Probing the Effect of Diversity in Agent Decision-Making Preferences on Group Performance via Computer Simulation

Xirui Zhao*

2021-05-23

Abstract

Some individuals are more impulsive and more intolerant to delay. Research has associated these preferences to worse life outcomes. But can a group of individuals provide any mechanism to mitigate or exploit the diversity in decision-making preferences? Here we use a simple agent-based model to examine the effect of diversity in learning rate and exploratory tendency and the role of communication.

Introduction

Individuals are diverse in their decision-making preferences/strategies. Some important differences include time-preferences, risk-preferences, learning rate, and the balance between exploration and exploitation. It is fairly well-established that a certain set of preferences are optimal (in the sense of utility-maximizing) for a single individual (or agent) foraging

*Supervising faculty: Jeffrey Erlich

in a given environment (Schoener (1987)). In humans, it is often written that individuals have pathological preferences. That is, people with very high risk-tolerance may be at risk for drug addiction; people with very low risk-tolerance end up missing out on important financial opportunities. However, humans evolved in groups, so we are interested in how diversity of preferences (including preferences that may seem pathological at the individual level) can contribute to group success. First, I will give some background on each of the preferences we set out to examine in this work and the general framework of agent-based modeling of reinforcement learning agents with diverse preferences.

Delay discounting means that compared with some present reward, people are indifferent only towards some larger future reward; in that sense, future rewards are discounted. Interpersonal disparities in delay discounting are reliable across time-horizons and verbal versus experiential tasks (Lukinova et al. (2019)) and stable over the life course (Keidel et al. (2021)). High level of delay discounting is moderately correlated with lower income, lower intelligence, and lower connectivity strength of inhibitory corticostriatal tracts (Keidel et al. (2021)). Haushofer and Fehr (2014) contends that poverty's effect on negative affect and stress may alter individuals' risk-taking and delay-discounting preferences and that this causal path can explain the persistence of poverty. Kable and Glimcher (2007) showed that the subjective value (the subject's actual evaluation of the external good) is associated with neural activity in the ventral striatum, medial prefrontal cortex and posterior cingulate cortex; it was also shown that delay discounting correlates negatively with future-oriented cognitive styles. Despite these findings, the exact cognitive and neural bases of delay discounting remain unclear.

Research on risk preferences shows that general risk tolerance is moderately heritable and slightly genetically correlated with household income and that genes near correlated SNPs are associated with genes for glutamate and GABA neurotransmitters (23and Me Research Team et al. (2019)).

One common feature about these studies is that they focus on the biological and cognitive causes and behavioral effects of individual decision-making preferences. This research attempts to illustrate the conditions under which within-group diversity in decision-making preferences (and other parameters) may provide a competitive advantage against more homogeneous groups.

Previously, Hong and Page (2004) showed that a functionally diverse group outperforms a group of less diverse agents with higher overall ability; they also proved a mathematical theorem to that effect. In their framework, the goal is to maximize a real-valued function of solutions; the agents are diverse in their perspectives (their internal representations of solutions), and heuristics (their search strategies in the solution space). However, the desirableness of diversity is built into the model: different heuristics can search only a subset of the solutions. The interpretation is that a group of best-performing agents from a fixed pool of agents are necessarily more homogeneous and therefore less likely find the best solutions. Furthermore, as the authors acknowledged, their model didn't consider learning and communication. Their problem-solving agents are endowed with fixed perspectives and heuristics and use a simple iterative algorithm to find the best solutions: for example, for heuristic (1, 7, 11) and starting point $x = x_0$, the agent repeatedly increment x by $1, 7$ or, 11 in any order until $f(x)$ is not increasing. To accommodate learning, our model is based on the n-armed bandit problem, a classical reinforcement learning problem. Reinforcement learning models have been used to explain animal behavior (Luksys, Gerstner, and Sandi (2009)). In the original n-armed bandit problem, an agent has to choose from n options at each decision step. For each option, the reward is drawn from a fixed probability distribution. The goal is to maximize the total reward over some decision steps. The agent has to learn the value of each option and, in the meantime, balance exploitation (choosing the option with the best estimate) and exploration (sampling from other options). Our model built on the n-armed bandit problem by investigating a group of agents and allowing agents within a group to communicate about their experiences.

The classical hypothesis linking reinforcement learning and neuroscience is the reward prediction error hypothesis of dopamine neuron activity (Sutton and Barto (2018)). Experiments have confirmed that dopamine neurons fire more than baseline after an unexpected reward and less than baseline after an undelivered yet expected reward (Schultz, Dayan, and Montague (n.d.)). In reinforcement learning algorithms, a policy π maps states to probabilities actions; a value function $v_\pi(s_t)$ maps a state s_t under a policy to the expected return. In the simplest temporal difference (TD) method, $v(s_t) \leftarrow v(s_t) + \alpha[R_{t+1} + \gamma v(s_{t+1}) - v(s_t)]$. The TD error $\delta_t := R_{t+1} + \gamma v(s_{t+1}) - v(s_t)$ (R_{t+1} is the next reward). Intuitively, it measures the difference between current estimate of $v(s_t)$ and the better estimate $R_{t+1} + \gamma v(s_{t+1})$ where the value of future states are considered and discounted by γ. Sutton and Barto (2018) explained the striking correspondence between TD error and dopamine neuron activity: during initial learning, similar to the dopamine neuron response to unexpected reward, the reward signal is zero until the rewarding state, at which TD error becomes $R*$; after learning, similar to the response to delivered expected reward (fires more just after the conditioned stimulus), the TD error is positive at the earlier predictive state but zero afterwards; when the expected reward failed to deliver, the TD error will be negative at that time. Unfortunately the TD error is not modeled in our simulation.

Methods

Agent-Based Modeling (ABM) models a problem as a group of interacting agents. It can implement any complex interaction of individual agents, investigate the evolution of the model and inspect its state at any instant. ABM can better represent non-linearity and heterogeneity of interactions than systems of different equations (Bankes (2002)) and is suited to study emergence (Macy and Willer (2002)). For these reasons, we use Agents.jl (Datseris, Vahdati, and DuBois (2021), https://github.com/JuliaDynamics/Agents.jl), an ABM library to model foragers and food source patches.

In our model, foragers seek to gather food. The forager scales the reward by a utility exponent $U = reward^{\rho}$, which accounts for risk tolerance. Foragers with higher ρ are more risk tolerant because exponential functions are concave for ρ less than 1. Each forager has an internal representation of the subjective value of all patches, $Q[i]$ for patch i, which is updated via a simplified Q-learning algorithm where the environment state is ignored: $Q[i] \mathrel{+}= \alpha(U - Q[i])$, a weighted average that exponentially forgets prior experiences. Higher learning rate α results in quicker learning. A forager can also tell others the reward it received at each step. At each decision step, a forager selects from the patches with probabilities based on $Q[i]$. First, $Q[i]$ is divided by the maximum across all patches. Then, a softmax function is applied to transform them to probabilities $P(\text{select patch } i) = \frac{e^{\beta Q'[i]}}{\sum_i e^{\beta Q'[i]}}$. Higher softmax temperature β causes more exploitation and less exploration because the difference between $e^{\beta Q'[i]}$s are larger. The initial maximum-normalization of $Q[i]$ is required to separate the exploitation-v-exploration dimension from risk tolerance. The reward from the patch is drawn from a normal distribution $reward \sim Normal(\mu_{\text{rew}}, \sigma_{\text{rew}})$.

The foragers are diverse in their learning rates, their softmax temperatures and their utility exponents. In a group of foragers, each of these parameters sample from a log-normal distributions to ensure a proper range. If $X \sim LogNormal(\mu, \sigma^2)$, $log(X) \sim Normal(\mu, \sigma^2)$. Note the mean and variance of a log-normal distributions are $\exp(\mu + \frac{\sigma^2}{2})$ and $(\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)$. The patches' μ_{rew} are drawn from a Poisson distribution, and σ_{rew} is a constant percentage of μ_{rew}. To model a dynamic environment, the μ_{rew} of a patch may be redrawn with probability *shock_prob* at each step.

The source code, including the random number generator seed and the Julia Pluto notebook for generating figures, is available at https://github.com/xiruizhao/foraging-abm

Results

The following figures are generated from a predetermined seed (see the Julia notebook).

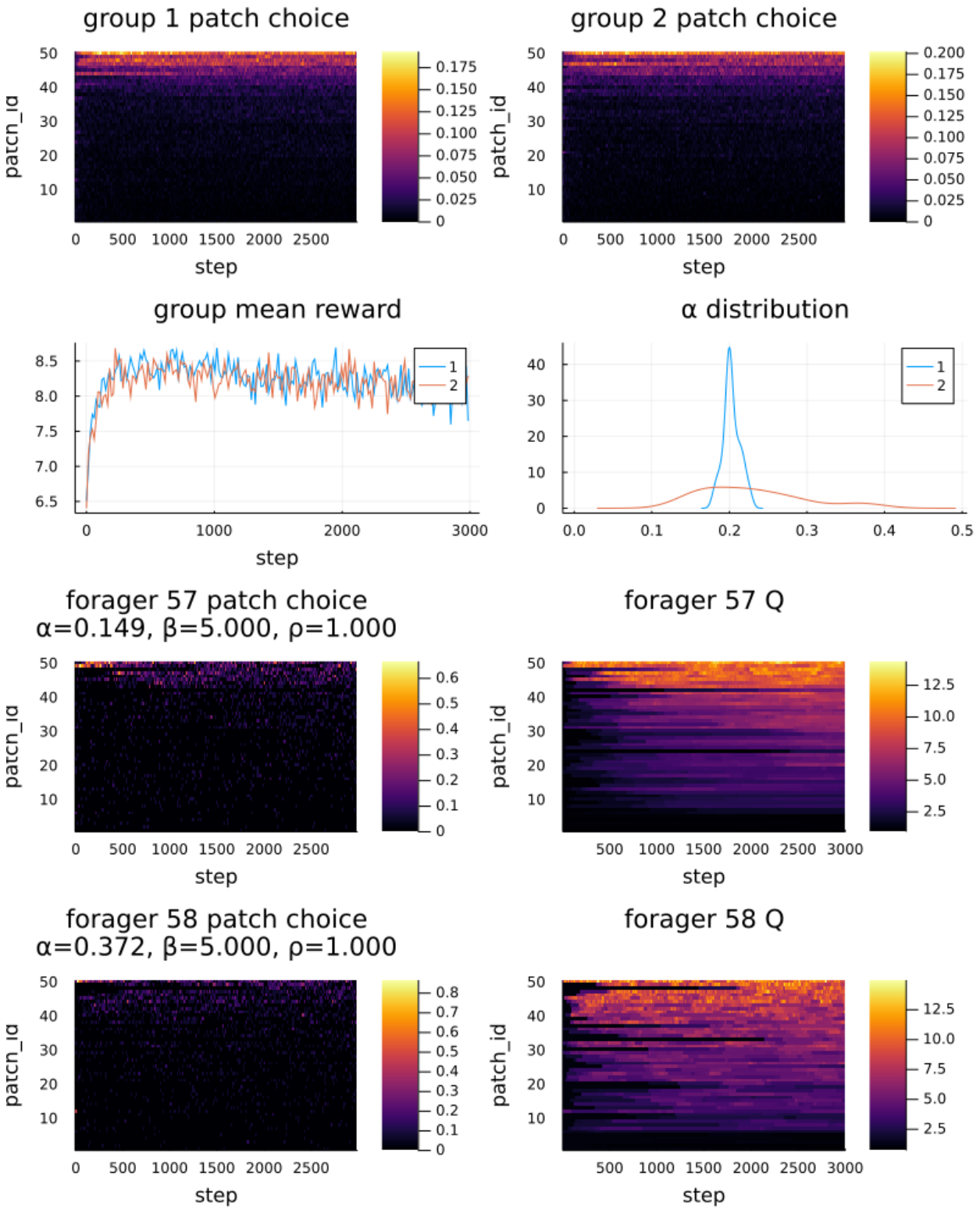
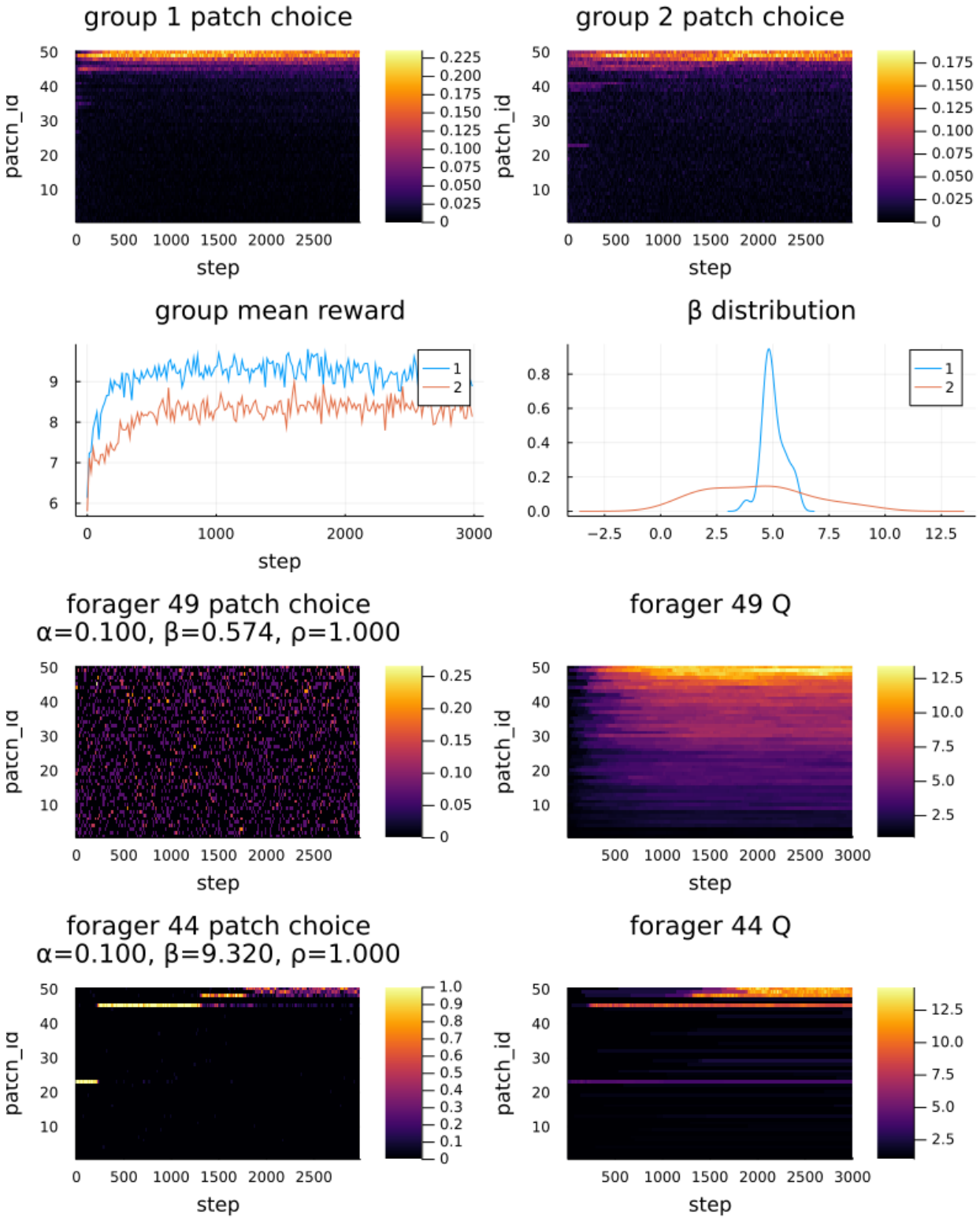


Figure 1: Heatmaps show the frequency of choosing a patch during a window of some steps. Patches: $\mu_{\text{rew}} \sim Poisson(5), \sigma_{\text{rew}} = 0.3 * \mu_{\text{rew}}$. Foragers: $\alpha \sim LogNormal, \beta = 5, \rho = 1$. Group 2 is more diverse in learning rate α. Forager 57 and 58 are, respectively, the slowest and fastest learner in group 2.

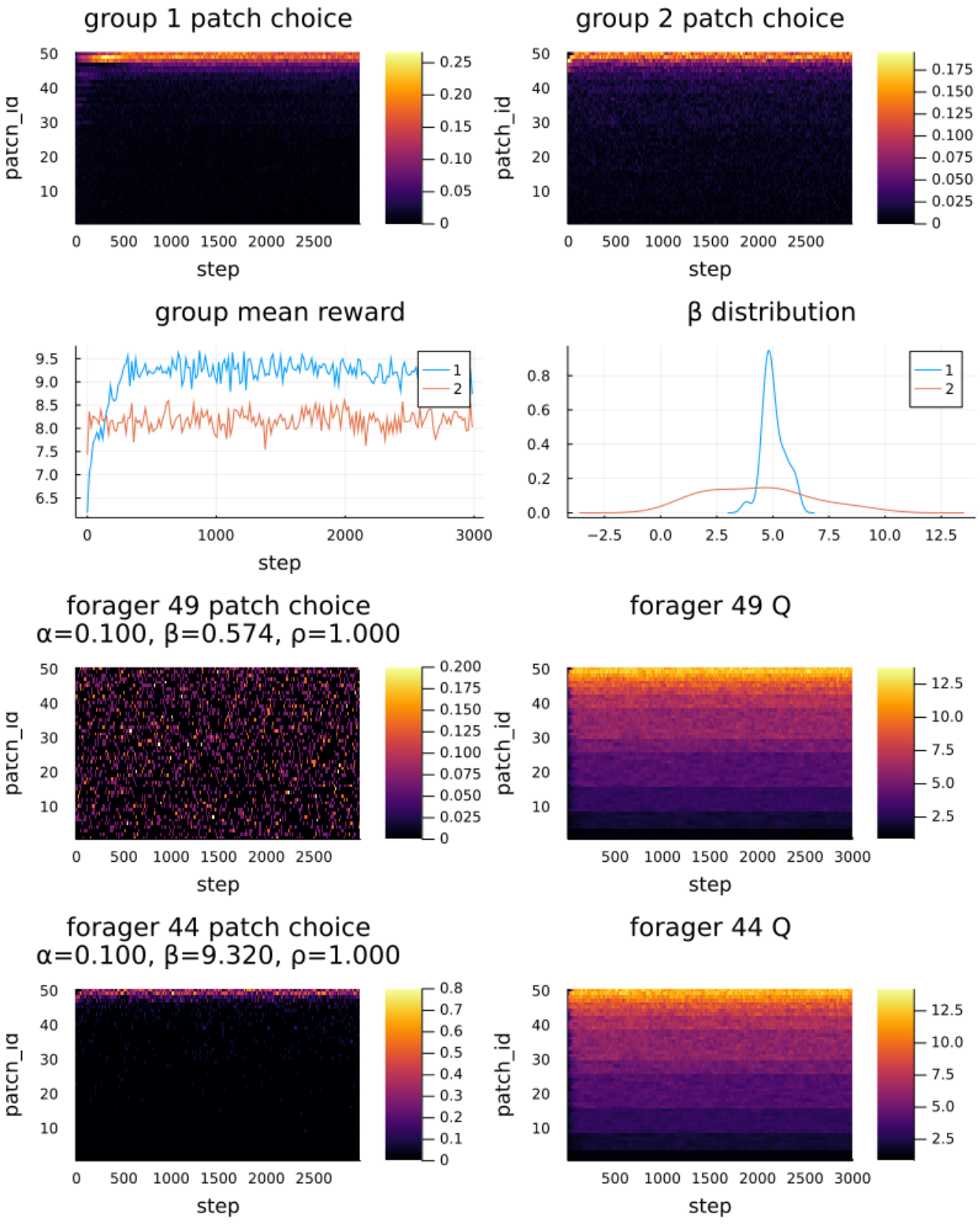
The contrast between forager 57 and 58's Q plots demonstrates the effect of learning rate.



Forager 49 is very exploratory and therefore disregards the values of patches even though it

Figure 2: Foragers: $\beta \sim LogNormal, \alpha = 0.1, \rho = 1$. Group 2 is more diverse in softmax temperature β. Forager 49 and 44 are, respectively, the most exploratory and most exploitative agent in group 2.

learned them accurately. Forager 44 is very exploitative and jumps to any patch and stays there for a while until it jumps again. It failed accurately learn the value of most patches because it didn't visit them. Group 2's choice plot is biased by its exploratory members.



The same initial conditions from the previous run is preserved. With communication, even though forager 44 is very exploitative, it learned from exploratory members' experiences

Figure 3: Foragers: $\beta \sim LogNormal, \alpha = 0.1, \rho = 1$. Group 2 is more diverse in softmax temperature β and enabled communication. Forager 49 and 44 are, respectively, the most exploratory and most exploitative agent in group 2.

and quickly lock onto the best patch. Group 2's performance is damaged by its exploratory members.

Discussions

Diversity can boost group performance via encouraging the division of labor.

Although our model explicitly included a parameter for risk preference (utility exponent ρ), we have yet to devise a proper way to analyze it. If the goal is to maximize total reward, the rational agent would be risk neutral ($\rho = 1$). Hence, ρ is set to 1 through the simulations. But there are scenarios where risk intolerance seems rationally required: for example, if an agent has to maintain a minimal reward per step and receives an additional penalty for failure, he might prefer a safer bet. That constraint actually goes both way for risk intolerance and risk seeking: if the minimal standard is sufficiently high to exceed the ordinary range of the safer bet, he would prefer to take a chance at the riskier one. Might a group benefit from diversity in this case? It seems possible if the group has to deal with different domains with different inherent reward "truncations." More advanced reinforcement learning algorithms can be used to properly consider reward stochasticity. Lowet et al. (2020) explores distributional reinforcement learning algorithms where the entire distribution of rewards are learned. The current model's reinforcement learning algorithm is rather simplified and a more advanced method would clearly separate the prediction problem (predicting the value of states and state-action pairs) from the control problem (improving the policy).

Despite discussing delay discounting extensively, we have yet to incorporate it in our model. On a high level, animals and humans discount future rewards because they can't satisfy immediate needs, which must be prioritized before future needs. In addition, excess present

rewards can also satisfy future needs. Andersen et al. (2014) concluded that humans generally discount exponentially at rate of 9%. This finding and the considerations just discussed can inform us to incorporate delay discounting in a more plausible way. Our model currently has no spatial dimension. the rewards could gain a temporal dimension by enabling spatial dimensions: further rewards take more time to retrieve.

Our model did not show that diversity in learning rates improved group performance. It was hypothesized that slow learners might be less sensitive to reward stochasticity and consistently choose to patch with largest μ_{rew}. That is indeed the case (Figure 1). However that effect is not pronounced in current settings and offset by whimsical fast learners. Fast learners can better respond to changing μ_{rew}. The ideal learning rate would be determined by how stochastic the environment is.

Diversity in softmax temperature generated some interesting results. The more varied group is performed worse because excessively exploratory members waste time exploring and excessively exploitative members fail to converge on the best patch (Figure 2). However, if the more diverse group enabled communication, they initially performed better but quickly plateaued due to the presence of more excessively exploratory members (Figure 3). This result is in line with research on division of labor. In this case, through sharing of experiences, some members are relegated to the role of explorers and others exploiters. Beshers and Fewell (2001) discussed some mechanisms for division of labor in social insects: 1) diverse response thresholds to external stimuli, 2) self-reinforcement of successful experiences and 3) information transfer. The mechanism in our model can be said to be information transfer combined with diverse internal decision-making parameters, functioning similarly to response thresholds: workers with lowest thresholds will specialize in low-stimulus tasks. It was also hypothesized that if experience decreases thresholds, a negative feedback loop could strengthen task specialization from relatively minor threshold differences. Such a feedback loop might also be possible in our model by allowing agents to differentially trust other agents' reports: if higher performing agents are given more credence, they might specialize

as exploiters.

One important limitation of this research is that diversity is not numerically measured. In Hong and Page (2004), because each heuristic is composed of distinct elements from a fixed set, they derived a natural diversity measure that averages the number of different elements between any pair of agents. However, in our model the parameters are real numbers. When we increased the σ^2 parameter of the log-normal distribution, the actual mean ($\exp(\mu + \frac{\sigma^2}{2})$) increased too. Although the log-normal distribution ensured that the parameters are positive, it also restricted the distribution to a particular shape. Hong and Page (2004)'s method may be helpful to solve this problem: we still have to assume a distributional form for each parameter, but we can first draw a large pool of agents and rank them by their performance when they explore the patches alone. Then we can select two groups with similar group average performances and directly use the variance as the diversity metric. This approach is better than keep μ or the mean of log-normal distribution constant because each parameter influences the performance nonlinearly. It makes little sense to keep the arithmetic average the same: for example, two agents with $\alpha s = [0.05, 0.15]$ cannot be regarded similar to another two agents with $\alpha s = [0.06, 0.14]$.

References

23and Me Research Team, eQTLgen Consortium, International Cannabis Consortium, Social Science Genetic Association Consortium, Richard Karlsson Linnér, Pietro Biroli, Edward Kong, et al. 2019. "Genome-Wide Association Analyses of Risk Tolerance and Risky Behaviors in over 1 Million Individuals Identify Hundreds of Loci and Shared Genetic Influences." *Nature Genetics* 51 (2): 245–57. https://doi.org/10.1038/s41588-018-0309-3.

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström. 2014. "Discounting Behavior: A Reconsideration." *European Economic Review* 71 (October):

15–33. https://doi.org/10.1016/j.euroecorev.2014.06.009.

Bankes, S. C. 2002. "Agent-Based Modeling: A Revolution?" *Proceedings of the National Academy of Sciences* 99 (May): 7199–7200. https://doi.org/10.1073/pnas.072081299.

Beshers, Samuel N., and Jennifer H. Fewell. 2001. "Models of Division of Labor in Social Insects." *Annual Review of Entomology* 46 (1): 413–40. https://doi.org/10.1146/annurev.ento.46.1.413.

Datseris, George, Ali R. Vahdati, and Timothy C. DuBois. 2021. "Agents.jl: A Performant and Feature-Full Agent Based Modelling Software of Minimal Code Complexity." http://arxiv.org/abs/2101.10072.

Haushofer, J., and E. Fehr. 2014. "On the Psychology of Poverty." *Science* 344 (6186): 862–67. https://doi.org/10.1126/science.1232491.

Hong, L., and S. E. Page. 2004. "Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers." *Proceedings of the National Academy of Sciences* 101 (46): 16385–89. https://doi.org/10.1073/pnas.0403723101.

Kable, Joseph W, and Paul W Glimcher. 2007. "The Neural Correlates of Subjective Value During Intertemporal Choice." *Nature Neuroscience* 10 (12): 1625–33. https://doi.org/10.1038/nn2007.

Keidel, Kristof, Qëndresa Rramani, Bernd Weber, Carsten Murawski, and Ulrich Ettinger. 2021. "Individual Differences in Intertemporal Choice." *Frontiers in Psychology* 12 (April): 643670. https://doi.org/10.3389/fpsyg.2021.643670.

Lowet, Adam S., Qiao Zheng, Sara Matias, Jan Drugowitsch, and Naoshige Uchida. 2020. "Distributional Reinforcement Learning in the Brain." *Trends in Neurosciences* 43 (12): 980–97. https://doi.org/10.1016/j.tins.2020.09.004.

Lukinova, Evgeniya, Yuyue Wang, Steven F Lehrer, and Jeffrey C Erlich. 2019. "Time Preferences Are Reliable Across Time-Horizons and Verbal Versus Experiential Tasks."

eLife 8 (February): e39656. https://doi.org/10.7554/eLife.39656.

Luksys, Gediminas, Wulfram Gerstner, and Carmen Sandi. 2009. "Stress, Genotype and Norepinephrine in the Prediction of Mouse Behavior Using Reinforcement Learning." *Nature Neuroscience* 12 (9): 1180–86. https://doi.org/10.1038/nn.2374.

Macy, Michael W., and Robert Willer. 2002. "From Factors to Actors: Computational Sociology and Agent-Based Modeling." *Annual Review of Sociology* 28 (1): 143–66. https://doi.org/10.1146/annurev.soc.28.110601.141117.

Schoener, Thomas W. 1987. "A Brief History of Optimal Foraging Ecology." In *Foraging Behavior*, edited by Alan C. Kamil, John R. Krebs, and H. Ronald Pulliam, 5–67. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4613-1839-2_1.

Schultz, Wolfram, Peter Dayan, and P Read Montague. n.d. "A Neural Substrate of Prediction and Reward," 8.

Sutton, Richard S., and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. Second edition. Adaptive Computation and Machine Learning Series. Cambridge, Massachusetts: The MIT Press.