# Social coordination perpetuates stereotypic expectations and behaviors across generations in deep multiagent reinforcement learning

Rebekah A. Gelpí [ID][a,b,c,*], Yikai Tang [ID][a,b], Ethan C. Jackson[a,b] and William A. Cunningham [ID][a,b,c,d]

[a]Department of Psychology, University of Toronto, 100 St. George Street, Toronto, Ontario, Canada M5S 3G3
[b]Vector Institute, 101 College Street, Toronto, Ontario, Canada M5G 1L7
[c]Schwartz Reisman Institute for Technology and Society, University of Toronto, 101 College Street, Toronto, Ontario, Canada M5G 1L7
[d]Department of Computer Science, University of Toronto, 40 St. George Street, Toronto, Ontario, Canada M5S 2E4
*To whom correspondence should be addressed: Email: rebekah.gelpi@mail.utoronto.ca
**Edited By** Attila Szolnoki

## Abstract

Despite often being perceived as morally objectionable, stereotypes are a common feature of social groups, a phenomenon that has often been attributed to biased motivations or limits on the ability to process information. We argue that one reason for this continued prevalence is that preexisting expectations about how others will behave, in the context of social coordination, can change the behaviors of one's social partners, creating the very stereotype one expected to see, even in the absence of other potential sources of stereotyping. We use a computational model of dynamic social coordination to illustrate how this "feedback loop" can emerge, engendering and entrenching role-consistent stereotypic behavior and then show that human behavior on the task generates a comparable feedback loop. Notably, people's choices on the task are not related to social dominance or system justification, suggesting biased motivations are not necessary to maintain these stereotypes.

**Keywords:** reinforcement learning, agent-based model, stereotyping, social norms, coordination

## Significance Statement

People often use social role stereotypes—generalized information about the roles associated with a social category—as a source of information about the skills and preferences of strangers. Social role stereotypes simplify the task of learning how different people behave but also create an expectation that group members will behave a certain way. When coordination is necessary for success, these expectations can engender a feedback loop creating the very stereotype one expected to see. Our computational model of social coordination shows how these feedback loops can emerge, and adults exhibit similar patterns of stereotypic behavior, irrespective of their preexisting attitudes towards social inequality.

## Introduction

The widespread presence of stereotyping—generalized, and often exaggerated, inferences about individuals based on their social category membership—is a broad social concern that presents a challenge to group diversity. Groups contain individuals who may have differing characteristics, skills, and needs; stereotypes threaten this by projecting similarity onto members of a social category that may not accurately reflect the diversity and variability of individuals within it. Many classic and contemporary theories of social cognition have sought to explain why otherwise well-meaning individuals might nevertheless engage in stereotyping. Some of these theories suggest that motivations such as promoting one's group at the expense of another's (1, 2) or justifying existing inequalities (3–5) conflict with norms against stereotyping; others argue that the limits of human cognition make overly simplistic generalization inevitable (6–8).

We argue that part of the reason stereotyping remains so prevalent despite the existence of norms against is because these stereotypes also result from the dynamics of social coordination, where social partners shift their behaviors to align with the preexisting expectations of the social environment in which they find themselves. In large societies, people must interact with strangers with whom they have almost no experience, requiring them to "go beyond the information given" (9) to make inferences about the traits, motivations, and beliefs of others based on secondary cues. Many social category representations include beliefs about

the social roles that individuals within the category are suited for ([10, 11]); these role-based stereotypes facilitate social cognition by simplifying one's representation of others, and ease social dilemmas about how to divide labor ([12]), but they can lead to a vicious cycle in which the target of an expectation who successfully coordinates with a social partner amplifies that partner's expectation that the target will continue to behave according to that expectation.

Using a combination of computational and behavioral work, we illustrate how such dynamics can emerge from the cognitive processes of individuals operating within a social system, and how agents in these systems can inadvertently contribute to a social structure that helps create, maintain, and strengthen social biases against individuals based on their social roles.

## Statistical regularities and social conventions

Modern human social groups are large and deeply complex. In order to successfully develop and maintain these societies, people must successfully learn to coordinate with one another, dividing labor between individuals to allow for specialization and large-scale cooperation (e.g. ([12–15])).

Coordinating successfully in such large groups poses a challenge: reliably knowing how to assign tasks with an individual requires understanding that individual's skill set, but as the size of a social group increases, representing each individual within it with high fidelity likewise incurs a greater cognitive cost. As finite cognitive resources constrain the number of simultaneous close social relationships we are able to entertain ([16–19]), several theories of social cognition have characterized people as "efficiency experts" (e.g. ([6–8])) for whom group-based generalizations result from an attempt to leverage the utility of low-cost category knowledge, thus avoiding the more effortful task of individuation. Generalizing these regularities to unfamiliar others, in turn, facilitates the development of shared knowledge and the ability to coordinate one's behavior.

Given these potential advantages, social norms or conventions that members of a group ought to engage in a particular behavior or share a certain belief can further simplify the space of expected behavior, easing coordination by making each group member more predictable ([20]). For example, computational accounts of the development and maintenance of conventions such as communication and cooperative foraging have emphasized the need not only to coordinate with individuals but to abstract these rules for interactions away from individual relationships and generalize them to a larger community (e.g. ([21])), and punish those whose behavior does not align with the norm (e.g. ([22–25])).

The codification of regularities into expectations and social conventions, while often providing substantial benefits to group members, can come with potential drawbacks, both to individual members within the group and also to the group as a whole. By compressing our highly complex social environments into easily interpretable generalized beliefs about social categories, individuals whose skills or preference do not match what a convention dictates can be unfairly disadvantaged.

Game theory has long used economic games such as the Bach-Stravinsky game (also known as the Battle of the Sexes; see e.g. ([26, 27])) to model how self-interested agents (and subgroups of agents) develop stable strategies for coordinating with one another. Coordination in such games reliably outperforms uncoordinated behavior, but in many cases, the rewards of coordination are not equally distributed. At the individual level, even when groups have an equal underlying distribution of skills

(i.e. which group specializes in which task is ultimately arbitrary), the fact that many specialized tasks within human societies require substantial investment creates strong incentives for individuals to coordinate, even when it is personally detrimental (e.g. ([28])). Coordination can also result in disadvantages for groups as a whole; for example, when coordination involves tasks with unequal rewards, one group is reliably disadvantaged ([12]), particularly when tasks provide surpluses ([14]) or when one group is smaller than another ([29, 30]). In all of these cases, while coordination may be useful in the sense that it results in better outcomes as a whole, it can come at the cost of disadvantaging those whose skills are not reflected by this simplified representation.

## Learning signals and feedback loops

Although game-theoretic analyses have provided a powerful tool for understanding how inequality can emerge from the need to coordinate, these models often do not capture the process by which individuals within a society learn and change their behavior. When learning agents attempt to coordinate, their actions also change one another's reasoning about optimal actions. For example, believing that one's partner prefers to clean and does not like to cook might make one more likely to offer to cook than to clean, all else being equal; however, this also means that one's partner is more likely to offer to clean than to cook in the future, even if the partner actually had no strong feelings to begin with.

In this way, learning systems can create positive feedback loops between expectations and outcomes. When these inferences involve generalized beliefs about social categories as a whole, they regularize one's experience of social groups more than would be expected by chance, strengthening the associations between a social category and an expectation not just for oneself, but also for newly introduced agents.

Another limitation to modeling coordination with game theory is that these models cannot capture agents' beliefs about regularities—whether agents believe groups have underlying differences, or whether regularities reflect an arbitrary convention agreed upon for the sake of efficiency or mutual benefit. If agents believe the outcomes of a coordinated system reflect an essential difference between groups, they may resist changes to that system and justify its resulting inequalities; however, this tendency may be mitigated if the convention is perceived as an arbitrary rule that it is possible to arbitrarily change.

One piece of evidence that agents' preexisting beliefs can contribute to the feedback loop described above comes from the phenomenon of "behavioral confirmation" observed for a number of real-world social categories, such as race and gender ([31, 32]). In one such task by Skrypnek and Snyder ([33]), male and female students were paired together and tasked with negotiating a division of labor to solve a set of stereotypically masculine and feminine tasks without face-to-face interaction. Male students allocated more masculine tasks to their partners when they believed they were negotiating with a male student; on the following task, the female students who had been perceived as male allocated more masculine tasks for themselves. Thus, the initial perceiver's stereotypic beliefs about men and women were not only confirmed by the target's behavior in the first task, but strengthened by the target's own subsequent choices. Therefore, if such role-based stereotypes can emerge from behavioral confirmation, then generalized conventions that disadvantage or misrepresent a group of individuals can emerge even among unbiased individuals with sufficient cognitive capacity to fully represent the complexity of their social interactions. Comparable dynamics can emerge from approach-contingent

learning biases, wherein initially negative exposures result in entrenched avoidance of a group, which prevents further learning (e.g. (34, 35)).

This possibility offers a potential answer to a fundamental question in social psychology: why people engage in stereotyping, despite endorsing moral norms against the use of stereotyping. In such situations, people may believe that they are simply drawing statistically appropriate conclusions about how groups are likely to behave (36, 37), even though their own professed belief may itself create, entrench, and strengthen the stereotype that individuals within a group are well-suited to a specific social role and will therefore behave role consistently.

We illustrate the development of role-based stereotypic conventions within populations of skilled agents that learn and change in response to their environment using multiagent deep reinforcement learning (MARL). MARL models are an emerging tool in the fields of artificial intelligence and cognitive science to understand social interaction and coordination problems (38–42), particularly those involving the emergence of social conventions (24, 25, 43), and open several new lines of inquiry that have not been possible using traditional game-theoretic approaches. Using our models, we test two hypotheses regarding the emergence of stereotypic conventions: (i) that these effects are most prominent in larger populations, as coordination around group labels is learned more quickly than the skills of individual agents and (ii) that learning agents develop a structure that is learned and perpetuated by subsequent generations of new agents, such that agents' initial expectations continue to affect the structure of the world even when those agents are no longer observed.

These models allow us to test whether agents optimizing their own anticipated rewards will develop and persist in stereotypic patterns of behavior. However, while this provides a baseline for agents optimizing expected utility, human beings can act against their economic interests to behave prosocially (e.g. (44–46)). Thus, people who are aware that an underlying convention is arbitrary in ways that disadvantage a group would be less likely to endorse that convention—but only to the extent that they are aware such a difference exists. By testing human participants on a comparable scenario to our MARL models, we can directly elicit explicit beliefs about the structure of the environment that they are learning in, allowing us to link the stereotypic actions that may result from social coordination, to the stereotypic beliefs that exist in our daily life. If human participants believe that the conventions observed in our tasks reflect a true, underlying essential difference, then participants on our task should not only make stereotype-consistent choices, they should do so irrespective of individual differences in tolerance for social inequality.

## Task setup

At a high level, we investigate the effect of this feedback loop on an environment in which there are differences in the benefits or needs of individual members within a social group. For example, a social group might have a larger number of people who are skilled lumberjacks, and a smaller number who are skilled miners. In a large enough social group, this group might be represented as "lumberjacks," establishing a norm that members of the group ought to chop wood. If the group's success depends on successful coordination, the mechanisms of expectation and enforcement—the "carrot" and the "stick" of convention formation—systematically disadvantage those who would be better served by mining, even as the majority is advantaged. Worse yet, if the disadvantaged miners do accede to the norm, engaging in

chopping wood despite their lack of skill, this strengthens the perceived regularity that members of the group are lumberjacks, and worsens the misrepresentation by the broader population of the underlying diversity of the social group.

We adapt the AI economist paradigm (47, 48) to model this environment. Agents in our studies participate in a produce-and-trade task where they collect resources from the environment, which they can use to trade with a market decider (described below) in exchange for rewards (Fig. 1). By making agents' rewards contingent on successful trades with an external market, this task allows for analyzing how the agents' and markets' simultaneous learning of the dynamics leads to patterns of social coordination.

Although the coordination demand between the agent and the market may seem to be trivially resolved, e.g. by communicating one's intentions in advance, thus preventing failures due to misaligned behaviors, this setup is identical to a typical coordination problem, and serves as an abstraction for situations where individuals actively select interactions with others about whom they know very little beyond preexisting beliefs about a social category. Thus, the market functions as a simplification of generalized society-wide beliefs about the degree to which an individual agent's social category membership predicts its conformity to a social role.

Each agent's input consists of its current reward value as well as its resource total (wood and stone). Each agent takes a single action per turn, with an action space of seven total possible actions. Three actions are independent actions: chopping wood (receive 1 wood), mining stone (receive 1 stone), or building a house (consume 1 wood and 1 stone, receive 15 points of reward). These actions do not involve the market decider, and their success rate depends on the skill levels of each agent (e.g. an agent with a skill level of 0.95 in chopping would successfully receive 1 wood on 95% of trials spent chopping).

The other four actions are social actions: buying wood, selling wood, buying stone, or selling stone. Successful selling actions require coordination between the agents and the market decider. Whenever an agent decides to sell a resource, the market decider has to predict the resource that the agent is selling; the transaction is only completed if the market decider makes the correct guess about what the agent is attempting to sell. After a successful transaction, the agent trades away 2 units of wood or stone, and receives 1 point of reward. Buying actions do not require coordination; if an agent buys a resource from the market decider, it loses 2 points of reward and receives 1 unit of wood or stone.

Unlike the agents, the market decider does not necessarily act at every timestep, nor is it limited to acting only once every timestep; rather, it only takes an action when agents opt to sell to the market decider. In each trade, the market decider's input consists of a unique 16-digit binary code corresponding to the agent's identity. The market decider receives positive or negative rewards depending on its prediction of agents' selling behaviors. It receives 1 point of reward when an agent successfully sells a resource to it, loses 1 point of reward if it does not correctly predict the agent's behavior, and loses 0.3 points of reward if it makes a successful prediction, but the agent does not have enough resources to make a successful transaction.

## Results

The design of the social group structure in our studies is aimed to enable the possibility of the emergence of social conventions. Thus, although agents can specialize in any behavior (mining, chopping, or building), we focus our analyses on the behaviors
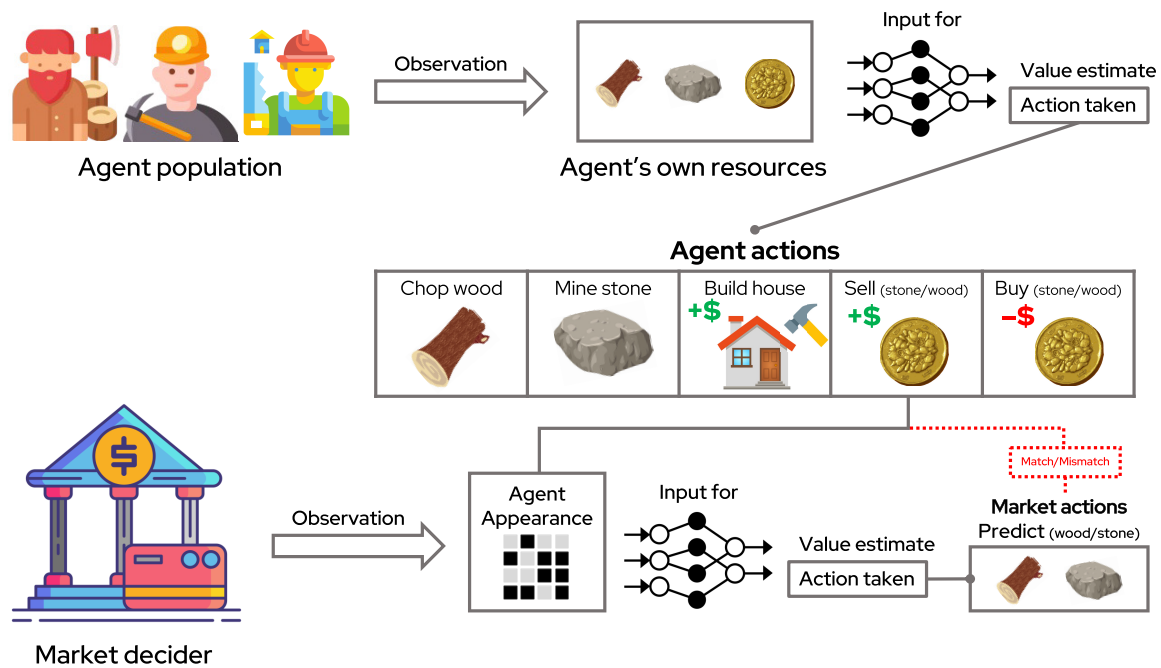
**Fig. 1.** Nontechnical schematic of task design. Agents (top) with varying skill specializations observe their own resources and input this information into a neural network to determine their actions. The market decider (bottom) observes the agent's appearance when it chooses to sell to the market, and input this information into its own neural network to determine whether to predict that the agent will sell wood or stone. Successful sales (and the agents'/market decider's reward) depend on the prediction matching the agent's action.

of mining and chopping specialist agents, and the degree to which they adhere to conventions in groups that are composed of a majority of choppers or miners. When selling to the market decider, these agents can behave according to their skill (by offering to sell the resource that they are skilled at collecting) or contrary to their skill (by offering to sell the resource that they are not skilled at collecting). Below, we examine the proportion of skill-based predictions by the market decider, and the proportion of skilled and unskilled behaviors by agents to measure the emergence of group-level conventions among the groups.

## Study 1: Individuation

To determine whether our model has the capacity to fully individuate all of the agents, we first conduct a series of baseline experiments, in which agents are assigned their identity codes randomly, and there is no correlation between these codes and the agent's skill. Thus, for the agents and the market decider to successfully coordinate, the market decider must learn to predict each agent's skill based on its unique identity code. If agents act according to their skills on this task, then any stereotyping effects observed in later studies must result from something other than a lack of model expressivity. Further, we can observe how differences in population size are reflected in the time it takes our models to learn individuals' skills.

Our results from this study suggest that the basic structure of the task is learnable by both the agents and the market decider, giving us a benchmark of how well our market decider model is capable of representing individuals' unique skill. Agents predominantly behave according to their skill; for example, skilled choppers mostly chop wood and sell wood, while skilled miners mostly mine stone and sell stone (Fig. 2A and B). Meanwhile, the market decider learns to expect that agents will behave according to their skilled behavior to a high degree of accuracy; the market decider's predictions match agents' skilled behavior 95% of the time (Fig. 2C).

The market decider is most accurate at making skill-based predictions when the size of the population is smaller, $F(3, 7988) = 1692.53$, $P < 0.001$, $\eta_p^2 = 0.39$, and the market decider's predictions become more accurate over time, $F(1, 7988) = 14127.42$, $P < 0.001$, $\eta_p^2 = 0.64$. However, the larger the population, the longer it takes for the market decider to reach a high level of accuracy, $F(3, 7988) = 615.39$, $P < 0.001$, $\eta_p^2 = 0.19$ (Fig. 3A). Broadly, we find that although the time taken to learn the population varies based on the group size, the market decider is able to predict how agents are likely to behave according to their true underlying skills. Due to the structure of the learning task in this condition, however, the market decider's accuracy in its predictions about how unobserved agents are likely to behave cannot change from 50% throughout training, because the skills and identity of unobserved agents is uncorrelated with the skills and identity of existing agents.

## Study 2: Regularity

As population sizes increase, people have fewer opportunities to learn about each individual within a population; correspondingly, people may rely on secondary cues such as group identity if these cues can predict any existing variance. To the extent that these secondary cues are used for prediction in coordination tasks, and these predictions lead to expectations that influence others' behavior, we should expect that when a secondary cue is statistically correlated with a skilled behavior, all agents with that cue should behave more conventionally, even if this means behaving in a way that conflicts with the agent's personal skill.

In study 2, we examine this by making each agent's skills statistically correlated with three of the digits of the agent's identity code, thus functioning as a group label. While 50% (the majority) of the agents with a given group label share a skill specialization, the remaining 50% is split evenly between the other two skill specializations (the minorities). In the groups where the majority of agents are chopping or mining specialists, this yields a ratio of
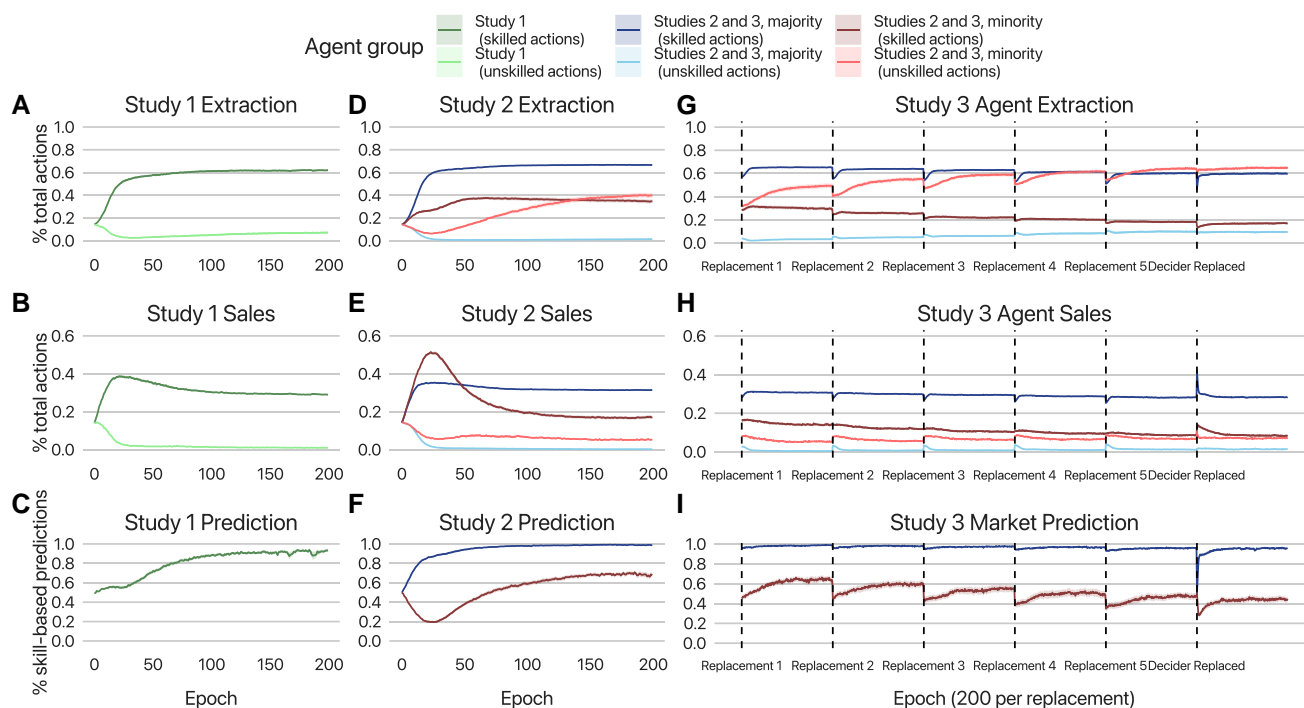
**Fig. 2.** Agent and market behavior for studies 1–3. Agent extraction and selling behavior, and market decider's predictions, for studies 1–3, with a population size of 300 agents. A, B) Agents' extraction (A) and selling (B) behavior in study 1. Agents learn to extract and sell resources that match their personal skills (dark green), and rarely extract or sell resources that they are unskilled at collecting (light green). C) Market decider's predictions in study 1. The market decider's predictions match the agents' personal skills 95% of the time. D, E) Agents' extraction (D) and selling (E) behavior in study 2. Majority agents almost exclusively match the stereotypic belief about the group as a whole, extracting and sell resources that they are skilled at collecting (dark blue) rather than the resources that they are unskilled at collecting (light blue). Minority agents, on the other hand, progressively extract and sell lower proportions of the resources they are skilled at collecting (dark red) and higher proportions of those they are unskilled at collecting (light red) over time. F) Market decider's predictions in study 2. The market decider almost always predicts that majority agents (dark blue) will act according to their skill but predicts that minority agents will often not act according to their skill (dark red). G, H) Agents' extraction (G) and selling (H) behavior in study 3. As agents are iteratively replaced (dashed lines), the patterns observed in study 2 strengthen; majority agents continue to mostly extract and sell resources they are skilled at collecting (dark blue) rather than at collecting (light blue). Minority agents' behavior becomes more stereotypic, mostly extracting the resources they are unskilled at collecting (dark red) rather than skilled at collecting (light red), and selling both skilled and unskilled resources at comparable rates. I) Market decider's predictions in study 3. As new agents are introduced (dashed lines), the market decider's predictions for the minority agents (dark red) becomes progressively less skill-based and more stereotypic, while its predictions for the majority agents (dark blue) are almost exclusively skill-based.

2:1 between the size of the chopping or mining specialist majority and the corresponding mining or chopping specialist minority, respectively (see Figs. S2–S5 for results with other ratios).

Critically, all individuating information from study 1 is still present, so it is possible for the model to fully learn each agent's skilled behavior. However, because both the market decider and the agents are learning the task dynamics simultaneously, if the market decider develops expectations about agents based on the group label before it has learned individual agents' skills, then agents may begin to display behavioral confirmation, establishing a stereotypic convention that regularizes the market decider's learning signal and precludes any ability to learn about individuals' unique skilled behaviors.

We find this behavioral confirmation effect emerges in study 2. The market decider quickly learns to predict that agents with a given group label will mostly attempt to sell the same resource that the majority of their group is skilled at collecting, i.e. predicting that most agents in the majority-miner group will bring stone, and most agents in the majority-chopper group will bring wood (Fig. 2F). As a result, agents in these conditions act according to the group label observed by the market decider, rather than their own personal skills: the minority group engages in a significantly lower proportion of skilled actions than agents in study 1 (extraction: $B = -0.209$, SE = 0.002, 95% CI = $[-0.213, -0.205]$,

$z = -122.66$, $P < 0.001$, Cohen's $d = -1.42$; sales: $B = -0.11$, SE = 0.001, 95% CI = $[-0.111, -0.109]$, $z = -83.02$, $P < 0.001$, Cohen's $d = -0.96$; Fig. 2D and E), and by contrast, majority group members engage in a slightly greater proportion of skilled actions than agents in study 1 (extraction: $B = 0.065$, SE = 0.002, 95% CI = $[0.061, 0.069]$, $z = 38.15$, $P < 0.001$, Cohen's $d = 0.44$; sales: $B = 0.037$, SE = 0.001, 95% CI = $[0.035, 0.041]$, $z = 27.90$, $P < 0.001$, Cohen's $d = 0.32$). Thus, even though the same market decider model was capable of learning individuals' skills to a high degree of accuracy, the confounded statistical cue leads the model to engage in stereotyped predictions.

Study 2 also illustrates how the complexity of learning each individual as population size increases affects the market decider's predictions when there is a statistical regularity. It is more difficult and time-consuming to learn the individual skill of each agent in a larger population. On the other hand, a statistical regularity—that more agents with a given group label have a particular skill—can be learned more quickly because of its simplicity. Further, it does not become more difficult to learn as the population size increases. In a population size of 30, the market decider predicts that agents with the minority skill in a group will behave according to their own skill most of the time, albeit still less often than the majority agents ($B = 0.080$, SE = 0.002, 95% CI = $[0.076, 0.085]$, $z = 43.01$, $P < 0.001$, Cohen's $d = 0.78$);
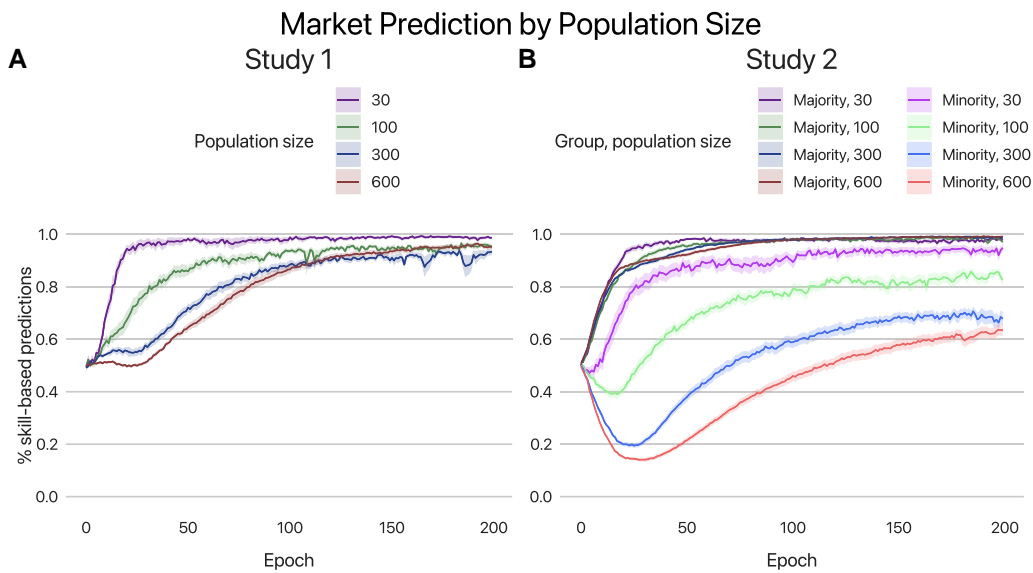
## Market Prediction by Population Size



**Fig. 3.** Market predictions with differing population sizes. In study 1 A), the market decider learns to predict that agents will sell the resource that matches their skill. Larger population sizes take longer to learn but does not affect the ultimate frequency of skill-based predictions. In study 2 B), the market decider's predictions quickly converge to skill-based predictions the majority group members (i.e. chopping specialists sell wood and mining specialists sell stone); however, the market decider's predictions for minority members depends on group size, such that the market decider makes mostly skilled-based predictions for the minority in the population size of 100, but almost exclusively stereotypic predictions for the population size of 600.

however, as the population size increases, the market decider's predictions, and consequently the agents' behaviors, become more highly stereotyped, with minority agents becoming much less likely to behave according to their skills than majority agents (all $z \geq 116.99$, $P < 0.001$, Cohen's $d \geq 2.14$; Fig. 3B).

As both the market decider and the agents are learning the task dynamics simultaneously, the fact that the market decider rapidly learns and applies a statistical cue before it is able to learn individuals' skilled behavior means that agents begin to change their own behaviors to match the market decider's predictions. In other words, these predictions result in a behavioral confirmation effect —the market decider's predictions lead agents to behave in a more stereotypic fashion. This behavior is, in turn, observed by the market decider, strengthening and entrenching the prediction it makes about these agents.

## Study 3: Generational transmission

In study 2, we demonstrate that behavioral confirmation can develop simply by including a statistical cue that is more easily learned than an individual's skill. This effect, in particular, disadvantages agents whose skill specialization is a minority within the larger group.

However, social environments are not static, and it is possible for underlying characteristics such as the correlation between a secondary cue and a skill to change over time. Stereotypic expectations developed over the course of history might be capable of maintaining stereotypic behaviors long after any underlying group differences have faded. In study 3, we examine how stereotype-consistent behavior can be maintained from stereotypic group-based expectations, even when groups have no underlying differences.

In this study, we continue to train the models from study 2, but begin to iteratively replace agents from the population, 20% at a time, with agents whose skill specializations are uncorrelated with the group label (i.e. an equal proportion of chopping and mining specialists). After all agents have been replaced, we additionally replace the

market decider with a new, untrained market decider. At this point, there is no longer any statistical regularity between the group label and the agents' skills, similar to study 1, and the new market decider has not developed expectations about any of the agents.

Because the initial market decider's expectations have already been established and reinforced, new agents face a strong pressure to behave according to the existing convention. Regardless of their true skills, new agents are rewarded if they can successfully coordinate with the market decider. As a result, although new agents introduced into the environment are evenly divided between chopping and mining specialists, these agents increasingly behave according to the stereotypic convention, and the difference between majority and minority agents' selling and extracting behavior becomes progressively more stereotyped between the initial replacement and the final replacement, all $B \geq 0.21$, $SE = 0.002$, $z \geq 109.02$, $P < 0.001$, Cohen's $d \geq 1.99$ (Fig. 2G and H). This pattern of increasingly stereotyped behavior by both minority and majority agents suggests that the behavioral confirmation of the market decider can create self-sustaining stereotypic behavior in the absence of underlying group differences. As the market decider's learning signal becomes progressively more stereotypic (Fig. 4B), the market decider, in turn, develops progressively more stereotypic expectations about the groups as agents are replaced, $B = 0.084$, $SE = 0.003$, 95% CI $= [0.078, 0.089]$, $z = 29.59$, Cohen's $d = 1.08$, even though the underlying differences are actually becoming smaller with each replacement (Figs. 2I and 4A).

Lastly, we find that upon replacing the market decider, the agents continue to engage in a pattern of stereotypic behavior. Due to the expectations of the previous market decider, and the convention that emerges as a result of this expectation, agents have learned that maximizing their rewards within the structure of the environment requires acting in accordance with the stereotypic convention; subsequently, the new market decider— learning to predict in an environment with no underlying group differences—nevertheless develops stereotypic expectations for the groups matching those of the previous market decider.
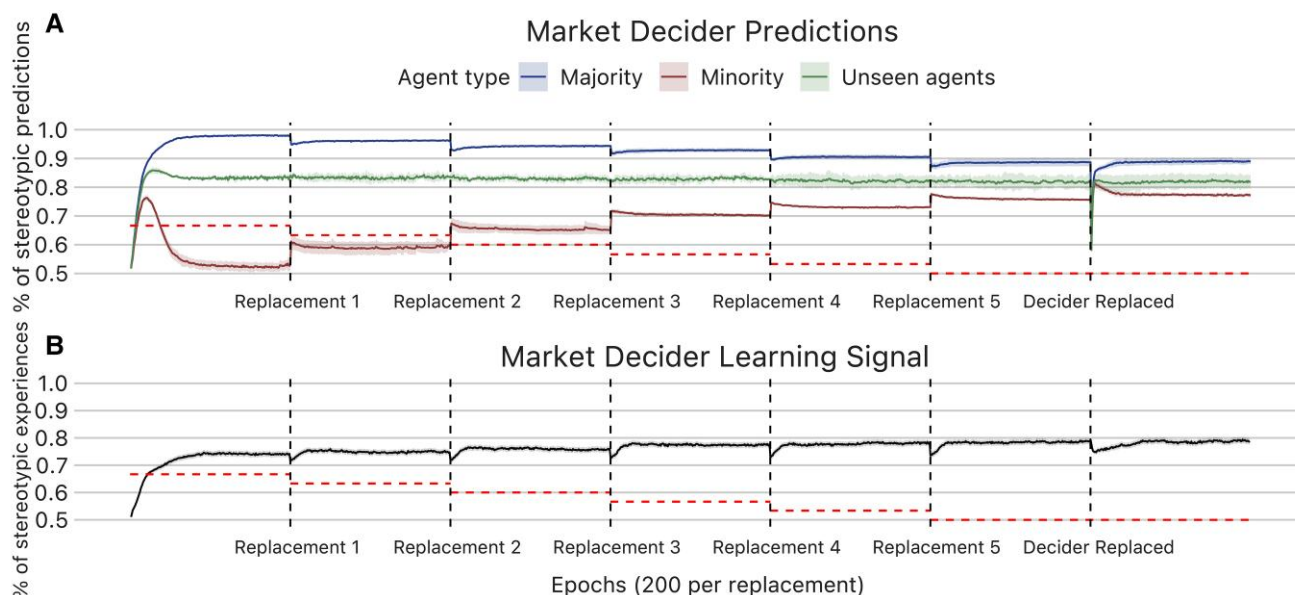
**Fig. 4.** Raw learning signal and market predictions for unobserved agents. A) Proportion of stereotypic predictions by the market decider for previously observed majority agents (blue), minority agents (red), and held-out unobserved agents (green) in study 3. The market stably predicts that unseen agents will mostly engage in stereotypic actions at a greater rate than the ground truth skill distribution (dashed red line), even as the underlying skill distribution changes. B) Market decider's raw learning signal for mining and chopping groups. As the market decider's expectations prompt minority agents to engage in against-skill behaviors, the decider's learning signal exaggerates the underlying difference in skill sets, and maintains it after the underlying difference disappears.

### Collective rewards.

As the market decider in study 1 learns to represent agents as individuals, the average collective reward obtained by agents increases. We find that in study 2, the collective reward obtained by agents once the agents and the market decider are fully trained is slightly smaller, but comparable to study 1 ($B = 0.13$, SE = 0.02, 95% CI = [0.08, 0.18], $z = 6.21$, $P < 0.001$, Cohen's $d = 0.35$). This effect is driven by the rewards obtained by the majority agents, who obtain a higher reward than comparable agents in study 1 ($B = 1.36$, SE = 0.055, 95% CI = [1.23, 1.49], $z = 24.79$, $P < 0.001$, Cohen's $d = 0.49$); meanwhile, minority agents obtain a lower reward ($B = -3.62$, SE = 0.055, 95% CI = $[-3.75, -3.49]$, $z = -66.04$, $P < 0.001$, Cohen's $d = -1.31$).

In study 3, the collective reward becomes lower as agents are replaced. By the time all agents and the market decider have been replaced and the market decider is fully trained, the collective reward obtained by the agents is significantly lower than in study 2 ($B = 1.39$, SE = 0.01, 95% CI = [1.37, 1.42], $z = 141.18$, $P < 0.001$, Cohen's $d = 3.65$; see Fig. S1).

## Study 4: Human trajectories

As we observed the potential for stereotypes to persist and entrench across generations of artificial agents with no motivations beyond maximizing their own potential returns, we assessed human performance on the same task. If human participants make predictions based on group-based expectations and match others' expectations through behavioral confirmation, we anticipate that these patterns of behavior will emerge on a task comparable to the simulated environments we developed. Importantly, we anticipate that these behaviors emerge due to the nature of achieving social coordination itself, rather than motivational biases; thus, we expect that the degree of stereotyping shown by participants in these studies will not be correlated to other potential factors associated with stereotyping, such as system justification (3, 4) or

social dominance orientation (5, 49). In other words, the degree to which individuals engage in stereotype-consistent reasoning or behavior in our tasks should not be correlated with the individuals' motivations to stereotype about social groups.

To test our predictions, we conducted a preregistered study in which 447 human participants completed a web experiment containing an adapted form of the environment from studies 1 to 3. In this task, participants were placed in the role of either an agent, collecting resources and attempting to sell them to the market decider, or a market decider, predicting the resources being sold by agents. Participants interacted with a market decider or agents whose behavior trajectories were collected from one run of study 3 and were randomly presented with a trajectory from the original agents or market, the agents or market after two replacements, or the agents and market after all five replacements. As most majority agents in study 3 were predicted to behave skill-consistently, and we wished to compare participants' behavior in the context of being stereotyped or successfully individuated, all participants participating as agents in the task played the role of minority agents. Participants had identical action spaces to the artificial agents, and participants' skillsets (wood or stone) were counterbalanced between participants. After completing the task, participants completed three social orientation scales: the System Justification Scale (SJS, adapted from Ref. (50)), the Social Dominance Orientation scale (SDO, (49)), and the Basic Social Justice Orientation scale (BSJO, (51)). Participants in the agent task additionally answered questions about their beliefs regarding the value of different resources, their learning of the task structure, and their performance relative to others. Participants in the market task answered questions about their beliefs regarding the percentage of agents in each group that were skilled at collecting different resources and the fairness of the world they encountered.

To assess the degree to which human participants demonstrated a similar response pattern to the artificial agents, we first tested human participants' actions on the agent task. As with the artificial agents, human participants assigned to agents for whom
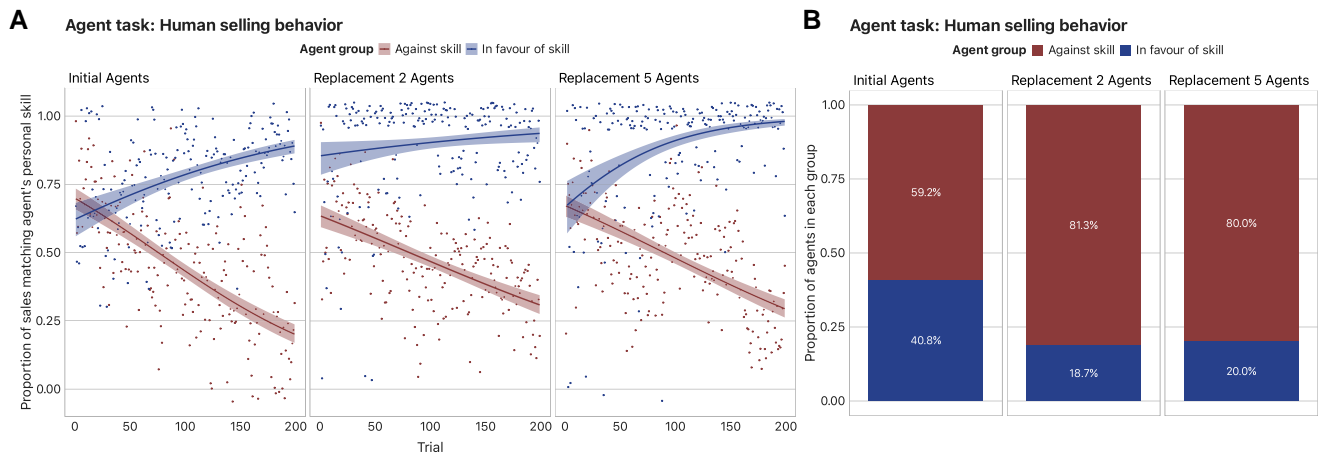
**Fig. 5.** Agent task with human participants. Participants were assigned to play the role of agents the produce-and-trade task in which the market predicted the agent would bring resources consistent with the agent's true skill (blue) or contrary to the participant's true skill (red). Participants predicted to behave skill-consistently by the market decider progressively made a higher proportion of skill-consistent sales over time, while this proportion declined for participants for which the market made mostly skill-inconsistent prediction (A). Further, the proportion of agents for which the market made skill-consistent predictions was higher for the initial agents than for the agents from replacement 2 and replacement 5 (B).



**Fig. 6.** Market task with human participants. Participants were assigned to play the role of the market decider in the produce-and-trade task and predict the resources brought by agents to the market. Across all three conditions, participants predicted that agents who formed the majority within their group would behave more skill-consistently than agents who formed the minority, and this pattern became stronger over time such that in later trials, participants made more stereotypic predictions about both the majority and the minority.

the market decider predicted would bring resources against the participants' true skill were less likely to bring resources they were skilled at collecting, $B = -2.41$, $SE = 0.171$, $z = -14.12$, 95% CI $= [-2.75, -2.08]$, $OR = 0.09$, $P < 0.001$, and became progressively less likely to do so in later trials, $B = -1.22$, $z = -15.84$, 95% CI $= [-1.37, -1.07]$, $P < 0.001$ (Fig. 5).

Likewise, human participants on the market decider task displayed a stereotypic effect consistent with the artificial agents. Participants predicted that agents would bring the same resource that a majority of the agents of the same color brought, making them more likely to make skill-consistent predictions for majority agents than for minority agents, $B = 1.97$, $SE = 0.03$, $z = -63.25$, 95% CI $= [1.91, 2.03]$, $OR = 7.17$, $P < 0.001$. This effect also became stronger in later trials, $B = 0.23$, $SE = 0.03$, $z = 7.47$, 95% CI $= [0.17, 0.29]$, $OR = 1.26$, $P < 0.001$ (Fig. 6).

On average, participants believed that most members of a group shared a single skill (e.g. agents with a purple body color are more likely to attempt to sell stone) corresponding to the most commonly observed action. When some or all of the agents had been replaced, they overestimated the proportion of the group that shared this skill, believing that the skill was more common in the group than it truly was (replacement 2: $M = 6.1\%$, $SE = 1.89$, $t(210) = 3.21$, $P = 0.002$; replacement 5: $M = 19.3\%$, $SE = 1.77$, $t(210) = 10.91$, $P < 0.001$). Further, beliefs about the degree to which groups of agents shared a single skill also predicted the degree to which participants made predictions consistent with that skill for both majority and minority agents. Participants who believed that the majority's skill was more common within a group were more likely to make skill-consistent predictions for majority agents, $B = 0.63$, $SE = 0.026$, $z = 24.46$, 95% CI $= [0.58, 0.69]$, $OR = 1.88$, $P < 0.001$ but were more likely to make skill-inconsistent predictions for minority agents, $B = -0.77$, $SE = 0.035$, $z = 21.98$, 95% CI $= [-0.84, -0.70]$, $OR = 0.46$, $P < 0.001$.

Participants also indicated that they believed the world was more fair than unfair for the aliens ($M = 3.33$, $SD = 0.78$, 1 = least fair, 5 = most fair), and 41% of participants indicated that the world

was very or somewhat fair, while only 13% indicated that the world was somewhat unfair. None indicated that the world was very unfair.

Participants' differences in SJS and SDO did not meaningfully predict the degree to which they engaged in stereotype-consistent decision-making (SJS: $b = -0.12$, 95% CI $= [-0.18, -0.06]$, 100% in ROPE; SDO: $b = 0.04$, 95% CI $= [-0.02, 0.10]$, 100% in ROPE). There was positive evidence that BSJO led participants to less strongly distinguish between the majority and minority agents ($b = 0.27$, 95% CI $= [0.20, 0.34]$, 0% in ROPE).

An analysis of the error rates indicated that being higher in BSJO did not lead participants to be more accurate about minority group members, but rather had higher error rates on majority group members, $B = -0.14$, SE $= 0.04$, 95% CI $= [-0.22, -0.06]$, $z = -3.32$, $P < 0.001$, and equivalent error rates on minority group members, $B = 0.01$, SE $= 0.04$, 95% CI $= [-0.07, 0.10]$, $z = 0.33$, $P = 0.74$. Although this may suggest that BSJO was associated with not using the agents' color to strategically undermine the structural inequality, this interpretation is less likely given that (i) higher BSJO participants thought that the world was more fair than other participants, $B = 0.22$, SE $= 0.05$, $t(211) = 4.27$, $P < 0.001$, $\eta_p^2 = 0.08$ and that (ii) individual differences in the degree to which these participants saw the game as fair or unfair did not moderate the effect of BSJO on error rates ($B = -0.24$, SE $= 1.40$, $t(209) = -0.17$, $P = 0.86$, $\eta_p^2 < 0.001$).

## Discussion

In this study, we demonstrate that agents rapidly regularize around group-based social conventions that facilitate a simple coordination strategy. This tendency was particularly strong in larger populations, in which learning the skills of each individual was more time-consuming and costly, initially providing a predictive advantage to a convention-based strategy. However, as this regularity became established, it became a learning signal in its own right, making it difficult to dislodge the convention for both agents and market deciders, as well as the human participants playing these roles.

Human participants behaved comparably to the predictions made by the models, aligning their actions with the market decider's expectations—even when these expectations involved collecting resources they were unskilled at. They also developed stereotypes based on appearance rather than individual skills. Critically, these role-based stereotypes were not used simply as a heuristic to maximize rewards, irrespective of the underlying structure of the world, but were seen as a reflection of the true state of the world (see e.g. (52)).

As participants considered the strong learning signal of the groups' distinct behavior to reflect real essential differences between groups, participants' choices were not affected by their attitudes towards real-world social inequality, that might otherwise have provided a motivation for individuals to behave stereotypically even though doing so could disadvantage minority members of a group. In fact, participants with higher scores in BSJO believed the structure of the world within our tasks was more fair than those with lower BSJO scores.

These findings suggest that role-based stereotyping may be particularly prone to emerging—and particularly resistant to change—in larger populations. In these contexts, one's experience with a single individual of a group is a small part of one's experience with the group as a whole. As a result, individuals within these societies may be more likely to perceive the differences they observe to reflect real underlying differences rather than an artifact of the demands of social coordination, especially when conformity to the expectations of a role is otherwise high. Modern social groupings, with sizes in the

millions and more, may be particularly vulnerable to the development and persistence of stereotypes, posing a unique threat to the potential for diversity and individuality within highly connected contemporary societies. Because these expectations dictate the structure of the subsequent coordination, expectations can amplify the associations between social roles and social categories and lead them to persist even if the factors that initially differentiated social categories disappear.

Role-based stereotypic conventions create a dilemma for agents whose skills do not match the expectation. If they continue to sell the resource they are skilled at collecting, they often fail to coordinate with the market decider or obtain any rewards. However, if they sell the resource they are unskilled at collecting, they reinforce the market decider's expectations. This dynamic compels agents to conform to the stereotype, suppressing their individual strengths.

The fact that human participants on our tasks made similar choices despite differing degrees of motivation to reject social inequality suggests that these humans' egalitarian attitudes and aversion to stereotyping were unable to prevent the emergence of similar degrees of unequal outcomes for the artificial agents, because most participants believed that the differences in the behavior they observed were reflective of a fair society in which groups are largely homogeneous and differ based on their skills, not simply because human beings were maximizing their own rewards.

Although the threat of stereotyping is often recognized as a social problem in that people identify and endorse moral norms against the use of stereotypes (36, 53), including across political ideologies (37), our results may shed light on why some forms of stereotyping, particularly role-based stereotypes, remain a ubiquitous feature of human societies. If behavioral confirmation can generate and perpetuate conventions that certain groups are more suited for certain roles or behaviors (e.g. (10)), then subsequent generations observe real-world confirmation of these stereotypic beliefs, making it seem as if the expectation was justified all along. Similarly, human participants on our tasks engaged in stereotypic behavior, but they did so because they believed that their beliefs about the agents were statistically accurate generalizations.

The fact that patterns of stereotypic behavior can remain in place even when the underlying group differences are no longer consistent with the equilibrium reached by the group-based convention suggests that role-based stereotypes may represent the "kernel" around which other kinds of stereotypes about social categories could form and accumulate. Thus, our findings should serve as a caution against succumbing to a "naturalistic fallacy" in stereotyping: the idea that observed differences between groups is due to an underlying or essential characteristic of the groups, rather than due to other explanations such as structural or historical factors. For example, the existence of the "gender-equality paradox"—the observation that gender segregation by occupation is stronger in gender-egalitarian countries—has sometimes been explained as revealing intrinsic differences between men's and women's interests when given a free choice to pursue a career (e.g. (54)). However, consistent with stereotype-based explanations for the gender-equality paradox (e.g. (55)), our data illustrate how beliefs about what is perceived to be a rational decision process driven by free choice, such as choosing one's career, could become biased by the development of these stereotypic norms.

The mechanisms by which perpetuation of stereotypic conventions can occur also provides us a method for understanding how inequalities persist. For example, although the coordination tasks we introduce in these studies are in principle

equitable with regards to the rewards provided to agents, in practice a proportion of agents were disadvantaged by the stereotype, because they were not able to live up to their full potential. Furthermore, many real-world instances of coordination result in divisions of labor that are unequal, particularly to structurally disadvantaged or numerically smaller groups (e.g. (29, 30)). Thus, the effects we observe in this work can be extended to understand when conformity to a stereotypic convention may emerge even when a group as a whole is disadvantaged by the convention, rather than disadvantaging a minority of agents within a group.

## Materials and methods

### Environment setup

The task for our agents $i \in I = \{1 \cdots N\}$ can be formalized as a partially observable Markov decision process (POMDP). Each agent $i$ has an observation size $v$, observation function $\mathcal{O}_i : \mathcal{S} \to \mathbb{R}^v$, and action set $\mathcal{A}_i$; for the producing agents, $v = 6$ and $\mathcal{A}_i = 7$, and for the market decider, $v = 16$ and $\mathcal{A}_i = 2$. State transitions are determined according to the transition probability $\mathcal{T}(\mathcal{S}' | \mathcal{S}, \mathcal{A})$, with each agent receiving rewards $r_i \in \mathcal{R}$ according to the reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.

At each timestep, the producing agents observe a six-digit vector indicating their own resource pool (quantity of coins, stone, and wood, and whether each respective quantity is above 0), and select an action from a list of seven possible actions (extract stone, extract wood, build, sell stone, sell wood, buy stone, buy wood). Simultaneously, the market decider observes the 16-digit binary vector representing the identity of the agent that attempts to sell stone or wood and selects one of two actions (predict the agent will sell stone or wood). Both the agents and the market decider must learn a policy mapping observed states to action probabilities based on the long-term expected discounted reward, $E_i[\sum_{t=0}^{\infty} \gamma^t r_{i,t}]$.

In all studies, the probability that certain actions (extracting stone, extracting wood, and extracting skill) will be successfully completed by agents is dependent on their skill level. Agents are instantiated with one of three specializations: chopping, mining, or building specialist. The choppers are good at chopping (0.75) and bad at mining (0.25) and building (0.05). The miners are good at mining (0.75) and bad at chopping (0.25) and building (0.05). The builders are good at building (0.95) and bad at chopping and mining (both 0.1; see also Figs. S6–S8 for results with other hyperparameters). Agents receive 15 points of reward for successfully building a house, receive 1 point of reward for successfully selling wood or stone, and lose 2 points of reward for buying wood or stone from the market decider.

## Agent architecture and training method

In all three studies, each agent and market decider comprised an independent neural network (including the agents serving as replacers in study 3). These networks were identical in architecture and were initialized with random weights at the start of each experimental condition. All models were structured using an actor-critic architecture (56). Each model consisted of an actor network producing an action policy and a critic network outputting a state value estimate. Both networks consisted of a three-layer fully connected multilayer perceptrons (MLP) (Actor and Critic) with 64 hyperbolic tangent units (tanh) in the first and third layers and 128 units in the second layer. The critic network outputs a scalar representing the estimated value of the current state. The actor network produces a probability over the actions to be chosen (seven actions for the agents and two for the market).

What the agents and the market decider observe shares the same structure across the three studies. At each step within an epoch, each agent observed the state of its current possessions, represented by a six-digit vector: the quantity of stones, the quantity of wood, the quantity of coins, whether the number of wood is >1, whether the number of stones is >1, and whether the number of coins is larger than 1. If any agent decides to sell resources to the market, the market observes a 16-digit binary code representing the identity of that agent. After transformations of the input observations, the agent's and the market's neural networks generate an action policy and an estimate of the current state. The observed state, the action policy, and the reward outcome of each time step are stored for training the neural networks. Each agent and market decider have a separate rollout buffer to store the experience trajectories. Specifically, we train the neural networks using the proximal policy optimization algorithm (PPO, (57)), with an Adam optimizer (58) (see Figs. S9–S11 for results with other algorithms). The learning rates are $10^{-3}$ and $5 \times 10^{-4}$ for the actor network and the critic network respectively, and the discount factor $\gamma$ is 0.9. At the end of each epoch, all agents and the market are trained with their own stored experiences. In detail, the stored trajectory of experiences in the current epoch are retrieved from the memory buffer, and the agents/market maximize the following objective:

$$L_t^{\text{CLIP+VF+S}}(\theta) = \hat{\mathbb{E}}_t[L_t^{\text{CLIP}}(\theta) - c_1 L_t^{\text{VF}}(\theta) + c_2 S[\pi_\theta](s_t)], \quad (1)$$

$$L_t^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)], \quad (2)$$

where $c_1$, $c_2$ are coefficients, and $S$ denotes an entropy bonus, and $L_t^{\text{VF}}$ is a squared-error loss $(V_\theta(s_t) - V_t^{\text{targ}})^2$. $V_\theta(s_t)$ denotes the generated estimate of the value of the current state, and $V_t^{\text{targ}}$ denotes the actual value of the current state. $\hat{A}_t$ denotes an estimator of the advantage function at timestep $t$. The term $r_t(\theta)$ denotes the probability ratio between the current stochastic policy and the old stochastic policy with which an agent collected the experience to learn from. The function clip() establishes a bound for the probability ratio term $r_t(\theta)$ within the interval $[1 - \epsilon, 1 + \epsilon]$. In our study, $c_1 = 0.5$, $c_2 = 0.01$, $\epsilon = 0.2$.

## Experimental design: human task

### Participants and design

Four hundred and forty seven adults (Sex: Male = 281, Female = 165, Not provided = 1. Age: $M = 38.34$, $SD = 12.65$, Range = [18, 74]. Nationality: United Kingdom = 184, United States = 117, Canada = 71, Australia = 33, Nigeria = 9, New Zealand = 8, South Africa = 4, Ireland = 4, India = 3, Poland = 2, Others = 12. Ethnicity: White = 321, Asian = 53, Black = 33, Mixed = 29, Other = 7, Not provided = 5) were recruited from the online platform Prolific. To be eligible, participants had to be native speakers of English and have an acceptance rate of at least 99% on the platform to qualify. Participants were paid £2.25 for completing the task, and a bonus of £0.25 for scoring in the top 10% of participants. This study was approved by the University of Toronto Research Ethics Board (#33582).

Participants were randomly assigned to complete either the market decider task ($n = 226$) or the agent task ($n = 221$) and were also randomly assigned to complete the task at either the early timepoint, before any agents were replaced ($n = 149$), at the middle time point, after 40% of the original agents were replaced

($n = 144$), or at the late time point, after 100% of the original agents were replaced ($n = 154$). As these conditions were completed between subjects, no participant performed both tasks or encountered the task at more than one time point.

According to our preregistered inclusion criteria, participants were excluded from analysis if their responses were multivariate outliers on Mahalanobis distance with an $\alpha$ of 0.001. Ten participants' responses were excluded from the agent task, and 13 were excluded from the market task, leaving a total of 424 final participants included in the analysis.

### Materials and procedure

After reading and agreeing to the consent form, participants completed either the market game or the agent game in a ReactJS web app.

In both tasks, the market behavior (predicting wood or stone for each agent approach) and the agent behavior (performing an action such as collecting, buying, or selling a resource or building a house) was drawn from one run of study 3. These behaviors were transformed into action probabilities for the market or agents that the human participants encountered in study 4.

#### Market game.

Participants playing the market game were first presented with instructions on how to play the game. They were told that they were going to encounter a number of aliens, whose identity could be determined by their body color and unique 12-pixel nametag. Participants then needed to predict the resources that each alien would bring to the market. They were told that successfully predicting an alien approach would gain them points, and that the goal in the task was to maximize the number of points obtained.

Then, participants completed 150 total trials of the game. On each trial, one alien would be shown and the participant selected whether they thought the alien brought wood or stone to the market. The individual alien that was shown on a given trial was determined according to the approach probability (how often each agent came to the market) and the resource probability (how often each agent brought wood or stone to the market) of agents in one run of study 3. The agent body color (purple, yellow, or cyan) was determined by the first three digits of the agents' 16-digit code identifier, and each agent was assigned a unique 12-digit identifier in place of the remaining 13 digits. As there were a total of 300 agents, not all identifiers were used.

Thus, agents who approached the market more often were more likely to be shown to a participant, and agents with a higher wood probability were more likely to bring wood than stone to the market.

After completing the task, participants were presented with a short survey. They were asked about their beliefs about the group's general skill at collecting wood or stone, the percentage of each group that was better at collecting wood or stone, and their beliefs about the fairness of the task, as well as three social orientation scales: the System Justification Scale (SJS), the Social Dominance Orientation scale (SDO), and the Basic Social Justice Orientation scale (BSJO).

#### Agent game.

Participants playing the agent game were first presented with instructions on how to play the game. They were told that they were a worker trying to obtain coins, which could be obtained by selling resources or building houses, and that each worker has a skill that makes them better at collecting different kinds of resources. Further, they were told that in order to successfully sell a resource, they needed to bring the same resource that the market decider predicted they would bring.

Participants' skill was randomly counterbalanced between being skilled at collecting wood or stone. No participants were skilled at building houses. Participants' chance of successfully collecting a resource depended on their skill, with a 75% chance to succeed at collecting their skilled resource and a 25% chance to succeed at collecting their unskilled resource. All participants had a 5% chance of successfully building a house.

At the beginning of the task, participants were assigned an agent identity code that was not visible to the participant. The market decider's predictions for the participant were computed for each epoch (from 1 to 200) as the average proportion of predictions that the agent would bring the resource they were skilled at collecting. Thus, the markets' prediction for agents on each trial would vary from 0 (always predict against the agents' true skill) to 1 (always predict consistently with the agents' true skill).

Participants completed 200 trials of the task, corresponding to 1 epoch of the task. Similar to the agents in studies 1–3, participants could attempt to collect wood, collect stone, buy wood, buy stone, sell wood, sell stone, or build a house. Participants received 1 coin for successfully selling a resource, and 15 coins for successfully building a house.

If participants attempted to sell to the market on a given trial, the market would make a prediction for the agent based on the probability that the agent would bring their skilled resource in the corresponding epoch.

After completing the task, participants were presented with a short survey. They were asked whether they believed the market needed more wood or stone, whether wood or stone was more valuable to own, how well they learned the structure of their environment, how well they believed they did relative to other players, and how frustrating and enjoyable they found the task. They were also presented with three social orientation scales (SJS, SDO, BSJO).

## Statistical analyses

Prior to analysis, the following variables of interest were collected or computed for studies 1–3. For the market, we used the proportion of the market's predictions that were consistent with the agent's skills for a given epoch and run. For each agent, we computed the proportion that the agent attempted to sell the resource it was skilled at collecting relative to the total number of sales attempted (skilled and unskilled), as well as the proportion of skilled extractions relative to the total number of extractions (skilled and unskilled) for a given epoch and run.

For the three simulation studies, we used linear mixed-effects regression (LMER) models which predict the degree of skill-based predictions by the market, or the degree of skilled extraction or skilled sales by agents. For all models, we first compute an omnibus ANOVA and compute $F$-test and partial $\eta^2$ values for the predictors of interest, and then compute follow-up pairwise contrasts adjusted using Tukey's method for multiple comparisons. The main analyses in studies 1–3 used linear mixed-effects models with the ratio of extracting, selling, or market prediction predicted by the epoch, agent group (majority, minority, or study 1 nonstereotyped baseline), population size (studies 1 and 2), and replacement (study 3).

Although population size is a numerical quantity, we analyze it as a factor as the values chosen are four discrete quantities and may display nonlinear effects at different population sizes. The values for epoch are scaled and normalized such that the zero point is placed at epoch 100. Note that in study 2, the group variable included the data from study 1 as a separate group, in order to

allow us to compare data from the majority and minority in study 2 to a nonstereotyped baseline.

Additionally, we recorded the average reward obtained per group (study 1: all agents, studies 2 and 3: mining and chopping specialist majorities and minorities) throughout the three studies. We then compute the population-level average reward obtained per run by the total population. These outcomes were compared using LMER models, with follow-up pairwise contrasts and effect sizes computed between groups (study 1 agents, study 2 majority, and study 2 minority) and between study populations as a whole (study 1, study 2, and study 3).

In study 4, we use a similar pipeline for assessing the data as in study 3. Due to a data collection error, actions were not recorded on the final trial for 100 out of the 221 participants in the agent task. The first 199 trials were not affected for any participants. Consistent with our preregistration, we used Mahalanobis distance to identify multivariate outliers with $\alpha < 0.001$, indicating careless or repetitive responses on the task or survey. Eleven participants were excluded from the agent task, and 13 participants were excluded from the market task, yielding a final sample of 210 participants on the agent task and 213 participants on the market task.

As participants performed one action per trial instead of multiple actions in one epoch (as in the simulation studies), we use a generalized linear mixed-effects regression (GLMER) with a binary outcome to analyze participants' actions, with effect sizes measured as odds ratios (OR) and follow-up pairwise contrasts computed for measures of interest. For the agent task, we focus on whether the participant's extracting or selling action was the skilled (1) or unskilled (0) resource. For the market task, predictions are either skill-consistent (1) or skill-inconsistent (0). For SJS, SDO, and BSJO scores, we compute and normalize participants' full scale scores before analysis. In study 4, we test the following analysis, where scale represents SJS, SDO, or BSJO:

$$\text{Ratio} = 1 + \text{epoch}*\text{group} + \text{replacement} \\ + \text{Scale} + \text{group: Scale} + (1 \mid \text{Participant}). \quad (3)$$

For our preregistered predictions of a null effect on SJS, SDO, and BSJO, we used a Bayesian GLMER using a region of practical equivalence (ROPE) test, using a prior of $b \sim \mathcal{N}(0, 5)$ on all parameters. For our models, we follow Kruschke (59) and interpret a negligible effect as $0.1 \times \text{SD}_y$, where the standard difference of a parameter expressed in the log odds scale is $\pi/\sqrt{3}$, yielding a ROPE of $[-0.18, 0.18]$.

Additional details and full analysis code are available in the Supplementary Material.

## Acknowledgments

The authors wish to thank Gillian Hadfield, Raphael Köster, Joel Z. Leibo, Paul Stillman, and the members of the Social AI Lab at the University of Toronto for their helpful feedback.

## Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

## Funding

## Author Contributions

Conceptualization: R.A.G., Y.T., E.C.J., and W.A.C; Data curation: R.A.G. and Y.T.; Software: R.A.G., Y.T., and E.C.J.; Formal analysis: R.A.G. and Y.T.; Validation: R.A.G.; Investigation: R.A.G. and Y.T.; Visualization: R.A.G.; Supervision: W.A.C.; Methodology: R.A.G., Y.T., E.C.J., and W.A.C.; Project administration: W.A.C.; Writing–original draft: R.A.G.; Writing–review and editing: Y.T., E.C.J., and W.A.C.

## Data Availability

Data for experiments 1–4 and study preregistrations are available at https://osf.io/zus7d/. The code used for the multiagent environment, agent architectures, and model training is available at: github.com/Yikai369.

## References

1 Brewer MB. 1999. The psychology of prejudice: ingroup love and outgroup hate? *J Soc Issues.* 55(3):429–444.

2 Tajfel H, Turner J. An integrative theory of intergroup conflict. In: Austin WG, Worchel S, editors. *The social psychology of intergroup relations* Brooks/Cole, Monterey, CA, 1979. p. 33–47.

3 Jost JT, Banaji MR. 1994. The role of stereotyping in system-justification and the production of false consciousness. *Br J Soc Psychol.* 33(1):1–27.

4 Jost JT, Banaji MR, Nosek BA. 2004. A decade of system justification theory: accumulated evidence of conscious and unconscious bolstering of the status quo. *Polit Psychol.* 25(6):881–919.

5 Sidanius J, Pratto F. *Social dominance: an intergroup theory of social hierarchy and oppression* Cambridge University Press, 2001.

6 Fiske ST, Neuberg SL. A continuum of impression formation, from category-based to individuating processes: influences of information and motivation on attention and interpretation. In: Zanna MP, editor. *Advances in experimental social psychology*. Vol. 23 Academic Press, 1990. p. 1–74.

7 Macrae CN, Milne AB, Bodenhausen GV. 1994. Stereotypes as energy-saving devices: a peek inside the cognitive toolbox. *J Pers Soc Psychol.* 66(1):37–47.

8 Sherman JW, Lee AY, Bessenoff GR, Frost LA. 1998. Stereotype efficiency reconsidered: encoding flexibility under cognitive load. *J Pers Soc Psychol.* 75(3):589.

9 Bruner JS. 1957. Going beyond the information given. *Contemp Approaches Cogn.* 1(1):119–160.

10 Eagly AH, Koenig AM. 2021. The vicious cycle linking stereotypes and social roles. *Curr Dir Psychol Sci.* 30(4):343–350.

11 Eagly AH, Wood W. Social role theory. In Van Lange PAM, Kruglanski AW, Higgins ET, editors. *Handbook of theories of social psychology*. Vol. 2. Sage, 2012. p. 458–476.

12 O'Connor C. *The origins of unfairness: social categories and cultural evolution* Oxford University Press, USA, 2019.

13 Binmore K. *Natural justice* Oxford University Press, 2005.

14 Henrich J, Boyd R. 2008. Division of labor, economic specialization, and the evolution of social stratification. *Curr Anthropol.* 49(4):715–724.

15 Sloman S, Fernbach P. *The knowledge illusion: why we never think alone* Penguin, 2018.

16 Dunbar RIM. 1998. The social brain hypothesis. *Evol Anthropol Issues News Rev.* 6(5):178–190.

17 Carron PM, Kaski K, Dunbar R. 2016. Calling Dunbar's numbers. *Soc Networks.* 47(1):151–155.

18  Tamarit I, Cuesta JA, Dunbar RIM, Sánchez A. 2018. Cognitive resource allocation determines the organization of personal networks. *Proc Natl Acad Sci U S A*. 115(33):8316–8321.

19  Tamarit I, Sánchez A, Cuesta A. 2022. Beyond Dunbar circles: a continuous description of social relationships and resource allocation. *Sci Rep*. 12(1):2287.

20  Wheeler NE, *et al.* 2020. Ideology and predictive processing: coordination, bias, and polarization in socially constrained error minimization. *Curr Opin Behav Sci*. 34:192–198.

21  Hawkins RD, *et al.* 2023. From partners to populations: a hierarchical Bayesian account of coordination and convention. *Psychol Rev*. 130(4):977–1016. https://doi.org/10.1037/rev0000348.

22  Burke MA, Young HP. Social norms. In: Benhabib J, Bisin A, Jackson MO, editors. *Handbook of social economics*. Vol. 1. Elsevier, 2011. p. 311–338.

23  Fehr E, Fischbacher U. 2004. Third-party punishment and social norms. *Evol Human Behav*. 25(2):63–87.

24  Köster R, *et al.* 2022. Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents. *Proc Natl Acad Sci U S A*. 119(3):e2106028118.

25  Vinitsky E, *et al.* 2023. A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings. *Collective Intelligence*. 2(2). https://doi.org/10.1177/26339137231162025.

26  Luce RD, Raiffa H. *Games and decisions: introduction and critical survey* Wiley, 1957.

27  Osborne MJ, Rubinstein A. *A course in game theory* MIT Press, 1994.

28  Hadfield GK. 1999. A coordination model of the sexual division of labor. *J Econ Behav Organ*. 40(2):125–153.

29  Bruner JP. 2019. Minority (dis)advantage in population games. *Synthese*. 196(1):413–427.

30  O'Connor C, Bright LK, Bruner JP. 2019. The emergence of intersectional disadvantage. *Soc Epistemol*. 33(1):23–41.

31  Klein O, Snyder M. Stereotypes and behavioral confirmation: from interpersonal to intergroup perspectives. In: Zanna MP, editor. *Advances in experimental social psychology*. Vol. 35. San Diego (CA): Elsevier Academic Press, 2003. p. 153–234.

32  Madon S, Willard J, Guyll M, Scherr KC. 2011. Self-fulfilling prophecies: mechanisms, power, and links to social problems: self-fulfilling prophecies. *Soc Personal Psychol Compass*. 5(8):578–590.

33  Skrypnek BJ, Snyder M. 1982. On the self-perpetuating nature of stereotypes about women and men. *J Exp Soc Psychol*. 18(3):277–291.

34  Allidina S, Cunningham WA. 2021. Avoidance begets avoidance: a computational account of negative stereotype persistence. *J Exp Psychol Gen*. 150(10):2078–2099.

35  Bai X, Fiske ST, Griffiths TL. 2022. Globally inaccurate stereotypes can result from locally adaptive exploration. *Psychol Sci*. 33(5):671–684.

36  Cao J, Kleiman-Weiner M, Banaji MR. 2019. People make the same Bayesian judgment they criticize in others. *Psychol Sci*. 30(1):20–31.

37  Vrantsidis TH, Cunningham WA. Your stereotypes, my beliefs about groups: belief differences underlie different views of the morality of group-based generalizations, under review.

38  Chen Y, Liu J. Social capital games as a framework for social structural pattern emergence. In: *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. New York (NY): Association for Computing Machinery, 2020. p. 309–316.

39  Jaques N, *et al.* Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In: *Proceedings of the 39th International Conference on Machine Learning*, PMLR 97. PMLR, 2019. p. 3040–3049.

40  Leibo JZ, Zambaldi V, Lanctot M, Marecki J, Graepel T. Multi-agent reinforcement learning in sequential social dilemmas. In: *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2017. p. 464–473.

41  Lerer A, Peysakhovich A. Learning existing social conventions via observationally augmented self-play. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Honolulu HI USA, 2019 Jan. p. 107–114.

42  Zhang H, *et al.* 2020. Bi-level actor-critic for multi-agent coordination. *Proc AAAI Conf Artif Intell*. 34(05):7325–7332.

43  Köster R, *et al.* 2020. Model-free conventions in multi-agent reinforcement learning with heterogeneous preferences. arXiv, arXiv:2010.09054. https://doi.org/10.48550/arXiv.2010.09054, preprint: not peer reviewed.

44  Fehr E, Fischbacher U. 2003. The nature of human altruism. *Nature*. 425(6960):785–791.

45  Gintis H, Bowles S, Boyd R, Fehr E. 2003. Explaining altruistic behavior in humans. *Evol Human Behav*. 24(3):153–172.

46  Henrich J, *et al.* 2001. In search of homo economicus: behavioral experiments in 15 small-scale societies. *Am Econ Rev*. 91(2):73–78.

47  Tang Y, Gelpí RA, Cunningham WA. Unequal norms emerge under coordination uncertainty in multi-agent deep reinforcement learning. In: *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*. Cognitive Science Society, 2023. p. 555–561.

48  Zheng S, *et al.* 2020. The AI economist: improving equality and productivity with AI-driven tax policies. arXiv, arXiv:2004.13332. https://doi.org/10.48550/arXiv.2004.13332, preprint: not peer reviewed.

49  Pratto F, Sidanius J, Stallworth LM, Malle BF. 1994. Social dominance orientation: a personality variable predicting social and political attitudes. *J Pers Soc Psychol*. 67(4):741.

50  Vargas-Salfate S, Paez D, Khan SS, Liu JH, de Zúñiga HG. 2018. System justification enhances well-being: a longitudinal analysis of the palliative function of system justification in 18 countries. *Br J Soc Psychol*. 57(3):567–590.

51  Hülle S, Liebig S, May MJ. 2018. Measuring attitudes toward distributive justice: the basic social justice orientations scale. *Soc Indic Res*. 136(2):663–692.

52  Bai X, Griffiths TL, Fiske ST. 2025. Costly exploration produces stereotypes with dimensions of warmth and competence. *J Exp Psychol Gen*. 154(2):347–357.

53  Mae L, Carlston DE. 2005. Hoist on your own petard: when prejudiced remarks are recognized and backfire on speakers. *J Exp Soc Psychol*. 41(3):240–255.

54  Stoet G, Geary DC. 2018. The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychol Sci*. 29(4):581–593.

55  Breda T, Jouini E, Napp C, Thebault G. 2020. Gender stereotypes can explain the gender-equality paradox. *Proc Natl Acad Sci U S A*. 117(49):31063–31069.

56  Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: *Proceedings of the 35th International Conference on Machine Learning*. PMLRM, 2018. p. 1861–1870.

57  Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. 2017. Proximal policy optimization algorithms. arXiv, arXiv:1707.06347. https://doi.org/10.48550/arXiv.1707.06347, preprint: not peer reviewed.

58  Kingma DP, Ba J. Adam: a method for stochastic optimization. In: Bengio Y, LeCun Y, editors. *3rd International Conference on Learning Representations*. ICLR, 2015.

59  Kruschke JK. 2018. Rejecting or accepting parameter values in Bayesian estimation. *Adv Methods Pract Psychol Sci*. 1(2):270–280.