Unequal Norms Emerge Under Coordination Uncertainty in Multi-Agent Deep Reinforcement Learning

Yikai Tang*, **Rebekah A. Gelpí***, **and William A. Cunningham**
Department of Psychology, University of Toronto
100 St. George Street, Toronto, Ontario, Canada M5S3G3
{yikai.tang,rebekah.gelpi}@mail.utoronto.ca,
wil.cunningham@utoronto.ca

Abstract

Successful social coordination requires being able to predict how the other people that one depends on are likely to behave. One solution to this dilemma is to establish social conventions, which constrain individuals' behavior but make prediction easier. Here, we develop a multi-agent deep reinforcement learning environment to investigate the costs associated with these conventions. In our produce-and-trade task, agents have varying production skills, but their actions must be predictable in order to be rewarded. Stronger norms improve the overall success of the group by improving the average rewards of the majority, but also systematically disadvantage agents whose specialization is in the minority of the group. Critically, this outcome is magnified by population size: as larger groups make it potentially more difficult to develop individualized representations of agents, minority agents become more likely to conform to a norm that is disadvantageous to them.

Keywords: reinforcement learning; agent-based model; social norms; coordination

Introduction

Modern humans live in very large groups, requiring us to engage in cooperation and coordination with each other to survive and succeed. Successful coordination requires two or more individuals to act in a systematically correlated or anticorrelated (i.e., complementary) fashion (O'Connor, 2019). This task can be easy if the intentions, goals, and abilities of the person whom we are going to interact with are clear. For example, when we are working with a close friend or family member, we can use our knowledge about them to easily generate an accurate prediction of their behavior in a given interaction. However, this information may not be available when we need to coordinate with an unfamiliar stranger. Because individuated knowledge about the goals and abilities of a stranger is difficult to acquire within a short time, we may struggle to predict their behavior in an interaction accurately. This problem may be exacerbated further in larger social groups, as we may be faced with cognitive limitations on our ability to manage individual social relationships beyond a certain point (Dunbar, 1998). As a result, coordination with novel partners is a challenging computational problem that we may regularly encounter in large social groups.

How can people efficiently overcome the action prediction challenge to achieve successful social coordination with uncertain and unfamiliar others? One potential solution is establishing a social structure or a convention in a population (Hadfield-Menell et al., 2019; Köster et al., 2022; Lewis, 1969; Vinitsky et al., 2021). Conventions constrain the space of actions that members of a population can perform, reducing variance and inducing regularity. With more regular behavioral patterns, the actions of a group of individuals become less uncertain, and prediction becomes more accurate and less demanding. A large body of research has emphasized that people use observed statistical regularities to resolve uncertainty about the world's causal and category structure (e.g., Griffiths et al., 2010; Tenenbaum et al., 2011), including within our social world (Constant et al., 2019; FeldmanHall & Shenhav, 2019; Fiske & Neuberg, 1990; Fleischhut et al., 2022).

Because these regularities can facilitate the ease of coordinating with potential social partners by making them more predictable (Wheeler et al., 2020), people might find it advantageous to use regularities—what one does—to derive prescriptive rules—what one ought to do. As young as 4 years old, children who observe patterns in a social environment use these to infer what kinds of behavior is normative (Roberts, Gelman, et al., 2017; Roberts, Ho, et al., 2017) and consider social categories to come with obligations to behave a certain way (Chalik & Rhodes, 2020; Foster-Hanson & Rhodes, 2019; Rhodes & Chalik, 2013). By punishing nonconformity with respect to a perceived regularity, the magnitude of the regularity can be exacerbated, creating the possibility for a positive feedback loop that strengthens a prescriptive norm (Burke & Young, 2011).

At the collective level, a group of individuals who show low variance and strong regularities should allow for more efficient coordination and lead to the easier attainment of social goals (Wheeler et al., 2020). When predictability and coordination are easily achievable, social interactions are likely to lead to dependable outcomes; their outputs, in turn, can be relied upon as inputs for other, potentially more complex social goals. As a result, people in such a group could cooperate with each other to produce more benefits within the same length of time than if they were restricted only to outputs that they could produce as an individual. In the same way that the presence of plentiful "silly rules" might help an agent learn about the social sanctions that come about from not following important rules (e.g., Hadfield-Menell et al., 2019; Köster et al., 2022), they could also help an agent learn that following a norm provides the benefits of predictability

*Authors contributed equally to this work.

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

and reliability, both of their own actions to others and others' actions to themselves.

However, the enforcement of social regularity is not without costs. Social norms, by definition, result in discouragements of non-conforming behaviors, systematically disadvantaging members of a social group who do not match the norm (e.g.,, Heilman, 2012). This can result in situations where a majority of individuals in a group behaving in the same way create an observable regularity, resulting in an established norm to behave in this more "predictable" way. The opportunity to improve the group's coordination by establishing the majority's behavior as a norm comes with unfairness – the majority groups win an advantage while the minority groups suffer a disadvantage (Jost et al., 2015; Sidanius & Pratto, 1999). For example, studies with bargaining models have shown that when social categories are in play, conventions may emerge with the concurrent phenomenon of inequality, even though all artificial agents in the models are programmed to maximize their own benefits (Amadae & Watts, 2022; Bruner, 2019; O'Connor, 2019). Bruner's (2019) and O'Connor's (2019) simulations also indicate that the relative size of minority groups plays a role in the emergence of minority-disadvantaged conventions.

Using a multi-agent reinforcement learning approach, we explored the emergence of conventions under social interaction uncertainty and its concurrent outcomes. We placed agents in a coordination task, where successful coordination brought more rewards than solo work. Agents could collect resources and consume them to build houses. They could also trade resources with a market to make more earnings. Agents belonged to different social groups defined by the type of identity cue used in trading, and they varied in their skills to collect different types of resources. We hypothesized that the demand for predictability in the trade process would regularize agents' behaviors and give rise to conventions. The emergence of conventions would come with a benefit at the collective level, but also came at the cost of inequality, particularly in environments with large populations as well as those where agents formed a numerical minority within a larger group.

Method

Environment Description

The present reinforcement learning (RL) environment was adapted from the AI economist paradigm (Zheng et al., 2020). A population of agents played a produce-and-trade task in the environment alongside a market decider, collecting resources and trading them with the market decider in exchange for coins. In this way, the market decider served as an abstraction for a typical coordination problem, where individuals need to accurately predict the actions of others about whom they know little. In each step, each agent could take one action in turn. The action space of the agents consisted of seven actions in both asocial and social domains. The asocial actions were chopping wood (receive 1 wood if successful), mining stone (receive 1 stone if successful), and building houses (receive 1 house and consume 1 wood and 1 stone if successful). Whether the actions were successful depended on the skill levels of the agents, which were set as fixed probabilities of succeeding at the action. Agents were created with one of three types of skillsets: choppers, miners, and builders. The choppers were good at chopping (0.95) and bad at mining (0.15) and building (0.05). The miners were good at mining (0.95) and bad at mining (0.15) and building (0.05). The builders were good at building (0.95) and bad at chopping and mining (both 0.1).

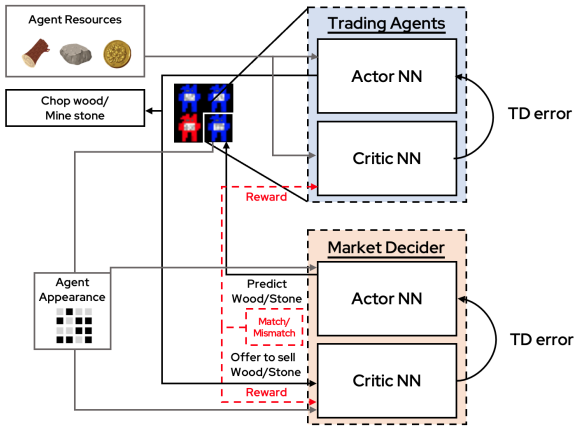

Figure 1: Environmental setup for the agents in our experimental task. Both the market decider (light orange) and the producing-and-trading agents (light blue) have their actions governed by a proximal policy optimization (PPO) actor-critic neural network architecture.

Social actions included selling wood, selling stone, buying wood, and buying stone. Successful selling actions required coordination between the agents and the market decider. When an agent decided to sell a resource, the market decider needed to guess what items the agent will sell (stone or wood). The transaction was completed only if the market decider correctly guessed the resource that the agent was trying to sell. After a successful transaction, the agent traded away 2 wood or stone units and received 1 coin. When an agent wanted to buy resources, the market decider did not need to make a prediction. Every successful purchase made by the agent cost the agent 2 coins.

Agents in RL tasks are typically programmed to maximize their rewards during learning. In the present simulated environment, agents were positively or negatively rewarded in the following three situations:

1. Received 15 points for successfully building a house
2. Received 1 point for successfully selling stone or wood to the market decider
3. Lost 2 points of reward for making a purchase

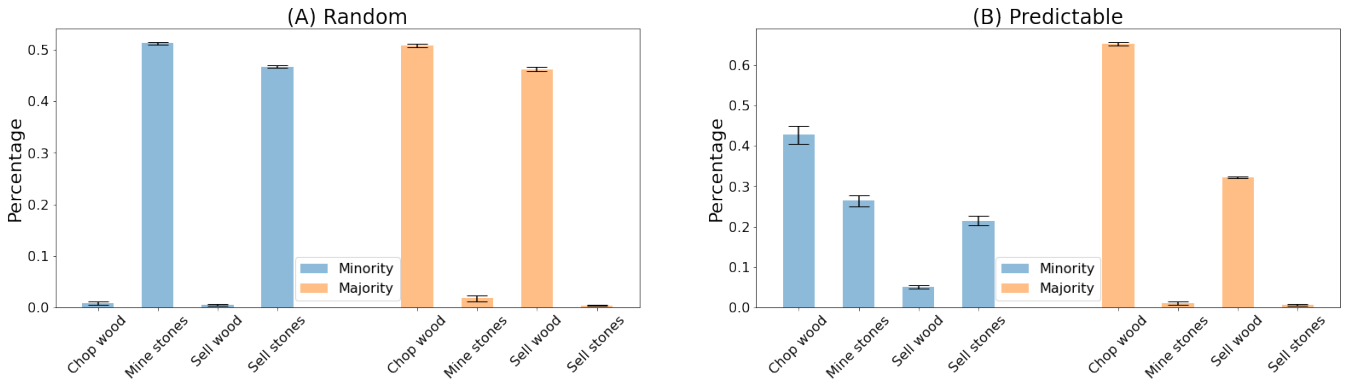

Figure 2: Distribution of behaviors by minority members (blue) and majority members (orange) for both the random decider (A) and predicting decider (B) conditions. When the decider's actions were random, agents focused exclusively on performing actions that favored their skill (e.g. here, minority agents specialized in producing stone, while majority agents specialized in producing wood). In the predicting condition, however, minority agents were much more likely to engage in actions that they were unskilled in. The frequency of this phenomenon varied across the ratio of minority to majority agents, as well as the population size (see also Figure 4).

The market decider received rewards or punishments through the selling-and-predicting process. It received 1 point when an agent successfully sold an item to it and lost 1 point when it failed to make an accurate prediction and 0.3 points when the prediction was correct, but the agent did not have enough possessions to make the transaction. The value of coins matches the value of rewards/punishments, and we set the coin system for the ease of collecting data of reward values.

The agents were placed in three groups of equal sizes: the Wood group, the Stone group, and the House group. Each group was named after the skill that the majority of agents in the group had, but a minority of members within the group had a different specialization. For example, choppers were a majority within the Wood group, while miners were the minority. The House group had a fixed proportion of choppers (0.1), miners (0.1), and builders (0.8) across all simulations. We included the House group to maintain a similar environmental structure to Zheng et al. (2020). In a similar vein, we designed a fixed proportion of builders (0.1) in both the Wood and the Stone groups. We focused on the chopping and the mining behaviors of agents to measure the emergence of normativity against the best interests of the agent. As a result, in this study, we focus our analysis on the behaviors and the collective rewards of the choppers and miners in the Wood and the Stone group, whose dominant skill types were chopping and mining.

In each trade, the market decider was presented with a 16-digit binary vector representing each unique agent. To represent group belonging, 3 of the digits were a one-hot code representing the agent's membership in a group (a group code associated with looking similar to agents that are better at chopping Wood, mining Stone, or building Houses). In addition to the group code, each agent also had a unique 8-digit binary identifier. The group membership codes were concatenated with the individual-unique codes, together with a 5-zero code that represented the agents' common category as "agents", to form a 16-digit identity code. During trades, the market decider used the identity codes to make predictions about what the agents wanted to sell.

Agent Architectures

We implemented nine neural network architectures, each of which was shared between members of a social group specialized for a particular skill. These networks were identical in architecture and were initialized with random weights at the start of each simulation. The market decider's policy was determined by a neural network with the same architecture. All models were structured using a soft actor-critic architecture (Haarnoja et al., 2018). The architecture consisted of two 3-layer fully connected multilevel perceptrons (MLP) (Actor and Critic) with 64 hyperbolic tangent units (`tanh`) in the first and third layer and 128 units in the second layer. The Critic model output an estimate of the value of the current observed state. The Actor model generated the action for the agents and the decider.

We trained the agents using the proximal policy optimization algorithm (PPO, Schulman et al., 2017), with an Adam optimizer (Kingma & Ba, 2015). The learning rates were 10^{-3} and 5×10^{-4} for the Actor network and the Critic network respectively. We used a discount factor of 0.9. The agents/decider maximized the following objective at the end of each timestep:

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

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

where c_1, c_2 are coefficients, and S denotes an entropy bonus, and L_t^{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^{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 $[1-\varepsilon, 1+\varepsilon]$. In our study, $c_1 = 0.5$, $c_2 = 0.01$, $\varepsilon = 0.2$.

On each timestep, agents observed the state of their current possessions, represented by a 6-digit vector: amount of stone, amount of wood, number of coins, whether the amount of wood is larger than 1, whether the amount of stone is larger than 1, and whether the number of coins is larger than 1. The agents' neural networks used this observed state to choose their actions.

Unlike the agents, the market decider did not necessarily act every timestep. It only took action when an agent wanted to sell something to it. The decider's observation space consisted of the identity code of the agent interacting with it; subsequently, the decider could choose one of two actions: guess that the agent is selling stone, or that the agent is selling wood.

Experimental Setup

Each instance of the simulation regime of agents playing the produce-and-trade task lasted for 50 turns. We manipulated the predictability of the market decider, the relative size of the majority members, and the total number of agents.

The behavior of the market decider differed depending on two conditions: random or predicting. In the random condition, the market decider made a random guess every time it interacted with an agent who intended to sell either wood or stone. In the predicting condition, the market decider made guesses about the agent's intentions based on what it had learned and observed.

The size of the majority (i.e., choppers in the Wood group) relative to the minority (i.e., miners in the Wood group) varied by condition; the ratio of the two groups was either 90:10, 80:20, 70:30, 60:40, or 50:50 (no majority). Finally, the total population of agents in the task was either 64, 128, or 256. For each combination of conditions, we conducted 30 simulations, and in total there were 2 (predicting vs. random) × 5 (majority size) × 3 (population size) × 30 = 900 simulations. We offered the builders in all groups with 6 coins every time a simulation started.

Outcome variables. The design of the social group structure was aimed to provide room for the emergence of group-based conventions. Generally, we considered a social behavioral pattern as evidence of convention emergence when the agents in the same group focused on the same type of behaviors (chopping or mining), irrespective of their skill. Specifically, we put our emphasis on the minority agents, and we operationalized the level of regularity as the difference between the number of steps doing the unskilled action and the skilled action. Taking the minority members of the Wood group as an example, they were skilled at mining, and we measured the difference between the proportion of time spent on mining and chopping. Because the Wood and the Stone group are completely symmetrical in terms of their composition, we collapsed our analysis across the Wood and Stone groups, analyzing the behavior of the majority and minority agents within both groups.

In addition to the regularity in the patterns of coordination-related behaviors, we were also interested in the concurrent outcomes, including the rewards earned both by individual majority/minority agents as well as at the collective level. We computed the average collective reward by calculating the weighted mean of the gained rewards by chopping and mining specialist agents in the Wood and the Stone group (i.e., not including the building specialists). As our pilot simulations showed that learning stabilized for agents by the time they had completed 100 steps, we used the last 10 steps of each simulation to calculate mean rewards and action frequency values and then pooled them to compute statistics.

Results and Discussion

When the market decider was allowed to act on its predictions, normative behaviors emerged. As a result, minority agents acted against their own skills when the decider was predicting the agents' choices. We found this pattern emerge across simulations where the ratio of the agents' skills in each group was not equal (i.e., a true majority and minority group existed): when the market was predicting, minority members were more likely to take actions that they were unskilled at ($M = 0.06$, $SD = 0.68$), which were the actions the majority of their group was skilled at, than when the market was random ($M = -0.51$, $SD = 0.11$), $t(1798) = -24.77$, $p < .001$. Further, this phenomenon was not because the minority members attempted to collect both sets of resources to build houses instead; when the market was predicting, minority members were more likely to attempt to sell unskilled resources (4.4% of actions) than to attempt to build houses (1.7% of actions), suggesting that the reward obtained from building houses was not the primary goal of minority agents conforming to the group norm.

For example, the minority agents in the Wood group, whose strength was mining stone, focused on what they were good at when the market was random, but worked against their skills (i.e., spent a large proportion of time on chopping wood) when the market was predicting. This stood in sharp contrast to their behavior when the market did not make predictions, where minority agents almost exclusively acted in accordance with their skillset. In contrast, the majority agents in both conditions adhered to their strengths (Figure 2). The frequency of wood sales did not increase in proportion to the increase in wood chopping because the minority agents had a

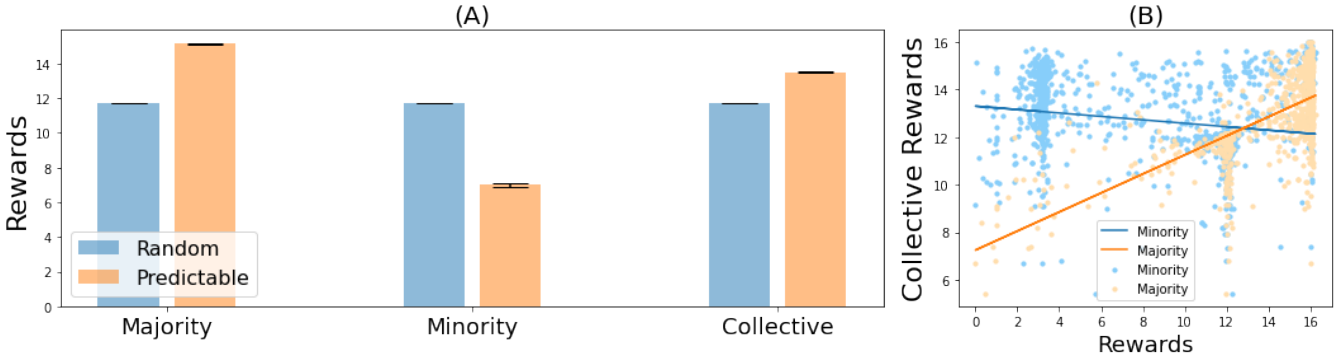

Figure 3: (A): Average rewards obtained by the majority, minority, and by collective groups. When the decider was random (blue), rewards were obtained equally by majority and minority group members. The predicting decider (orange) resulted in higher rewards for the majority agents and lower rewards for the minority agents, resulting in a greater average collective reward. (B) Majority rewards were positively, and minority rewards were negatively, associated with collective rewards.

low success rate for chopping wood (0.1). As a result, even though these agents attempted to collect wood on a plurality of timesteps, they still had very little to sell in the market.

Successful group-level coordination led to greater collective benefits, but at the expense of inequality. The average collective rewards—that is, the weighted average of minority and majority group members—was significantly higher when the market was predicting compared to when it was not, $t(1798) = 19.44,\ p < .001$ (Figure 3A). However, the benefits for the whole group sacrificed the interests of the minority members. They gained significantly fewer rewards in the predicting condition than what they received in the random condition, $t(1798) = 19.44,\ p < .001$, even though their objective was to maximize their own benefits.

The majority members achieved more when the market was predicting than when it was random, $t(1798) = -27.79, p < .001$. In Figure 3B, we plot the relationship between majority/minority rewards and the collective reward obtained by the entire group. There was a negative relationship between minority rewards and collective rewards ($b = -0.07,\ SE = 0.01,\ p < .001$), but a positive relationship between majority rewards and collective rewards ($b = 0.40,\ SE = 0.01,\ p < .001$).

Norms favoring the majority's skillset were most pronounced in the largest groups. Consistent with the hypothesis that group-based conventionality is more necessary in groups where strangers must interact and there is difficulty individuating all members of the group, we found that minority group members were more likely to conform to stereotyped group-based behavior (working against their skill and rather performing actions expected from visual group membership) as population size increased. Specifically, when controlling for the ratio between the number of majority and minority agents and the interaction between the ratio and population size, a larger population size strengthened the minority agents' preference for doing the group-mainstream work against their own skills, $b = 0.07,\ SE = 0.01,\ p < .001$. These effects were more pronounced as the relative size of the minority group decreased with respect to the total group size (see Figure 4).

General Discussion

In the current simulation study, we have demonstrated that normativity can arise in RL agents' behaviors when they are placed in a context where successful coordination brings more rewards than engaging in asocial behaviors, but demands that agents behave predictably during interactions in order to obtain the rewards. As a result, agents who were negatively impacted by these norms nevertheless conformed to them, in order to maximize their rewards by seeming predictable to an interacting agent.

These results provide a computational account for social psychological theories of the emergence and maintenance of unjust norms, even by those who are disadvantaged by their continued existence (Jost et al., 2004; Jost et al., 2008; Sidanius & Pratto, 1999). They also align with a growing body of work using agent-based models to understand norm emergence through the use of punishment (Gavrilets & Richerson, 2017; Köster et al., 2022; Vinitsky et al., 2021; Yaman et al., 2022). In this study, we show that agents do not need to incur explicit punishment in order for normative pressure to shift their behaviours; the sustained disadvantage of defying an expectation that one will behave another way results in agents conforming to a personally less rewarding action policy.

These findings are consistent with theories of social cognition within Bayesian and predictive processing frameworks (Clark, 2013; FeldmanHall & Shenhav, 2019; Koster-Hale & Saxe, 2013; Wheeler et al., 2020). Effectively minimizing prediction error in an uncertain social world is much easier when one's social partners behave consistently; however, demands that group members behave consistently in social interactions are likely to have inegalitarian social consequences for heterogeneous groups. This is particularly true as groups become larger—making it more cognitively demanding to

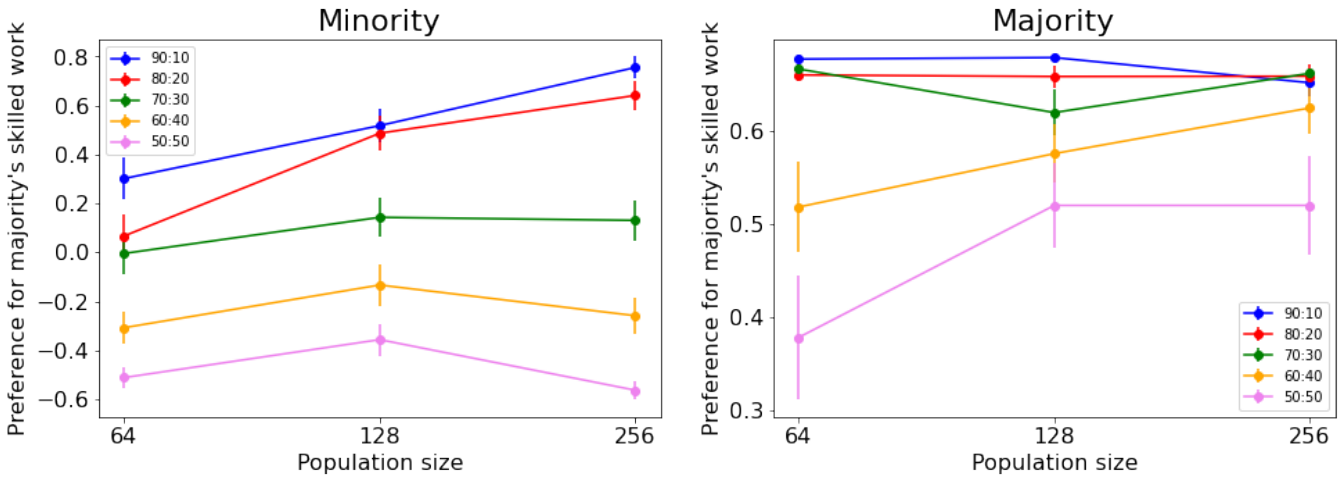

Figure 4: Summary of behaviors performed by Minority (left) and Majority (right) groups in the Predicting Decider condition, across population sizes and group proportions. Minority agents were more likely to engage in the majority's skilled work in larger groups, when their own proportion was smallest. Majority agents were always more likely to engage in their own skilled work.

process group members as individuals—as well as when minority members are a smaller proportion of the group, or when group members are simply not attended to due to existing negative impressions or stereotypes (e.g., Allidina & Cunningham, 2021; Fazio et al., 2004).

In future research, we hope to consider more complex interactions between the emergence of norms and the skills and rewards that give rise to them. Agents in this simulation had a fixed skill, and were in principle able to produce resources of equal value, even if it was more difficult for agents working against their own skill level to produce the resources. However, humans exist within communities in which people are not merely born with skills but learn to specialize deeply, distributing physical and cognitive labor (Fernbach & Light, 2020; O'Connor, 2019; Saunders, 2022). Further, the rewards that are available to members of a minority group can themselves become unequal (Bruner, 2019), potentially exacerbating existing inequalities.

Thus, by allowing agents' skills to vary and modifying reward structures, we plan to illustrate other potential consequences of norm emergence on disadvantaged groups. For example, such groups might be pressured into taking less rewarding actions because an advantaged group has already fully accessed the limited resources that offer this reward. More ecologically complex simulations will further formalize how such unequal arrangements of resources can develop or be perpetuated, and potentially provide insight into how to rectify them as well.

Acknowledgments

This work was partially funded by grants from the Natural Sciences and Engineering Research Council of Canada [RGPIN-2018-05946] and the Social Sciences and Humanities Research Council of Canada [SSHRC-506547] to W.A.C. and a CGS-D (Canada Graduate Scholarship – Doctoral) Fellowship to R.A.G.

References

Allidina, S., & Cunningham, W. A. (2021). Avoidance begets avoidance: A computational account of negative stereotype persistence. *Journal of Experimental Psychology: General*, *150*(10), 2078.

Amadae, S. M., & Watts, C. J. (2022). Red Queen and Red King Effects in cultural agent-based modeling: Hawk Dove Binary and Systemic Discrimination. *The Journal of Mathematical Sociology*, 1–28.

Bruner, J. P. (2019). Minority (dis)advantage in population games. *Synthese*, *196*(1), 413–427.

Burke, M. A., & Young, H. P. (2011). Social norms. In *Handbook of social economics* (pp. 311–338). Elsevier.

Chalik, L., & Rhodes, M. (2020). Groups as moral boundaries: A developmental perspective. *Advances in child development and behavior*, *58*, 63–93.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181–204.

Constant, A., Ramstead, M. J. D., Veissière, S. P. L., & Friston, K. (2019). Regimes of Expectations: An Active Inference Model of Social Conformity and Human Decision Making. *Frontiers in Psychology*, *10*.

Dunbar, R. I. M. (1998). The Social Brain Hypothesis.

Fazio, R. H., Eiser, J. R., & Shook, N. J. (2004). Attitude formation through exploration: Valence asymmetries. *Journal of personality and social psychology*, *87*(3), 293.

FeldmanHall, O., & Shenhav, A. (2019). Resolving uncertainty in a social world. *Nature Human Behaviour*, *3*(5), 426–435.

Fernbach, P. M., & Light, N. (2020). Knowledge is shared. *Psychological Inquiry*, *31*(1), 26–28.

Fiske, S. T., & Neuberg, S. L. (1990). A Continuum of Impression Formation, from Category-Based to Individuating Processes: Influences of Information and Motivation on Attention and Interpretation. In *Advances in Experimental Social Psychology* (pp. 1–74). Elsevier.

Fleischhut, N., Artinger, F. M., Olschewski, S., & Hertwig, R. (2022). Not all uncertainty is treated equally: Information search under social and nonsocial uncertainty. *Journal of Behavioral Decision Making*, *35*(2).

Foster-Hanson, E., & Rhodes, M. (2019). Normative Social Role Concepts in Early Childhood. *Cognitive Science*, *43*(8).

Gavrilets, S., & Richerson, P. J. (2017). Collective action and the evolution of social norm internalization. *Proceedings of the National Academy of Sciences*, *114*(23), 6068–6073.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364.

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor [ISSN: 2640-3498]. *Proceedings of the 35th International Conference on Machine Learning*, 1861–1870.

Hadfield-Menell, D., Andrus, M., & Hadfield, G. (2019). Legible normativity for ai alignment: The value of silly rules. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 115–121.

Heilman, M. E. (2012). Gender stereotypes and workplace bias. *Research in Organizational Behavior*, *32*, 113–135.

Jost, J. T., Banaji, M. R., & Nosek, B. A. (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political psychology*, *25*(6), 881–919.

Jost, J. T., Ledgerwood, A., & Hardin, C. D. (2008). Shared reality, system justification, and the relational basis of ideological beliefs. *Social and Personality Psychology Compass*, *2*(1), 171–186.

Jost, J. T., Sterling, J. L., & Langer, M. (2015). From "is" to "ought" and sometimes "not" compliance with and resistance to social norms from a system justification perspective. *Journal of Cross-Cultural Psychology*, *46*(10), 1287–1291.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio & Y. LeCun (Eds.), *3rd International Conference on Learning Representations, ICLR 2015*.

Köster, R., Hadfield-Menell, D., Everett, R., Weidinger, L., Hadfield, G. K., & Leibo, J. Z. (2022). Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents. *Proceedings of the National Academy of Sciences*, *119*(3), e2106028118.

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: A neural prediction problem. *Neuron*, *79*(5), 836–848.

Lewis, D. K. (1969). *Convention: A Philosophical Study* (Vol. 20) [Issue: 80 Pages: 286]. Wiley-Blackwell.

O'Connor, C. (2019). *The Origins of Unfairness: Social Categories and Cultural Evolution*. Oxford University Press.

Rhodes, M., & Chalik, L. (2013). Social Categories as Markers of Intrinsic Interpersonal Obligations. *Psychological Science*, *24*(6), 999–1006.

Roberts, S. O., Gelman, S. A., & Ho, A. K. (2017). So It Is, So It Shall Be: Group Regularities License Children's Prescriptive Judgments. *Cognitive Science*, *41*, 576–600.

Roberts, S. O., Ho, A. K., & Gelman, S. A. (2017). Group presence, category labels, and generic statements influence children to treat descriptive group regularities as prescriptive. *Journal of Experimental Child Psychology*, *158*, 19–31.

Saunders, D. (2022). How to put the cart behind the horse in the cultural evolution of gender. *Philosophy of the Social Sciences*, *52*(1-2), 81–102.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal Policy Optimization Algorithms.

Sidanius, J., & Pratto, F. (1999). *Social Dominance: An intergroup theory of social hierarchy and oppression*. Cambridge University Press.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, *331*(6022), 1279–1285.

Vinitsky, E., Köster, R., Agapiou, J. P., Duéñez-Guzmán, E., Vezhnevets, A. S., & Leibo, J. Z. (2021). A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings. *arXiv preprint arXiv:2106.09012*.

Wheeler, N. E., Allidina, S., Long, E. U., Schneider, S. P., Haas, I. J., & Cunningham, W. A. (2020). Ideology and predictive processing: Coordination, bias, and polarization in socially constrained error minimization. *Current Opinion in Behavioral Sciences*, *34*, 192–198.

Yaman, A., Leibo, J. Z., Iacca, G., & Lee, S. W. (2022). The emergence of division of labor through decentralized social sanctioning. *arXiv preprint arXiv:2208.05568*.

Zheng, S., Trott, A., Srinivasa, S., Naik, N., Gruesbeck, M., Parkes, D. C., & Socher, R. (2020). The AI Economist: Improving Equality and Productivity with AI-Driven Tax Policies.