌟

⚠️ **Hurry up!! Only 1 seconds left to make your choice!** 🚨

Success: 3 Failure: 1	Success: 3 Failure: 1	Success: 15 Failure: 1	Success: 4 Failure: 1	Success: 4 Failure: 1
Success: 3 Failure: 2	Success: 3 Failure: 1	Success: 6 Failure: 1	Success: 3 Failure: 2	Success: 2 Failure: 2

(e) Choices screen.

Figure S8: Screenshots of the experimental interface.

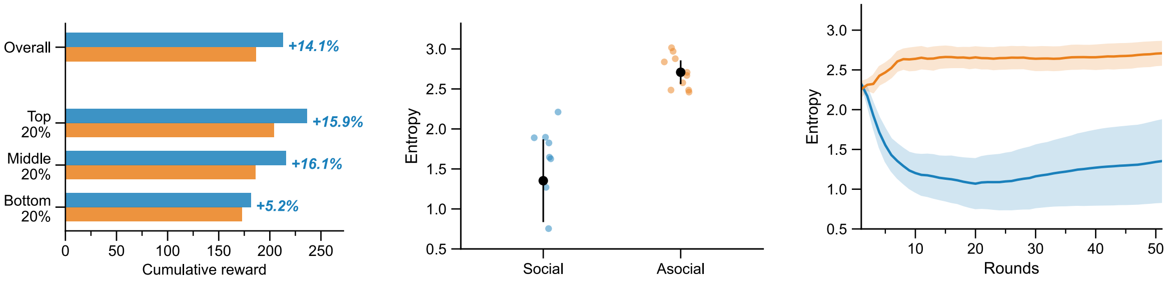
Exit survey. Participants completed the following exit survey at the conclusion of the experiment. Responses were required to receive full compensation.

```
Please submit the following code to receive your bonus: [Player ID].
Your final bonus is in addition to the $1 base reward for completing the HIT.

(Page 1 of 2)
- Age: 1-100
- Gender: Female, Male, Other (Please specify)
- Race: African, Asian, Caucasian, Latin/x, Native American, Mixed-Race,
   Other (Please specify)
- Education: Did not graduate from high school, High School, Some College,
             College, Graduate Professional School, Other (Please specify)
- Primary Country/Region of Residence
- Political Orientation (1 = Extremely Conservative, 6 = Extremely Liberal)

(Page 2 of 2)
- Have you participated in similar experiments before? Yes/No
    If yes: Please share details such as when, where, who organized it (e.g.,
    university, company, etc.), and a brief description of the experiment.
- Would you like to participate in follow-up studies? Yes/No
- How engaging was the game? (1 = Not engaging, 5 = Very engaging)
- How clear were the instructions? (1 = Very unclear, 5 = Very clear)
- How did you approach decision-making in this experiment?
    (e.g., based purely on success, based purely on failure,
    weighting success more, weighting failure more,
    relying on observations, analyzing trends, intuition, trial and error)
- What key factors influenced your decisions?
    (e.g., consistency, personal experience, group experience)
- What motivated you to participate in this experiment?
    (e.g., monetary incentive, interest in research, curiosity)
- Did you encounter any technical issues or distractions? Yes/No
    If yes: Please describe when, what happened, and any error messages.
```

5.2 Pilot study

Before launching the main experiment, we conducted a pilot study to validate the design with a smaller group size. The game was identical to the main setup, except that each group consisted of five participants instead of ten. We recruited $N = 200$ participants in total, organized into 10 groups per experimental condition (equal vs. unequal productivity × social vs. asocial learning). Data collection took place between February 12 and February 16, 2025. Participants were recruited from Amazon Mechanical Turk via CloudResearch. Eligibility required being born in the United States, at least 18 years of age, and proficient in English. Recruitment targeted a balanced gender distribution, and participants were screened using CloudResearch's quality filters (approval rating $\geq$ 90%). Repeat participation was disallowed.

Overall, results closely mirrored those observed in the main experiment, detailed in Figure S9. When worker quality varied, social learning improved hiring efficiency: average rewards increased by 14.08% (95% CI = [6.79, 21.11], $p = 0.002$). When all groups were equally productive, however, participants still converged disproportionately on a single group of workers ($b = -1.356$, 95% CI = [–1.856, –0.856], $p < 0.001$). Entropy declined monotonically throughout the game, ending 2.12× lower in the social learning condition than in the asocial learning condition, corresponding to a 50.06% reduction in entropy. We presented this pilot results at the 11th International Conference of Computational Social Science in Norrkoing, Sweden.



(a) Cumulative reward.

(b) Final-round entropy.

(c) Entropy trend.

Figure S9: Pilot study results.

5.3 Preliminary analyses

With the scaled-up sample in the main experiment, we first examined whether participant demographics systematically influenced our intended treatment outcomes by conducting a balance test.

Descriptive statistics Table S3 reports demographic characteristics across all participants. The sample was broadly representative of U.S. online panels: age ranged from 18 to 100 (mean = 37.8). Gender was nearly balanced (54% female, 48% male, < 1% other). The majority identified as Caucasian (71%), with smaller proportions identifying as African (9.7%), Latinx (6.3%), Asian (6.2%), and Mixed (4.2%). Education levels were heterogeneous, with 46% reporting college graduation, 25% some college, 17% graduate/professional school, and 10% high school. Political orientation leaned slightly liberal, with 25% moderately liberal, 22% moderately conservative, and the remainder distributed across other categories. These distributions indicate that the recruitment strategy acheived a balanced and diverse participant pool of the US population, comparable other online studies.

Balance tests. We next tested whether demographic variables systematically predicted decision outcomes. Specifically, we regressed final-round entropy on age, gender, race, education, and political orientation, alongside the experimental manipulation of communication. The model explained little variation ($R^2 = 0.045$, Adj. $R^2 = 0.019$). None of the demographic predictors consistently reached statistical significance, indicating that participant background did not drive observed patterns. By contrast, the social learning manipulation remained a robust predictor ($b = -0.207$, $p = .005$). Together, these results confirm that differences in entropy reflect the causal effect of social learning, rather than imbalances in demographic composition across conditions.

Table S3: Participant demographics and OLS results for final-round entropy

Category	Value	Percent	95% CI	*p*
Age	Mean (Range)	37.8 (12–100)	[–0.001, 0.011]	0.132
Gender	Female	53.8%	—	—
	Other	0.1%	—	—
	Male	45.4%	[–0.123, 0.176]	0.724
	Prefer not to say	0.7%	[–2.366, 0.683]	0.279
Race	African	9.7%	—	—
	Caucasian	70.6%	[–0.229, 0.234]	0.982
	Latinx	6.3%	[–0.278, 0.498]	0.579
	Asian	6.2%	[–0.517, 0.215]	0.417
	Mixed-Race	4.2%	[–0.136, 0.696]	0.187
	Native American	0.6%	[–1.190, 1.272]	0.947
	Other	1.2%	[–0.418, 0.921]	0.460
Education	College graduate	46.3%	—	—
	Some college	25.1%	[–0.118, 0.243]	0.496
	Graduate/Prof. School	16.8%	[–0.429, –0.020]	0.031
	High school	10.4%	[–0.273, 0.216]	0.818
	Did not graduate HS	0.5%	[–0.311, 1.807]	0.166
	Other	0.2%	[–1.079, 3.128]	0.339
	Prefer not to say	0.7%	[–0.518, 1.725]	0.291
Political orientation	Extremely Conservative	4.8%	—	—
	Moderately Liberal	25.2%	[–0.261, 0.456]	0.594
	Extremely Liberal	22.3%	[–0.124, 0.613]	0.194
	Slightly Liberal	18.1%	[–0.107, 0.631]	0.164
	Slightly Conservative	15.4%	[–0.009, 0.741]	0.056
	Moderately Conservative	11.0%	[–0.356, 0.440]	0.836
	Prefer not to say	3.1%	[–0.142, 0.990]	0.142
Experimental condition	Social (vs. Asocial)	—	[–0.352, –0.062]	**0.005**

5.4 Additional analyses

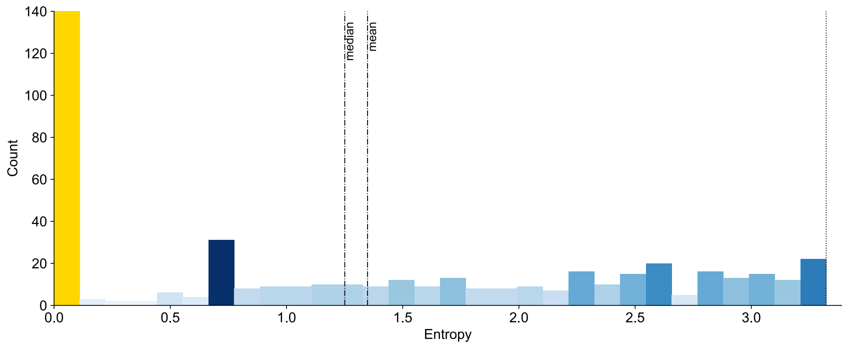
5.4.1 Allocation

Overview. To measure participants' beliefs about the relative productivity of candidate groups, we administered a group-allocation task at the end of the hiring game (see Section 5.1). Participants distributed 100 hypothetical positions across the ten candidate groups. This task provides a direct quantitative measure of perceived group productivity beyond what can be inferred from hiring choices.

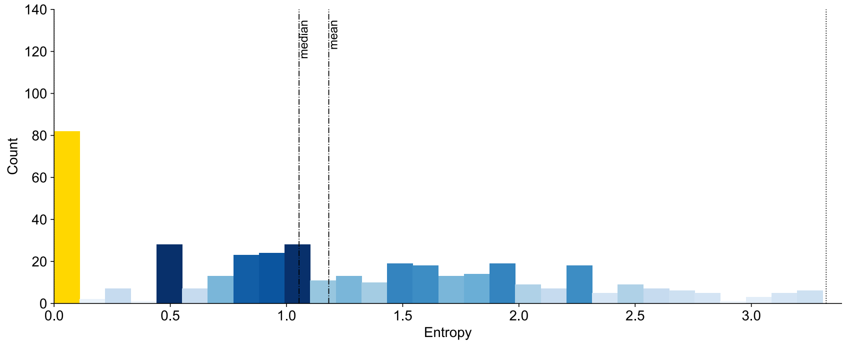
Methods. We analyzed only the equal-productivity conditions so that any differences in allocations reflected learning dynamics rather than true productivity gaps. Allocations were analyzed at both the individual and group level using the entropy metric introduced in Section 2. Higher entropy indicates dispersed allocations, reflecting a belief that the groups are roughly equally productive; lower entropy indicates concentrated allocations, reflecting a belief that only one or a few groups are more productive.

Results. Figure S10 shows the distribution of individual entropy. In the social learning condition, 19.9% of participants allocated all hires to a single group (entropy = 0). Most others also concentrated heavily, with indices clustered between 0.5 and 2.0 and very few approaching the maximum value (3.32). By contrast, in the asocial learning condition, 30.5% of participants showed entropy = 0 and the distribution was more diffuse, with many participants spreading hires in line with the true underlying distribution. Figure S11 displays session-level patterns. Average group-level entropy was lower in the social condition ($M = 1.17$) than in the asocial condition ($M = 1.35$), and in some sessions collapsed to zero, indicating complete convergence on a single group across all candidate groups.

Summary. The group-allocation task corroborates the main behavioral findings: social learning systematically reduces entropy, both within individuals and across groups, making participants more likely to converge on a false belief, whereas asocial learning preserves greater dispersion and more closely reflects the actual underlying distribution.



(a) Asocial learning: individual-level allocation entropy.



(b) Social learning: individual-level allocation entropy.

Figure S10: Distribution of individual-level Shannon entropy from the group-allocation task. Lower entropy values indicate more concentrated allocations (stronger belief in a single group), while higher values indicate more even allocations across groups. Color intensity (blue gradient) reflects the frequency of observations within each entropy bin, and yellow highlights correspond to cases of exact zero entropy (complete concentration on a single group).

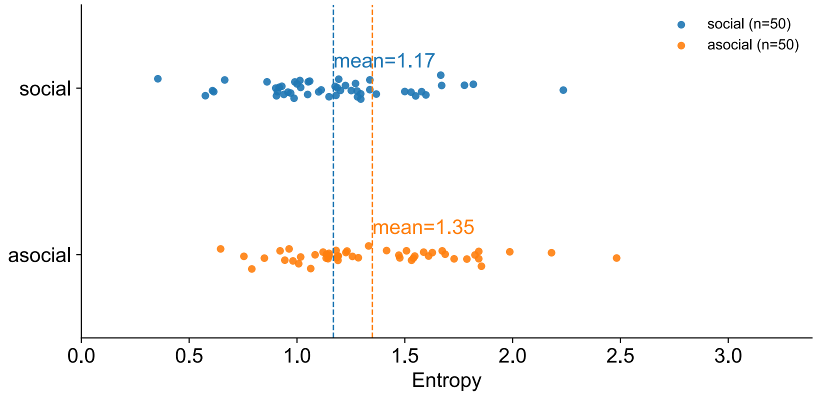


Figure S11: Group-level entropy from the group-allocation task. Each point represents one group's average allocation across 10 participants in a repeated session. Blue points indicate the social learning condition, orange points indicate the asocial learning condition. Dashed vertical lines denote the mean entropy within each condition.

5.4.2 Text analysis

Overview. To understand how participants themselves described their decision strategies, we analyzed the open-ended responses to two exit-survey questions:

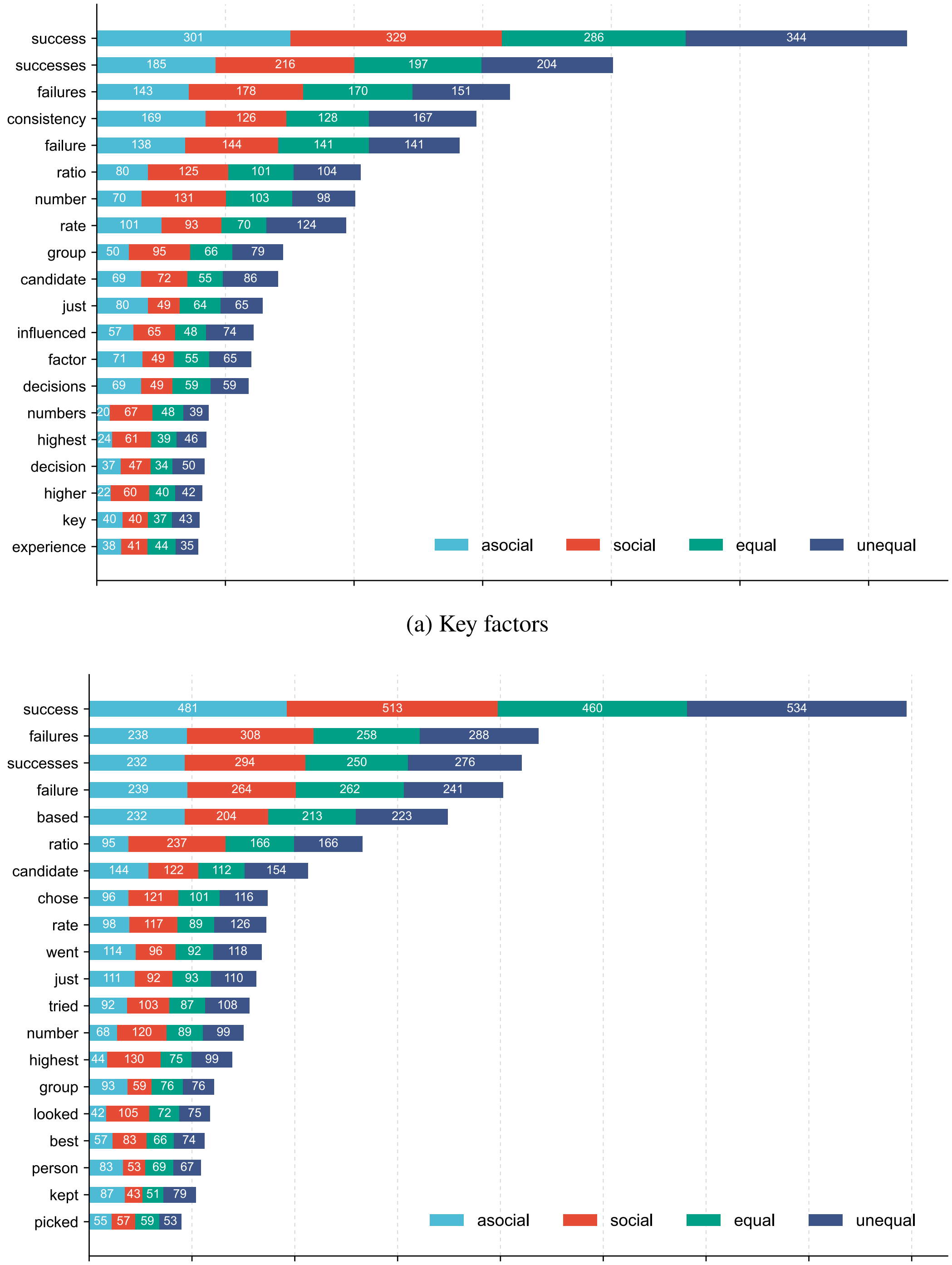
- *"How did you approach decision-making in this experiment?"* (strategies)
- *"What key factors influenced your decisions?"* (key factors).

The goal was to characterize how participants describe their own reasoning and whether language differs across *social* vs. *asocial* learning and *equal* vs. *unequal* reward settings.

Methods. Texts were lowercased, punctuation and numbers removed, and tokenized. We removed standard English stopwords and a short domain-specific stoplist (e.g.,"candidate", "round", "group" when used as UI labels). Tokens were lemmatized to merge verb and noun forms, and frequent or bigrams with high Pointwise Mutual Information (PMI), such as win stay, trial error, and group info were identified (e.g., "win stay", "trial error", "group info"). Minor synonym mapping merged close variants (e.g., "success"/"win"; "observations"/"history"). For each question (strategy, key factors) we then computed term counts and normalized rates $r_w = 1000 \times \frac{c_w}{\text{total tokens}}$, reporting the top unigrams and bigrams by normalized frequency to create an overall lexical profile. Finally, we compared frequency distributions across experimental conditions—(i) social vs. asocial learning and (ii) unequal vs. equal reward structures—to identify which terms were more salient in each setting and to examine how participants emphasized different strategies or factors depending on the learning environment.

Results. As illustrate in Figure S12, the overall frequency distributions across all participants revealed a consistent lexical pattern. For both key factors and strategies, the most common words were "success/successes," followed by "failure/failures," "consistency," "ratio/rate," and "highest/higher." These patterns indicate that participants overwhelmingly prioritized past "success," often framing their decision strategy as hiring the option with the "highest success." References to "failure" and to success-to-failure "ratios" also appeared frequently, suggesting that participants tracked comparative performance rather than relying on single outcomes. Condition contrasts further refined this picture. In the social vs. asocial contrast, "success" remained the most frequent term in both settings, but the secondary emphasis differed: in the asocial condition, "consistency" ranked above "failure," whereas in the social condition this ordering was reversed. Moreover, references to "number" appeared nearly twice as often under social learning as under asocial, and mentions of "highest" were about four times more frequent. This pattern suggests that in social environments, participants were particularly responsive to the absolute "number" of successes—large group-level counts appeared to amplify salience and guided decision-making more strongly than personal routines of "consistency." Finally, in the equal vs. unequal reward comparison, differences were minimal. The overall profiles were stable across reward structures, reinforcing that social learning, rather than baseline productivity differences, was the primary driver of participants' stated decision strategies and key factors.

Implications. These patterns show that social learning amplified attention to absolute success counts. Large group-level numbers of successes became especially salient and guided decision-making more strongly than individual routines of consistency, regardless of the baseline reward structure.



(a) Key factors

(b) Strategies

Figure S12: Top-10 word frequencies in participants' open-ended responses: (a) *key factors* and (b) *strategies*, shown across the four experimental conditions.

5.4.3 Analyzing trial-by-trial responses

Overview. To uncover how participants adapted their choices from round to round, we analyzed every hiring decision using two complementary approaches. First, a rule-based labeling framework classified each decision according to the type of information used and the immediate pattern of reinforcement. Second, a model-based inference estimated, probabilistically rather than heuristically, the relative influence of personal evidence, group evidence, and previous-round outcomes. Together, these analyses reveal how social learning reshapes the micro-level strategies that drive convergence.

Methods. We used two complementary analytic approaches detailed as follows.

Rule-based labeling framework. Each decision received two complementary labels. A "judgment label" indicated the informational basis for the choice, and a "pattern label" captured how the current decision related to the participant's previous outcome. Missing or inapplicable cases were coded as N/A. In the asocial learning condition, participants had access only to their own history of choices and outcomes. A decision was labeled "exploit" if it selected the option best supported by private evidence, defined as the arm with the highest cumulative number of successes, the most favorable success–failure ratio, or the fewest accumulated failures. Any other observed choice was labeled "explore". Pattern labels were derived solely from the immediately preceding round: repeating a successful choice was "Win–Stay", repeating a failed choice was "Lose–Stay", switching after a failure was "Lose–Shift", and switching after a success was "Switch–After–Win". In the social learning condition, participants additionally observed group-level cumulative outcomes. Judgment labels were extended to incorporate collective information. A choice that matched the most popular option from the previous round was labeled "majority-biased"; a choice that aligned with the arm best supported by cumulative group-level evidence (highest number of successes, best success–failure ratio, or fewest failures) was labeled "social-exploit"; and any other observed choice was coded as "contrarian". Pattern labels were assigned exactly as in the asocial condition, based solely on the individual's own prior choice and reward. This labeling framework yields individual-level profiles of strategy use, for example the proportion of "exploit" versus "explore" or "majority-biased" versus "contrarian" choices, which can be aggregated across conditions to compare decision-making tendencies between asocial and social learning environments.

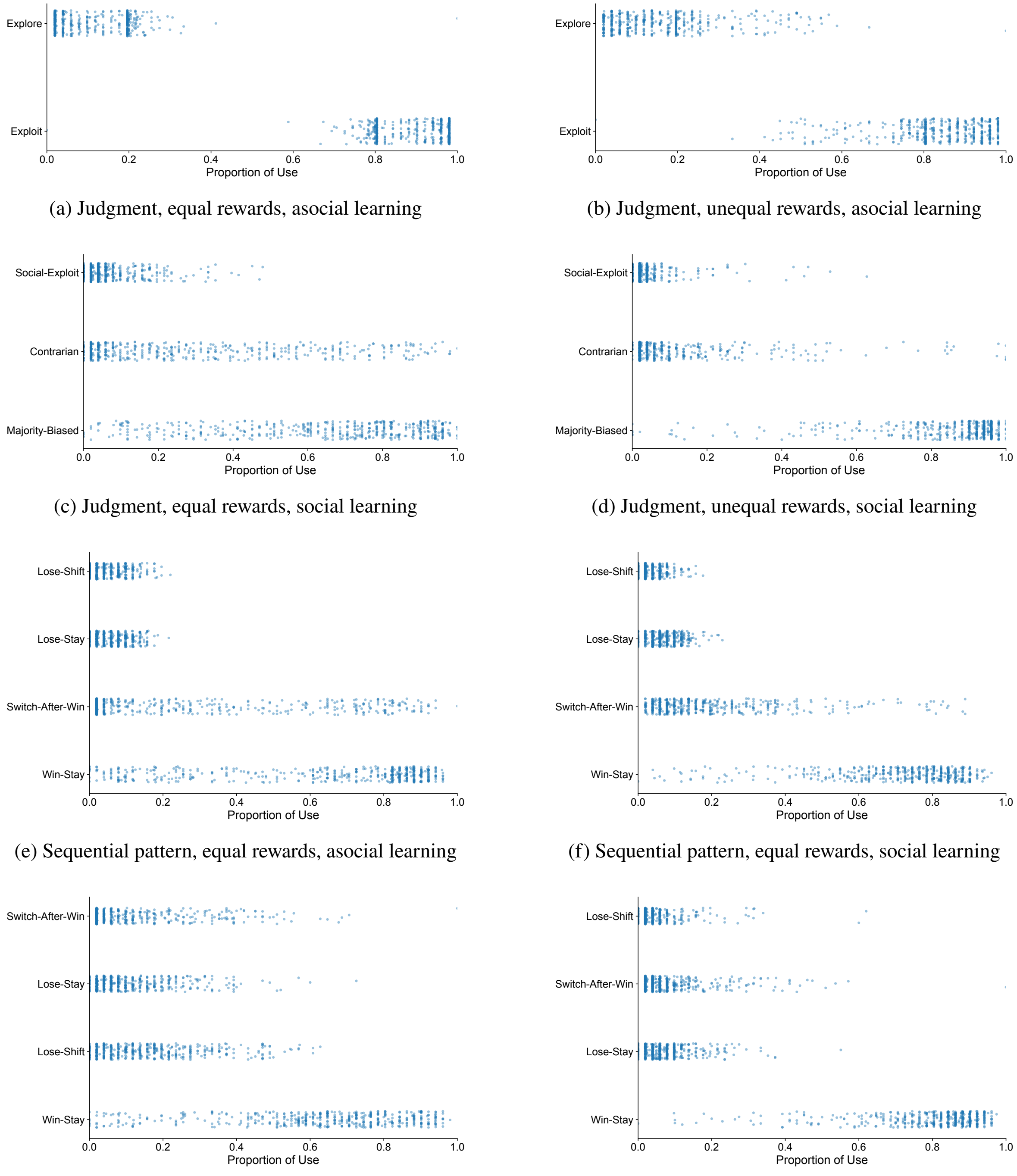
Model-based inference of decision rules. For each participant and each round we constructed a state representation with three components: "own evidence"—the cumulative successes and failures from that participant's past choices; "group evidence"—the cumulative successes and failures aggregated across all participants and available only in the social learning condition; and "previous-round reinforcement"—whether the participant repeated or switched relative to their last choice and outcome. We then generated posterior draws from Beta distributions Beta(α, β) for each arm, separately for own and group evidence. For each mechanism, the candidate "optimal" arm or arms were identified as those with the highest sampled values across 200 posterior draws, and the participant's observed choice was compared to these candidates. If the choice matched the own-optimal arm(s) in a majority of draws, it was labeled "own-based"; if it matched the group-optimal arm(s) in a majority of draws, it was labeled "group-based"; and if it repeated the prior arm after a success, it was labeled "previous-round." Choices that fit none of these rules were classified as "other." Because a single decision can reflect multiple influences, we reported both "concurrent" labels—allowing multiple categories for the same choice—and "exclusive" labels, which applied a fixed priority: "previous-round" > "own" > "group" > "other".

Results. Labeling decisions across the four experiments revealed systematic contrasts between learning conditions, as detailed in Figure S13.

Rule-based labeling. In the asocial learning condition, "exploitative" judgments dominated across both reward settings: the proportion of "exploit" decisions was substantially higher than "explore", confirming that participants primarily relied on private evidence. The exploit rate was even higher under the equal-reward setting, suggesting that when all candidates were equally good participants were less inclined to test alternatives and instead reinforced initial preferences. Despite this aggregate pattern, considerable heterogeneity remained across individuals, especially in the unequal-reward setting: some participants displayed stronger exploratory tendencies while others quickly locked into exploitation. These observations imply that private reward histories shaped idiosyncratic strategies, but in the aggregate equal rewards encouraged even stronger reliance on exploitation. In the social learning condition, judgments were shaped predominantly by collective signals. Across both reward settings, "majority-biased" decisions overwhelmingly outnumbered "social-exploit" and "contrarian" choices, underscoring the strength of conformity to group behavior. The frequency of majority-biased decisions was even higher under unequal rewards, consistent with rapid convergence once a superior group became apparent. At the individual level, however, distributions of strategies remained more dispersed than in the asocial case, suggesting that participants weighed collective evidence differently. Even so, inequality systematically emerged in the social condition, reflecting the amplifying force of majority dynamics. Turning to sequential patterns, "Win–Stay" dominated in all four experiments, consistent with limited exploration overall. Yet differences emerged between conditions. In the social condition the proportion of Win–Stay responses was higher and clustered tightly around 0.7, showing stronger reinforcement from shared outcomes. In contrast, the asocial condition showed a more uniform distribution across participants, with some individuals displaying extremely high reliance on "Switch–After–Win" (exceeding 0.9). By comparison, "Lose–Stay" and "Lose–Shift" occurred at relatively similar rates, suggesting that failures did not decisively push participants toward consistent switching. These contrasts highlight how social learning reinforces convergence by amplifying successful choices, particularly when rewards are unequal and one group is perceived as superior.

Model-based inference. The model-based analysis corroborated and extended these labeling results. Reliance on "own evidence" was similar across conditions (social: 26.3%, asocial: 28.1%), indicating that participants consistently incorporated their own success histories. The key divergence emerged in "group evidence": under social learning, 20.5% of choices aligned with group-level posteriors, almost double the rate in asocial learning (11.9%). Reinforcement from prior successes also increased under social learning, with "Win–Stay" reaching 71.2% compared to 62.6% in asocial learning. Conversely, the share of "other" strategies declined (20.3% vs. 29.1%), suggesting that social learning reduced random or inconsistent choices, making participants' behavior more consistent. To examine heterogeneity across participants, we plotted the distribution of decision weights at the individual level. Each participant's choices were decomposed into the proportion of decisions attributed to "own-based", "group-based", "previous-round", or "other" rules (Figure S14). Although individual reliance varied widely—from near-exclusive dependence on one rule to more balanced mixtures—systematic differences between social and asocial conditions nevertheless emerged. Under social learning, reliance on group evidence shifted upward and "Win–Stay" proportions clustered more tightly, indicating stronger convergence despite persistent heterogeneity across individuals.

Summary. Taken together, these analyses demonstrate that social learning systematically shifts weight toward collective evidence and amplifies reinforcement from prior successes, producing stronger convergence dynamics relative to the more fragmented strategies observed in asocial learning.



(a) Judgment, equal rewards, asocial learning

(b) Judgment, unequal rewards, asocial learning

(c) Judgment, equal rewards, social learning

(d) Judgment, unequal rewards, social learning

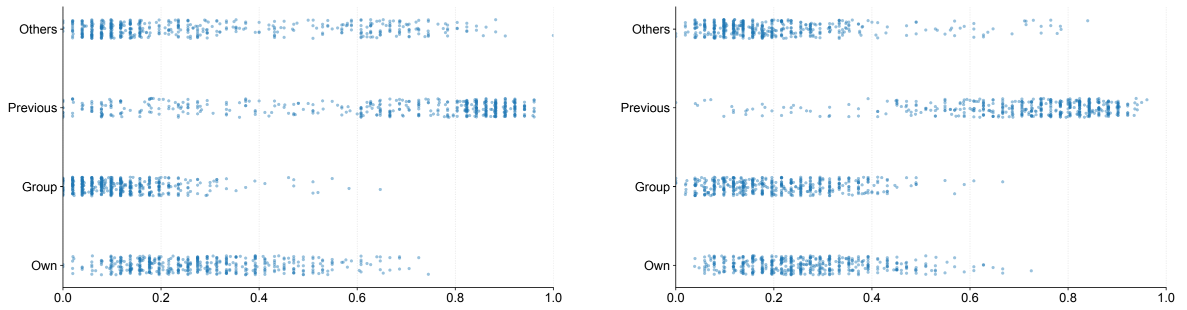
(e) Sequential pattern, equal rewards, asocial learning

(f) Sequential pattern, equal rewards, social learning

(g) Sequential pattern, unequal rewards, asocial learning

(h) Sequential pattern, unequal rewards, social learning

Figure S13: Judgment and sequential pattern labels across all experimental conditions.



(a) Asocial condition.

(b) Social condition.

Figure S14: Individual-level distributions of decision rule reliance. Each point represents one participant's proportion of choices attributed to a given rule.

References

[1] A. Almaatouq, J. Becker, J. P. Houghton, N. Paton, D. J. Watts, and M. E. Whiting. Empirica: a virtual lab for high-throughput macro-level experiments. *Behavior Research Methods*, 53(5):2158–2171, 2021.

[2] X. Bai, S. T. Fiske, and T. L. Griffiths. Globally inaccurate stereotypes can result from locally adaptive exploration. *Psychological Science*, 33(5):671–684, 2022.

[3] A. Bardhi, Y. Guo, and B. Strulovici. Early-career discrimination: Spiraling or self-correcting? *Duke University and Northwestern University Working Paper*, 2020.

[4] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[5] A. Krishnamurthy, K. Harris, D. J. Foster, C. Zhang, and A. Slivkins. Can large language models explore in-context? *Advances in Neural Information Processing Systems*, 37:120124–120158, 2024.

[6] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[7] P. Landgren, V. Srivastava, and N. E. Leonard. Distributed cooperative decision making in multi-agent multi-armed bandits. *Automatica*, 125:109445, 2021.

[8] D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, Z. Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

[9] R. S. Sutton, A. G. Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.