Who, When, How (Not) to Imitate? The Role of Imitation in Collective Intelligence, and Its Implications on the Design of Socio-Technical Systems

by

Eunseo Dana Choi

B.A. in Statistics, B.A. in Economics,
and Kellogg Certificate in Managerial Analytics
Northwestern University(2019)

Submitted to the Institute for Data, Systems, and Society & Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degrees of
MASTER OF SCIENCE IN TECHNOLOGY AND POLICY
and
MASTER OF SCIENCE IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2024

Authored by: Eunseo Dana Choi
Institute for Data, Systems and Society and
Department of Electrical Engineering and Computer Science
September 29, 2023

Certified by: Dylan Hadfield-Menell
Assistant Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by: Frank R. Field III
Senior Research Engineer, Sociotechnical Systems Research Center
Interim Director, Technology and Policy Program

Accepted by: Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Who, When, How (Not) to Imitate?

The Role of Imitation in Collective Intelligence, and Its Implications on the Design of Socio-Technical Systems

by

Eunseo Dana Choi

Submitted to the Institute for Data, Systems, and Society & Department of Electrical Engineering and Computer Science
on September 29, 2023, in partial fulfillment of the
requirements for the degrees of
MASTER OF SCIENCE IN TECHNOLOGY AND POLICY

and

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

Abstract

Humans collectively demonstrate coordination and progress on a massive scale, building, adapting, and thriving under the rules of different institutions. Researchers posit social learning as a mechanism for overcoming individual limitations, quickly adapting to environments, passing knowledge across generations, and enabling rapid cumulative cultural evolution. This thesis demonstrates how multi-agent learning (MAL) can facilitate counterfactual experiments that shed light on the performance of different social learning. Simulations present that the details of who, when, and how to imitate affect group fitness in distinct ways based on the size and homogeneity of the group: 1. unbiased imitation works well in homogeneous groups as long as there is a minimum age for agents to be imitated; 2. imitation strategies based on models' complete action history instead of their recent actions, although similar, can attain very different levels of group fitness; 3. very high levels of imitation probability (up to 98% in some cases) may be efficient for group learning. Results from this thesis complement and contradict accepted results from the literature. By explicitly comparing the mechanisms that govern the success or failure of group learning, findings from multi-agent learning can provide essential guidance for the design of socio-technical systems.

Thesis Supervisor: Dylan Hadfield-Menell
Title: Assistant Professor of Electrical Engineering and Computer Science

Acknowledgments

I am grateful to my advisor, Dylan Hadfield-Menell, for his commitment to my intellectual development and his consistent provision of direct, constructive feedback. Reflecting upon our weekly meetings, I realize that each one left me with renewed enthusiasm and a clearer perspective on my research.

I extend my appreciation to Elsa Olivetti for embracing me at the outset of this academic journey. Her unwavering support has served as a driving force, motivating me to push the boundaries of my research.

I thank Rakshit Trivedi, Rui Jie Yew, Neil Gaikwad, Julian Manyika, John Ryter, Karan Bhuwalka, Luca Montanelli, and Pinar Ozisik for engaging and stimulating research discussions over the past two years. They made doing research at MIT all the more enjoyable. Max Tan deserves special mention for his indispensable assistance with experiments and meticulous citation work. I thank Amy Cheung for her perceptive feedback on my writing. I also wish to acknowledge Barbara DeLaBarre for her exceptional administrative support, ensuring that no important details were overlooked and unnecessary distractions were minimized.

To my friends beyond the Algorithmic Alignment Group and the MIT community, your patience and understanding during this journey have provided much-needed respite and a sense of normalcy, for which I am truly grateful.

Finally, I want to express my deepest gratitude to my family—my mom, Kyoung, my dad, Jin, and my sister, Sophia—for their boundless love. You have been my unwavering pillar of support throughout this journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Reverse-Engineering Collective Intelligence

Insights from reverse-engineering collective intelligence could inform streamlining adaptation processes and improve performance in socio-technical systems. Social scientists posit that collective intelligence has propelled the evolution of human civilization and shaped intricate normative infrastructures. By delving deeper into the factors that facilitate or hinder group learning and collaborative problem-solving, we can better understand when and how groups evolve, thrive, and fail. This thesis focuses on social learning, hypothesized as the primary mechanism for efficient cultural accumulation and transmission [53, 100, 70, 69, 67].

Social learning is pervasive across various aspects of human society [69, 31], from students learning from their parents to companies carefully selecting their board of directors. In today's world, social computing technologies like content recommendation systems are channels for cultural transmission, facilitating efficient communication and imitation [68, 17, 2]. Furthermore, interactions among humans and agents are now expanding to two-way hybrid social learning. Algorithmic agents and AI systems, particularly large language and vision models (LLMs and LLVMs) producing cultural artifacts [121], already harness social learning from human labellers and the internet during training to acquire cultural knowledge. Humans learn from AI solutions as well, as exemplified by the alignment of humans and AI algorithms in the game of Go and chess [19].

In pursuit of collective knowledge and shared progress, unlocking the full potential of social learning demands a nuanced exploration that goes beyond its surface. We need a more comprehensive discussion on cases of (in)effective social learning [33, 27, 68]. Gaining insight into when and why social learning works, and conversely, when it does not, is critical to optimizing group outcomes. By dissecting the contextual factors that influence the effectiveness of social learning, we can pave the way for a more informed and strategic approach to utilizing social learning.

1.2 Thesis Statement

Multi-agent learning (MAL) facilitates counterfactual experiments that shed light on the performance of different social learning to complement and contradict accepted results from the literature. The details of who, when, and how to imitate affect group fitness in distinct ways based on the size and homogeneity of the group:

1. unbiased imitation works well in homogeneous groups of learners, as long as there is a minimum age for agents to be imitated;

2. imitation strategies based on models' complete action history instead of their recent actions, although similar, can attain very different levels of group fitness;

3. very high levels of imitation probability (up to 98% in some cases) may be efficient for group learning.

At its core, our thesis contributes to the development of robust theoretical foundations in the realm of collective intelligence. By explicitly comparing and evaluating the mechanisms that govern the success or failure of group learning, this thesis provides essential guidance for the proactive design of sustainable and efficient normative infrastructures across multiple domains.

1.3 Social Learning and Cumulative Cultural Evolution

Researchers from various academic disciplines, including evolutionary biology, cultural anthropology, and psychology, have studied social learning as a mechanism behind how people acquire and transmit cultural knowledge. The term "social learning" covers learning from

others through observation, learning directly through social interaction, teaching, or being influenced by others' arguments. Experts have built formal evolutionary models and run simulations to form theories on the role of social learning in cultural evolution across generations.

Previous studies suggest that indiscriminate, unbiased, high-fidelity social learning may not necessarily lead to cumulative cultural evolution represented by an increase in group performance over generations [14, 38, 11]. For example, in frequency-based Rogers's model [103], individual performance from social learning depends on others' choice of learning strategies. Roger's paradox warns that the average fitness of a population with unbiased social learners is no greater than the average fitness of a population consisting entirely of individual learners; the population would need a subset of individual learners to explore, innovate, and accumulate new information.

On the other hand, simulations from Rendell et al.'s study [99] demonstrated that high levels of social learning without model bias can still benefit population fitness. In their simulations, social learning agents copied random models that chose to EXPLOIT, or by their definition, pull an arm in the previous timestep. Imitated models tend to selectively perform actions that return high rewards from their repertoire of known actions, thereby providing a non-random selection of information to imitators. Rendell et al. argue that social learning is useful when it contributes to adaptive filtering, helping the population retain adaptive and abandon maladaptive traits.

What seems to matter for cultural learning is the micro-level details of when and how learners imitate. With results from a theoretical model supported by ethnohistorical and archeological examples, Henrich [49] argues that group innovativeness is largely determined by the size of the population and the level of interconnectedness, rather than individuals' novel inventions. A large population in one generation would enable lucky errors and incremental additions for the next generation to refine and extend. An increase in the flow of information (i.e. high interconnectedness) would ensure details of useful ideas are paid attention to and diffuse across the network. Henrich cites empirical learning experiments showing that group payoff rises significantly when high-quality social information is available for frequent imitation.

Beyond unbiased social learning, previous works have also explored who individuals should rely on social learning or learn on their own. Richerson and Boyd [16] suggest that natural selection favors social learners who can distinguish cultural variants and copy the most successful models. Subsequent empirical studies provide evidence that humans employ a more nuanced set of selective social learning strategies (selective SLS) rather than blind, unbiased imitation.

However, whether that be interventions in lab experiments [112] or passive observations from ethnographic research [52], most empirical works focus on one or a few social learning strategies (SLS) in human learning within one fixed environment (domain/context/population) due to cost and time constraints. Further research is needed to holistically examine learning decisions and environmental factors that address who, when, and how individuals should (not) imitate to bring about successful group outcomes.

1.4 Why This Methodology

This thesis aims to provide a new level of concreteness in understanding the role of imitation in collective dynamics. Individuals with informational goals are modelled with agent-based models (ABMs); the ability to faithfully model specific characteristics with ABMs can help test the effect of different learning decisions in emergent group behaviors [77, 71, 22, 99]. Such computational models of goal-driven agents serve as a valuable sandbox to generate and test hypotheses, aggregate data into more extensive theory, and thereby enable rigorous exploration and expansion of cognitive and social science concepts.

Characteristics of each learning agent have similarities to human behaviors and belief updates shaped by positive/negative feedback and sampling from an internal distribution [8]. Thompson sampling [114] is used to represent the individual decision-making process behind choosing an arm for asocial learners. Social learning agents imitate others to choose an action and integrate its reward information with previously learned information when updating policy weights [115, 87].

Multi-armed bandits are used to model problems of learning, aggregating, and transmitting complex cultural skills, knowledge, and norms. Such environment with discrete action space allow direct measurement of group action alignment or behavioral diversity which are posited

have implications on group problem-solving and collective action. By distilling accumulated wisdom from micro-level interactions, findings from bandit simulations act as an initial foray into my comprehension of humans' adaptive abilities across different institutional contexts.

1.5 Scope

1.5.1 Objective

By comparing various distributed social learning algorithms and their effects on group performance, this thesis aims to provide further insights into when they facilitate or hinder group learning. The primary objective of this project is not to provide solutions for cooperation or propose methods to enhance state-of-the-art learning algorithms. While the set of social learning algorithms considered in this thesis is based on plausible models of human behaviors from the social learning literature, identifying the most faithful algorithm to the human cognitive process is out of the scope of this project.

1.5.2 Environment

This thesis aims to investigate the learning dynamics of goal-directed learners across multiple generations who independently pursue homogeneous tasks. The task involves learning, aggregating, and transmitting complex skills or cultural information like social norms. Each agent learns to navigate the same environment across its lifespan. The group shares an informational goal, but the rewards they observe are unaffected by the actions of other agents. There is no pre-defined set of expert demonstrators in our considered environments; each imitating agent learns from another learning agent. An example of this environment is humans learning norms from observing others' behaviors; policy designers would be interested to see whether humans will sustain an inter-generational equilibrium or whether they will shift over time.

The thesis does not explore scenarios that involve a shared environment with strategic interactions; this type of environment introduces interdependence among agents, where the actions and rewards of other agents influence their own rewards. An example of this scenario is a sequential social dilemma, where agents with differing motives must learn to adapt and coordinate. The environment presents another complex problem different from the one

currently considered in this thesis.

1.5.3 Levels of Analysis

This thesis considers individual-level learning processes; the use of distributed computation explains how macro-level phenomena emerge from individual-level algorithms [69]. When analyzing technology in relation to society, we want to consider its impact beyond the direct individual user of a system. Our levels of analysis motivate further research on coherent system design processes connecting an individual (user), communities, and society at once.

1.6 Thesis Organization

The rest of this thesis is organized as follows:

- Chapter 2 organizes previous research related to our work.
- Chapter 3 covers the technical background relevant to the following chapters. It also provides additional motivation for studying social learning with multi-agent learning by reviewing cultural technologies that utilize social learning to leverage, transmit, or shape cultural knowledge.
- Chapter 4 presents two observations: 1) unbiased imitation with a minimum age filter helps learning in homogeneous groups and 2) learning environments and strategies conducive to a greater quantity of accurate, social information are more effective for improving group learning efficiency among these unbiased learners.
- Chapter 5 of the thesis highlights that 1) selective imitation can be efficient for group learning even when the probability of imitation is set at high levels but 2) adapting imitation strategies to available model characteristics is crucial.
- Chapter 6 considers belief-based imitation, an alternative imitation strategy based on the model's complete action history, in small groups of agents facing a linear contextual bandit. Results validate that imitation may be efficient for group learning even at very high levels of group imitation probability, provided that agents' model biases are adaptive to the available imitation strategy.

- Chapter 7 discusses how multi-agent learning could and should be used to examine the long-term societal impact of policy and design choices for socio-technical systems. It concludes by briefly discussing potential avenues for future research.

Chapter 2

Related Work

2.1 Comparing Social Learning Strategies

Previous analyses of evolutionary and game theory models suggest that social learning is advantageous when imitated models are accurate, and most of the population is not social learning [103, 14, 11]. When the proportion of non-hybrid social learners increases, imitating loses value due to reduced reliable information from asocial learners. In other words, to optimize group learning, certain individuals should consistently generate information through non-social means while others should use social learning sparingly and complement it with their independent observations [65, 38].

Selective imitation among (hybrid[1]) individual learners seem to be crucial for group benefits – filtering social information based on quality before imitating [49]. Previously studied social learning strategies involve decisions like when (e.g., "copy-when-uncertain" heuristics) and who (e.g., "copy-successful-individuals" heuristics) to learn from [70, 62, 16]. In-depth studies from many fields of social science explore different tactics and principles that enhance the effectiveness of social learning; a combination of empirical investigations and computational methodologies have shown that specific use of social information can cascade throughout the group and, as a result, qualitatively alter group outcomes [13]. For example, extensive experiments have shown that those who selectively copied people based on their performance

[1]Hybrid social learners oscillate between individual learning and social learning. While many seminal theoretical works [14, 103] consider social learning and individual learning as distinct alternative strategies, humans are capable of doing both.

in the previous timestep, compared to those who did not copy or copied randomly, retained knowledge of the most complex but most efficient algorithm across generations [113]. However, most studies study one or few social learning decisions side-to-side in specific group environments over a limited time period (in terms of generations) due to time and cost constraints [85, 101, 84, 21, 71, 99, 78, 82, 34, 77]. Humans have demonstrated the use of hybrid strategies that combine multiple learning decisions, including who and when to imitate across different learning environments [79, 87, 115].

While diverse arrays of selective learning strategies have been discussed, more work could be invested in explicitly and rigorously comparing learning strategies to identify important learning parameters for group learning. Previous works have tried to collect and organize theoretical and empirical evidence that predict and report different animal and human learning strategies in different contexts [70, 54, 65, 61]. The most related work to this thesis is a competition organized by Rendall et al. [99], which used multi-armed bandit simulations to expansively explore the relative strengths of a wide range of learning strategies submitted as entries. All learning strategies in this study assume agents to choose among three different actions at every time step: 1. (EXPLOIT) pulling an arm and receiving its actual payoff, 2. (INNOVATE) learning about previously unexplored behavior and its payoff, and 3. (OBSERVE) learning noisy information about demonstrators' actions and their corresponding rewards where demonstrators were random agents who played EXPLOIT in the previous round. The study reports that winning strategies relied heavily on social learning (OBSERVE), explored less, spent more time exploiting learned knowledge, and estimated rates of environmental change to identify still relevant actions. Learning strategies from this competition did not consider model biases.

In contrast, this thesis is different in that it details how environment and agent learning characteristics together influence the relative effectiveness of different social learning strategies. First, Chapter 5 examines a broader set of prestige-biased strategies, addressing from whom one should learn. Moreover, learning agents in this thesis make decisions about which arm to pull at every time step to balance exploration and exploitation; in other words, agents can only INNOVATE by pulling an arm, and they can only OBSERVE and learn (noisy) reward information about a demonstrator's action after pulling an arm. Imitation is *one way of decision-making and learning* that facilitates exploration and exploitation.

2.2 Who to Imitate

Payoff bias, or copying based on recent payoff achieved by the demonstrator, is commonly studied and is effective in adopting and transmitting complex skills [112]. In the real world, however, reliable payoff information for every action others have taken is not always available.

Prestige cues, or model-based biases (copying based on others' status), are more practical alternatives to payoff biases in learning and adaptation [102, 64]. Through an online learning experiment in which participants could copy others to score points in a general knowledge quiz, [18] shows that prestige acts as an indirect cue of success when there is no reliable information about success cues.

However, how prestige is measured varies across studies, and their relative effects remain unclear. In animal learning, prestige is considered in terms of rank and is assumed to correlate with the model's success or skills [70, 21]. [51] considered prestige as someone who displays culturally valued skills and receives much deference. Previous studies demonstrate that model biases can interact or be deployed simultaneously [94]; for example, children consider both age and competence when choosing who to copy [119]. These inconsistent definitions of prestige signals across studies and limited discussion on relevant learning environments in which they are helpful leave designers of agent interactions with the question of how to use prestige signals to facilitate group learning given a domain. Direct and relevant measures of prestige should be better defined to compare model biases and evaluate the effect of prestige on group fitness.

Experiments in Chapters 5, and 6 consider different model biases to compare with asocial learning and unbiased imitation: skill (score summary), age, or accounting for both (metrics: age_skill, age_restricted_skill). Agents are ranked based on these metrics and have imitating agents follow agent(s) ranked highest at each timestep. Age_skill bias accounts for both age and skill of potential models and imitates agents with the highest cumulative reward. Social learning groups with age_restricted_skill bias apply an age filter before model selection, only considering agents that have played for longer than 20 timesteps and choosing the one with the highest mean reward. By considering both skill bias and age_restricted_skill bias, the thesis compares collective outcomes from imitating an agent with the highest average score versus imitating an agent with the highest average score who has gained a minimum

learning experience.

2.3 How to Imitate

The method of imitation also matters. Research from developmental psychology has observed over-imitation, where people copy even causally irrelevant demonstrations. Scientists have found this behavior in specific [90, 76] and across [80] age ranges. This thesis assumes faithful imitation is possible in bandit games; the following chapters mainly consider explicit imitation in which agents pull the same arm previously pulled by a model. Explicit action-based imitation is defined as imitating arm most frequently pulled by the imitated model from a pre-defined time frame. Implicit belief-based imitation is also considered. Implicit imitators pull arm based on temporary belief distribution, which is independently formed based on observed actions pulled by model imitated at each timestep; models' actions are weighted differently based on their age when choosing which arm to pull.

Experiments in this thesis extend previous work and explore potential cues from agents' goal-directed actions that may be useful for inferring their knowledge in a sequential decision-making task [46, 47]. Chapters 4 and 5 look at the interplay between the amount of models' action history and other decision factors in group learning, varying how much models' actions are considered for imitation. Imitating agents pull the most often pulled arm by the imitated model in the previous timestep $t-1$, the last N previous timesteps, or the entire action history. In Chapter 6, agents directly adopt models' belief distributions, which get updated as their experience accumulates and are essentially a summary of their complete action history.

2.4 When to Imitate

Environments constantly change, and they do not go in one pre-designated direction; as a result, learning strategies may need to be chosen adaptively. Some posit that prestige bias is favored only when there is a positive relationship between status and performance; others predict that age should be associated with prestige only when the environment is stable and age has a positive linear relationship with skills [53, 61]. Henrich and Broesch point out that people copy others more in domains where there is significant variation in knowledge [50]; in

cases where the variation is small, and most individuals already possess shared information, social learners would benefit less from imitating (prestigious) individuals. Theoretical models and empirical evidence also suggest that social learning is more likely to be effective in larger groups [15, 86, 109, **?**, 66, 118, 49]. However, this may be dependent on a complex interplay of group conditions like the difficulty of a given task, the accuracy of the model, the accuracy of the average population, the number of available learning sources, the level of uncertainty in an environment, the amount of conformity within the group, interconnectedness, or network structure [91, 115, 83, 34, 77, 71, 10].

Chapters 4, 5, and 6 broadly evaluate model biases and explore when they may be (mal)adaptive for multi-agent learning by examining their performance across different environmental conditions. First, a set of selective social learning strategies is tested, varying the population average imitation probability to represent the level of network efficiency or model dependency; a more efficient network would facilitate a higher group imitation probability and greater access to social information. Then, selective social learning strategies are compared across group sizes and varied learning abilities. By exploring policy performance in bandits with a discrete action space, the amount of behavioral diversity is measured to evaluate the effect of social learning on collective knowledge both in terms of quality and quantity; Herfindahl-Hirschman Index (HHI) of agents' action, or action convergence, measures agents' level of conformity.

2.5 Why Might Social Learning (Not) Work?

Social learning could facilitate the accumulation of diverse knowledge, [39, 32] which tends to increase with group size [81, 49]. [56] suggest that diversity in information, bias, and learning/decision-making strategies can reduce errors cascading throughout the group and help with group performance. Diversity in knowledge and learning strategies is also predicted to lead to cultural variance and greater innovation [86, 48, 49]. However, diversity alone is posited to be an insufficient condition for group performance if the quantity does not correlate with the quality of information [96].

On the other hand, social learning could decrease variance and increase bias within the group. In their computational study analyzing characteristics of groups following different

social learning strategies, Rendell et al. observed that frequent social learning strategies lead to a greater amount of knowledge endured over time for the group but a smaller amount and uneven distribution of behaviors expressed by the group due to exploitation [100]. Disproportional social influence of a select few members facilitated by selective imitation of observable behaviors can potentially bias the crowd and lead to undesirable group outcomes [75]. Although there is a lack of consensus on the effect of network structure, [71] posit that a decrease in the speed to reach majority solutions, moderated by less efficient network structures, could be a solution to improve group performance, especially in error-prone networks.

2.6 Social Learning in Reinforcement Learning Paradigm

Reinforcement Learning (RL) has been considered to be a useful framework for studying the human brain and mind [40, 59, 116]. A more complex test bed than ABMs, RL models allow the computational representation of potential problems and solutions for goal-driven agents navigating a new, uncertain environment.

Previous RL research has studied imitation or observational learning from expert demonstrations. In off-policy, optimal policies are estimated and learned based on experiences collected using a different policy. Inverse reinforcement learning (IRL) aims to infer the underlying reward function that would explain the model behavior, given observed actions taken by a demonstrator over time in an environment and assuming implicit optimality in their actions [104, 1, 55, 74]. For these strategies to be effective, learners must overcome the challenge of identifiability, or the need to accurately map between the reward structure and demonstrators' actions, since different reward functions can lead to the same actions [89].

Instead of being given the optimal policy, some studies (including this thesis) consider agents learning from another learner. This paradigm is increasingly reconsidered in recent studies as it is more realistic and may even offer more helpful information for learning than what is already an optimal policy; the fundamental assumption here is that the learning policy of an observed model will improve over time [104, 120, 57]. [44] demonstrates that learning within this framework is achievable when the demonstrator initially explores an unfamiliar environment, progressively enhances performance, and then prioritizes exploitation based

on an approximately optimal policy. Studying how policies evolve or how trajectories are modified can offer valuable insights into identifying the sub-optimality of a policy.

Chapter 3

Background

3.1 The Stochastic Multi-Armed Bandit

The multi-armed bandit problem (MAB) represents a sequential learning and decision-making process. At each time step, a decision-maker must choose from a finite set of actions (the bandit's arms). Each action has an unknown reward-generating process or distribution. The corresponding reward is immediately observed. In the next time step, the decision-maker can either exploit by pulling an arm that has already been visited and has a high estimated success probability or explore by pulling arms that have uncertain/unexplored success probabilities. Exploitation maximizes immediate performance while exploration accumulates new information about the environment. To succeed in this task, agents must find a balance between exploiting the known arms of high rewards and exploring unknown arms that may offer even greater rewards.

Rewards are generated at each time step, independently of past actions. Performance is often measured by regret, which measures the difference between the best possible reward of an arm it could have pulled and the actual accrued reward.

3.1.1 Beta-Bernoulli Bandit

Beta-Bernoulli bandits have binary rewards $\{0, 1\}$. At each time step, the agent decides which arm to select based on its beliefs about each arm. The closed-form update rules of

the Beta distribution and the use of conjugate priors enable efficient and analytical updates of the beliefs about the success probabilities. The key components of a Beta Bernoulli bandit can be defined as follows:

- **Action Space:** Let $\mathcal{A}$ denote the set of possible actions or arms. Each action $i \in \mathcal{A}$ is associated with an unknown success probability μ_i, representing the likelihood of receiving a reward of 1 when selecting arm i.

- **Reward Model:** The unknown expected reward for each arm i is given by $\mathbb{E}[r_i] = p_i$. The reward for arm i in timestep t, denoted as $r_{i,t}$, follows a Bernoulli distribution with parameter p_i:

$$r_{i,t} \sim \text{Bernoulli}(p_i)$$

  where p_i is the unknown probability of success for arm i. The true reward probabilities form a vector $p = (p_1, p_2, \ldots, p_K)$. In stationary bandits, p is fixed across timesteps.

- **Learning Objective:** The agent aims to learn an effective policy that maximizes the expected total reward over time by adapting the choice of arms based on observed rewards and updated beliefs.

3.2 Linear Contextual Bandit

A linear, stationary contextual bandit [4] is a sequential decision-making problem that extends the classic bandit setting to incorporate contextual information.

At each time step t, a learning agent faces a set of K distinct arms, and each arm i is associated with a context vector $x_{i,t} \in \mathbb{R}^d$ that provides additional information about the arm. The agent decides at each time step which arm to select based two types of information: 1) the observed contexts (in forms of d-dimensional feature vectors) at the current timestep and 2) historical pairs of chosen context and its reward.

The learning agent faces the decision to exploit existing knowledge by selecting the arm that maximizes the expected reward based on its $\hat{\mu}^*$ (i.e. current belief about μ^*) or to explore by selecting an arm to gather more information about its reward potential. Over time, it

collects data about the relationship between contexts and rewards, allowing the agent to predict a high reward arm with $\hat{\mu}^*$.

The key components of a linear, stationary contextual bandit can be defined as follows:

- **Context Space:** Let $\mathcal{X}$ be the context space, representing the set of possible context vectors. Each context vector $x_{i,t}$ provides contextual information about arm i at time step t. These contexts can be chosen adaptively according to design intentions or the learning environment.

- **Action Space:** The agent can choose from a set of K possible actions or arms, denoted as $\mathcal{A} = \{1, 2, \ldots, K\}$.

- **Reward Model:** The rewards are determined by an underlying true parameter vector $\mu^* \in \mathbb{R}^d$ that is unknown. For a chosen arm i at time step t and its context $x_{i,t}$, the expected reward is given by $< \mu^*, x_{i,t} >$. In stationary bandits, μ^* is fixed across time steps.

- **Learning Objective:** The agent's goal is to learn an effective policy $\pi : \mathcal{X} \to \mathcal{A}$ that maximizes the cumulative reward over a sequence of time steps while adapting its decisions based on the observed contexts and historical rewards.

3.3 Thompson Sampling

Thompson sampling [114], or Bayesian posterior sampling, is a policy commonly used in decision-making problems to balance exploration and exploitation. The method grew out of interest in planning for research, specifically for better use of existing data to guide planning for future data collection. For example, we want to continue minimizing the number of times a treatment is sub-optimal when we compare and assign treatments to individuals. The method became widely adopted after several studies demonstrated its efficacy empirically [43, 42, 26, 105] and theoretically [3, 4].

The sampling method initially forms priors or current beliefs about each arm (i.e., its probability of being optimal). The algorithm then independently samples from these distributions at each time step and selects the arm with the largest sampled value. Based on the observed

reward of the arm, the algorithm updates posterior distributions over each arm's unknown true reward probabilities (or action). The underlying idea towards more adaptive and efficient decision-making is to 1) estimate the uncertain information or distributions about model parameters, 2) make decisions according to the assumed prior beliefs, and 3) update these distributions as more (reward) data is generated and collected. Conjugate models have analytically tractable posterior.

Algorithm 1 Thompson Sampling

```
Initialize parameters and prior distributions for each arm.
for t = 1, 2, ... do
    for each arm i do
        Sample μ̂_i^(t) ~ P(μ_i | r_1, ..., r_{t-1})
    end for
    Choose arm j with the highest sampled value
    Observe the reward r_t by pulling arm j.
    Update the arm's posterior distribution using the new observation.
end for=0
```

3.3.1 Thompson Sampling in Beta-Bernoulli Bandits

With Thompson Sampling, an agent forms initial beliefs about the success probabilities for each arm, represented as prior distributions. The prior distribution for arm i is denoted as $Beta(\alpha, \beta)$, where α and β are the hyperparameters of the Beta distribution. After observing the reward r (either 0 or 1) from pulling an arm at each time step, the decision maker updates their beliefs about the success probability using a closed-form update rule. If $r = 1$, the updated distribution becomes $Beta(\alpha + 1, \beta)$, reflecting the incorporation of a success. If $r = 0$, the updated distribution becomes $Beta(\alpha, \beta + 1)$, reflecting the incorporation of a failure. For arm i after t rounds, the posterior distribution is given by:

$$\hat{p_{i,t}}|r_{i,1:t-1} \sim \text{Beta}(\alpha_i + \sum_{j=1}^{t-1} r_{i,j},\ \beta_i + t - \sum_{j=1}^{t-1} r_{i,j}) \tag{3.1}$$

where $D_{i,t-1} = r_{i,1:t-1} = r_{i,1}, r_{i,2}, \ldots, r_{i,t-1}$ is a series of past observations about arm i. Agents pull the arm with the highest $\hat{p}$. Thompson sampling policy asymptotically achieves optimal solution in Bernoulli bandits [63].

3.3.2 Linear Thompson Sampling in Linear Contextual Bandits

With linear Thompson Sampling, an agent assumes linearity in rewards; rewards of a linear contextual bandit are assumed to be sampled from

$$R_t|\mu^*, \mathcal{X}_t \sim N(<\mu^*, \mathcal{X}_t>, \sigma^2\mathbb{I}_d)$$

At timestep = 0, precision matrix $A_0 = \mathbb{I}_d$ and $\mathcal{X}_0 = b_0 = 0_d$. At each time step t, an agent samples $\widetilde{\mu}_t$ from $N(\widehat{\mu}_t, \sigma^2 A_t^{-1})$ and choose arm i, or its the context $(x_{i,t})$, maximizing ${\widetilde{\mu}_t}^\top x_{i,t}$ given $\mathcal{X}_t$ and past observations $D_{t-1} = \{(x_s, r_s)\}_{s=1:t-1}$.

After observing the reward $r_{i,t}$ from pulling arm i, it then updates its belief about μ:

$$\widehat{\mu}_t = A_t^{-1} b_t \tag{3.2}$$

where precision matrix $A_t = A_0 + \sum_{\tau=1}^{t-1} x_{i,t} {x_{i,t}}^\top$, $b_t = b_0 + \sum_{\tau=1}^{t} x_{i,\tau} r_{i,\tau}$.

3.4 Cultural Technologies Characterized by Social Learning

This section examines cultural technologies that facilitate or leverage social learning, thereby harnessing, disseminating, or shaping cultural knowledge. The term "social learning" covers learning from others through observation, directly through social interaction, teaching, or being influenced by others' arguments. Technology and policy designers can proactively predict and shape long-term collective outcomes by modelling different details of social learning strategies that a group of (human and/or artificial) agents employ at a micro-level.

Computer-Mediated System Facilitating Human-Human Cooperation

Content recommendation systems and social media platforms (e.g. Reddit) have mediated social learning among humans online, facilitating the transmission of human culture. A search engine is a primary destination for individuals seeking political information [35]; products it offers are cultural artifacts reflecting collective knowledge. Comment sections in online discussion forums serve as an online public sphere that enables interactions among various users, including those who initiate the discussion, respond to it, and observe. Users'

initial comments often elicit follow-up responses and reactions, such as likes, shares, comments, emoji reactions, flagging, or follows.

Design choices around curating and ranking media on these platforms shape the dynamics of users' attention and cultural consumption [23]. Items are placed within the critical window based on some platform curation/content-scoring mechanism, which directs most users' attention and engagement [108]. For social media platforms, the number of user reactions (e.g., up/downvotes, the "Insightful" emoji on LinkedIn, the "Care" emoji on Facebook, and the proposed Respect button [110]) are often used to rank the order of comments. Over time, the accumulation of these reactions influences users' perception of the public opinion on critical issues [93, 88, 72] and their own value judgments about the relevant topic [111]. The type of accumulated reactions can also significantly impact the dynamics of user interactions; exploring user-generated labels (UGLSs) that are adaptive at a comment-level, [68] characterizes the effects of information rich-reach trade-offs across reaction designs. While ranking based on some pre-defined signals can filter high-quality information, it can also pollute the information ecosystem by amplifying bias [97] or inaccurate content; previous research has studied optimization-driven systems as a driving mechanism behind echo chambers and polarization [123, 20, 107].

Interactive Social Interfaces Between Human and AI

Interactive social interfaces often learn from and leverage various types of human behavioral inputs before and after deployment to align system behavior with user preferences or social norms:

- Individual content recommendations are generated by mechanically incorporating users' history of past interactions; recurrent models like LSTM that model users' hidden states have been deployed on YouTube [12].

- Often, paid human raters that pass some minimum performance qualifications produce training data that classify positive and negative content for content moderation and ranking.

- A recent study proposed a cost-efficient way to collect people's preferences and train reward models for large algorithmic systems: they collected user reactions to Reddit

comments as signals for helpful responses [37]. The metric they optimized for would be relevant in systems with informational goals like question-answering context.

- Another study presented improvement in the performance of a chatbot by having it learn from implicit signals in human dialogue rather than explicit button presses like up/downvotes with practical scalability issues; they collected human interaction data with the interface to develop rewards the system optimized for and trained it with reinforcement learning [58].

- Similar to simulated agents from this thesis that infer high reward arms from other agents' action trajectories, users' trajectories of items they observe and react to (over a long horizon) are often valuable information for algorithmic systems to predict their (slowly evolving) preferences [23].

- Previous research has demonstrated the relationship between the performance of a large language model and the context window size limitation that constrains the length of users' behavioral sequences [122].

As such, user feedback and behavioral trajectories can offer distinct and useful information to interactive systems. However, their effects and efficacy can vary by modes of collection and usage context [60, 24]. The complex interplay of specific types and amount of feedback, data collection methods [9], and use in the context of a real system motivates further research to effectively provide feedback for a given system.

Generative Foundation Models Shaping Culture

Large Language (and Vision) Models (LLMs, LLVMs) or generative foundation models are both products and transmission channels of human culture [121]. As products, they encapsulate the collective knowledge, values, and perspectives embedded within past texts and data that fuel their training. They emerge as tangible results of cultural creation, wielding the potential to reflect and mold the very cultural landscape they inhabit. In response to user prompts, model outputs can transmit linguistic patterns, biases, and societal norms ingrained within the texts on which they have been schooled [2].

More importantly, these models can shape human culture through their generative capability

[19]. However, we have yet to determine where we are headed with this perpetual ebb and flow of cultural evolution. Low-effort and cost-effective mass content production after the introduction of highly accessible interfaces powered by generative foundation models (e.g., ChatGPT and Midjourney) has raised concerns about permissible use cases and the future of creative and information landscapes:

- The proliferation of AI-generated art is altering the landscape for artists to upload and share their art, providing evidence for humans' unbiased, indiscriminate imitation of outputs from generative AI models. Once considered safe havens for artists to showcase their portfolios and interact with their audience, art portfolio sites and marketplaces like DeviantArt and ArtStation are seeing a rapid change in the percentage of human versus AI art. To avoid having their artworks trained, artists actively pull down their works and even remove their accounts, which seems to be the only viable way to protect their future works (Appendix A-1).

- Amazon is faced with ramifications of AI-generated content. For instance, The New York Mycological Society has warned of AI-generated foraging books for beginners on Amazon that contain false and potentially deadly information (Appendix A-2). As another example, Amazon's young adult romance best seller ranking is infiltrated with AI-generated content as well.

- Additional example on changing dynamics of content production is provided by the decline in Stack Overflow's traffic over the past year and a half—a nearly 50% decrease in the number of questions and votes received by posts(Appendix A-3).

The above examples across different content platforms demonstrate dynamic interactions establishing a high probability of two-way social learning between humans and interfaces powered by generative foundation models. These interactions prompt reflection on how technology reshapes and affects the quality of the public cultural knowledge pool. We need a more comprehensive evaluation that analyzes user interactions with technology at a micro-level and connects its impact on the community at large so that we can outline adaptive standards and boundaries for content and consumption practices. These standards will then inform platforms to develop low-cost mechanisms for enforcement and compliance as rapid content creation of generative AI models scales up the number of taxing conflicts and

ramifications.

Chapter 4

Comparing Random Social Learning Strategies

4.1 Overview

Using multi-agent, multi-armed bandits, this section simulates the multi-generational learning dynamics of a group of agents facing an independent, homogeneous task with a shared informational goal. The study aims to explore the effectiveness of imitation in learning; more specifically, it assesses nuanced variations of unbiased imitation.

Results show that even unbiased imitation can facilitate knowledge accumulation and transmission, benefiting homogeneous agents both at an individual and collective level. However, the efficacy of imitating unbiased learners depends on the nuance of how and when it is applied. The performance of unbiased imitators is more efficient and reliable when age filter is applied for model selection, especially when they imitate based on models' complete action history at a high group imitation probability. Age filter reduces the risk of imitating suboptimal models, and imitating based on the complete action history maximizes the transfer of accurate information among agents.

4.2 The Environment

Multi-agent, multi-armed bandits (MAMAB) is two-folded in nature; it is an exploration-exploitation problem and an information accumulation problem [69]. To be successful in a bandit task, it is important for an individual to effectively balance exploration and exploitation. Initially, exploration is vital to identify the optimal arm. The optimal strategy involves rapidly discovering and exploiting the arm with the highest expected reward to maximize cumulative returns (or minimize cumulative regrets) over time. For a group of social learners to harness collective intelligence and successfully solve a bandit problem, it is important to effectively pool and transmit information across generations, given agents' partial beliefs or knowledge about the environment.

4.2.1 Beta-Bernoulli bandits

In this section, agents face multi-armed, Beta-Bernoulli bandits. In following experiments, uniform distribution $Beta(1,1)$ is initially assumed for each bandit arm i. Arms' reward probabilities are also sampled from $Beta(1,1)$. At time t, with $\alpha_i(t))$ successes and $\beta_i(t)$ failures in $k_i(t) = \alpha_i(t) + \beta_i(t)$ plays of arm i, the posterior of μ_i is updated as $Beta(\alpha_i(t) + 1, \beta_i(t) + 1)$.

Rewards are private; agents can only observe their reward at the current timestep. They do not have access to information about the rewards obtained by other agents. Agents have access to and can discriminate two valuable sources of information: 1) the action trajectories pursued by other agents up to the present time point $\{\tau_{i,0:t}, \ldots, \tau_{n,0:t}\}$ 2) prestige level.

Based on a pre-defined prestige signal, social learning agents imitate the most prestigious model at each timestep with a fixed imitation probability. When imitating, an agent pulls the most frequently chosen arm by the imitated model.

Asocial learners use Thompson Sampling (See Section 3.3.1) to select and pull an arm. Social learners who are not imitating at timestep t choose an arm based on Thompson sampling [114]. All agents have an average lifespan of 50 timesteps (sd = 10 timesteps). When agents "die," agents' action and payoff repertoires are reset, and agents are reborn. I did not consider the reproduction of agents and the inheritance of parental fitness.

4.3 Experiments

Through a series of empirical experiments using multi-agent, multi-armed bandits, group performance is compared across a variety of learning settings and stimuli.

All social learning agents oscillate between imitation and Thompson sampling (asocial learning) based on some pre-determined group imitation probability (imitate_prob).

Different variations of group random imitation or unbiased social learning are considered in this chapter: random, random_multiple, restricted_random. When a group randomly imitates, each agent imitates the same model randomly sampled from the population at each timestep. In the case of restricted_random, everything stays the same except now there is an age filter; agents are only sampled to be a model if they are above minimum age (lived for more than 20 timesteps). When there is no eligible model to be imitated in the population, agents do not imitate and instead learn independently using Thompson Sampling. When imitating at random_multiple, the group may not necessarily share the same model while each agent imitates one random model at each timestep.

At each timestep, imitating agents observe the model's action history to choose how to imitate; different access to its action history is considered the complete action history, actions taken by the model in the last ten timesteps, or the last action of the model. The most frequent arm pulled by the model in a chosen timeframe is pulled for imitation.

This chapter considers an environment where all agents have the same learning ability and observe correct rewards for arms pulled. It investigates different learning contexts for groups facing the 100-armed bandit problem, varying group sizes: 7, 50, and 100 agents. I consider different levels of network efficiency, varying the fixed probability of imitation: f = 20%, 40%,60%,80%,98%.

The study additionally considers the varied distribution of learning abilities (i.e. decision accuracy) within the group where a subset of agents tend to incorrectly update beliefs about arms upon observed rewards of pulled arms (learning_abilities = Beta(1,1) and Beta(0.5,0.5)). An agent's learning_ability φ closer to 0 means higher probability ($\tilde{\text{Ber}}(\varphi)$) of making errors in updating beliefs. The type of error is consistent across all agents: agents perceive positive rewards as non-positive rewards and update beliefs about the reward func-

tion accordingly. Agents' learning abilities gradually improve over their lifespans; each agent can increase its learning ability by 0.05 at most by the end of its lifespan.

See Table 4.1 for parameters considered in our experiments. Each trial is 5000 timesteps. The performance of learning agents is averaged across 15 trials.

Table 4.1: Experiment Parameters

Parameter	**Options**	**Description**
[WHO]	random, random_multiple, random with age restriction	Model Selection
[HOW]	Choose the most frequent actions from the model's complete action history, from the model's last ten actions, the last action	Imitation Type and Time Frame
[WHEN]	{0.2, 0.4, 0.6, 0.8, 0.98}	Imitation Probability (Network Efficiency)
	7 agents, 50 agents, 100 agents	Group Size
	All agents observe true rewards, Beta(1,1), Beta(0.5,0.5)	Distribution of Population Learning Ability
	100 arm bandit	Arm Size

4.3.1 Social Learning Policy

Our social learning algorithm takes several input parameters, including the number of agents (num_agents), the type of model bias (status_type), the population imitation probability (imitate_prob), the learning ability (learning_ability), the average agent lifespan (avg_lifespan), the number of timesteps (num_timesteps), and the number of arms (num_arms).

If Bernoulli(imitate_prob) returns positive, a social learner imitates as long as there are eligible models (e.g, agents above some minimum age in the case of restricted_random). A random model is sampled from the same group for all social learners in that time step to imitate. The agent then chooses and pulls an arm based on the previous actions of the selected model:

$$\arg\max_{a \in A} \sum_{i=T-t}^{T-1} \delta(a, a_i) \tag{4.1}$$

where:

A represents the set of bandit arms,

t is the amount of model's action history considered,

T is the current timestep,

$\delta(a, a_i)$ is an indicator function that equals 1 when a is equal to a_i and 0 otherwise.

Given an agent's pre-determined learning_ability, sampled from a chosen population learning distribution, if Bernoulli(learning_ability) returns negative, the agent observes an arm reward of 0 regardless of the actual reward and updates its belief based on the observation. Each agent's learning ability gradually improves with age by 0.0001 percentage points.

At the end of its lifespan, an agent's accumulated experience(action and reward history) all reset.

4.3.2 Evaluation

The primary measure of group performance is the mean regret summed and averaged across all agents, which can be expressed as:

$$\text{Individual Mean Regret for Agent i} = (\sum_{t=1}^{T(i)} \text{Optimal Reward}_t - \text{Reward}_t)/T(i) \tag{4.2}$$

where:

$T(i)$: Age of agent i

Optimal Reward_t : Maximum possible reward at timestep t (i.e. 1 in Beta-Bernoulli Bandit)

Reward_t : Reward obtained by agent i at time step t

$$\text{Group Mean Regret} = [\sum_{t=T-s}^{T} (\sum_{i=1}^{N} \text{Mean Regret}_{i,t})/N]/T \tag{4.3}$$

where:

T : Trial Length (i.e. 5000 timesteps)

s : Time period considered for evaluation (i.e. 2500 timesteps)

N : Number of agents

$\text{Mean Regret}_{i,t}$: Individual Mean Regret for Agent i at timestep t.

Group mean regret for each learning policy is averaged across 15 trials.

The performance of social learners is compared against the performance of asocial learners with the same average population lifespan (lifespan = 50 timesteps, marked in black), asocial learners with a longer average lifespan (lifespan = 250 timesteps, marked in gray), and immortal asocial learners (marked in gold).

The level of action convergence at a group level is measured for additional analysis. A normalized Herfindahl-Hirschman Index (HHI) is used to measure action convergence. HHI

calculates the concentration of arms pulled by all agents at the end of each episode. HHI of 0 means agents are exploring and have been pulling all arms equally. HHI of 1 indicates that agents are exploiting and pulling from just one arm. Group action convergence at timestep t is thus calculated as

$$\text{HHI}_t = \frac{(\sum_{k=1}^{K} p_{k,t}^2) - \frac{1}{K}}{1 - \frac{1}{K}} \tag{4.4}$$

where:

K is the number of bandit arms,

$p_{k,t}$ is the proportion of action k across all actions pulled by a group of agents throughout individual lifespans

By considering all actions pulled by each agent across its lifespan, we can measure the collective level of exploitation.

To compare group performance with model performance, model reward is also calculated, summing and averaging the imitated model's mean reward (skill) across timesteps. If the group imitates multiple models at the same timestep, model rewards are averaged among these models.

4.4 Results and Discussion

Social learning via unbiased imitation can be more effective for group learning than asocial learning with a longer learning period. Given that models imitated in this group of learners are selected at random, model performance among unbiased learners could be considered an average knowledge of an individual. Imitation turns out to be helpful for individual learning as well; model rewards among unbiased imitators employing this strategy (Top of Appendix A-4) in settings above 50% group imitation probability are generally comparable to that of asocial learners with a longer lifespan. Results imply that unbiased imitation can not just be beneficial for cumulative shared welfare, but also for individual agents as well, although the details of when and how agents imitate matter.

4.4.1 Unbiased imitation, coupled with a minimum age filter, proves reliable across a range of homogeneous group sizes - even when the network is highly efficient or exhibits a high probability of group imitation.

High efficient network or high group imitation probability tends to produce more homogeneous group behavior (Figure 4-1), although this can vary by nuanced methods of imitation. Results partially correlate with [100] on group patterns that follow with an increase in social learning; knowledge increases but the amount of expressed behavior decreases. Such homogeneity across agents' actions either increases efficiency or risk in group learning depending on the quality of the imitated strategy. Groups exploiting models at a high imitation probability can be susceptible to bias, narrowing focus on previously explored options that are not necessarily optimal. On the other hand, imitation could be an effective information aggregation/consensus mechanism, helping a group overcome individual variations in beliefs and partial knowledge about an objective truth. When agents randomly imitate from a single learning source at each timestep in a very efficient network, the success of group learning is highly dependent on its model (Appendix A-6). Consistent with previous works [70, 14, 85], selective imitation of higher performing demonstrators is conducive for group adaptation. Boyd and Richerson [14] posit that a repeated process– of group average increasing to and exceeding the skill of the most skilled person– enables cumulative cultural adaptation.

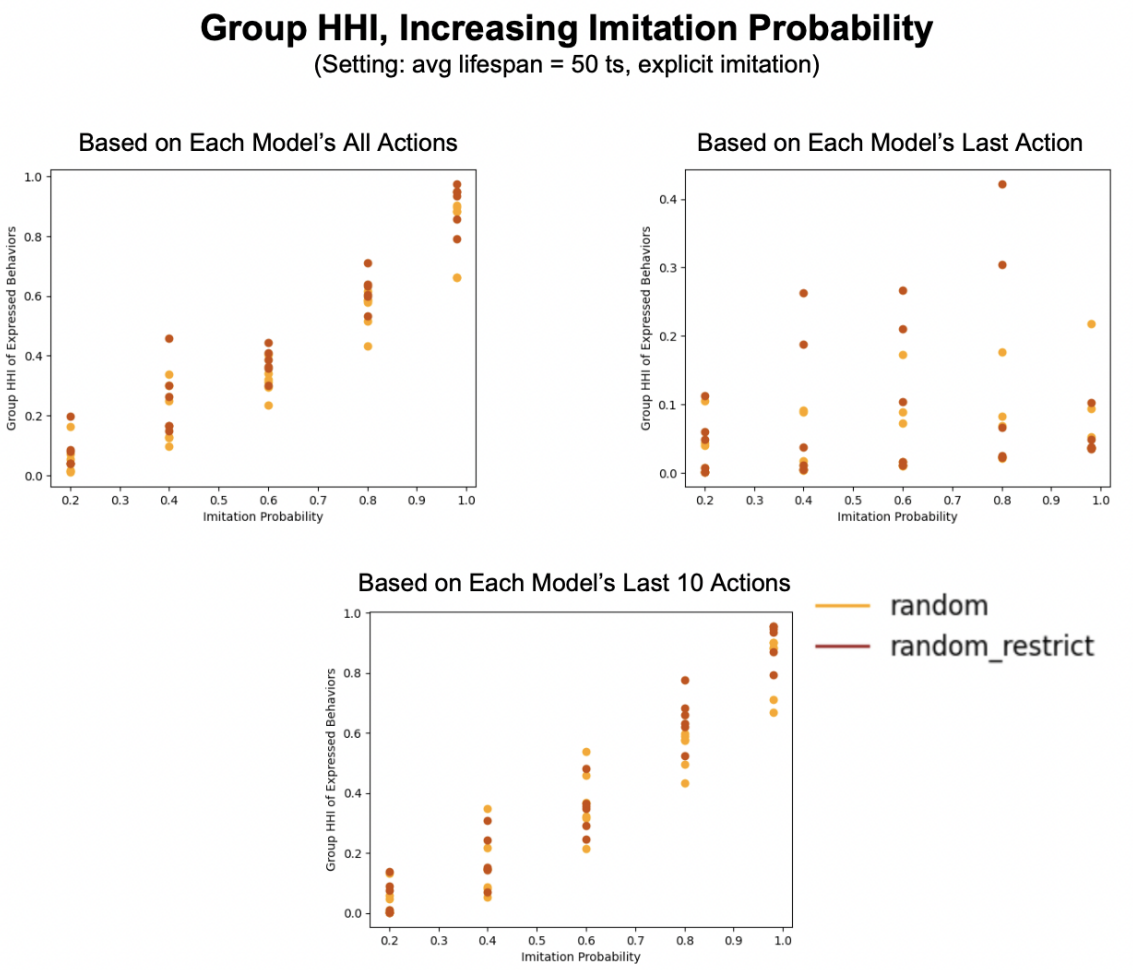

Figure 4-1: (Y-axis) **The level of group action convergence in terms of HHI across the last 2500 timesteps** averaged across trials, each 5K steps long. Higher HHI represents a higher convergence of arms pulled at a collective level. (X-axis) Group imitation probability, or network efficiency.

Observations from (Figure 4-2) suggest that applying an age filter before model selection could be a reliable way to improve the performance of homogeneous unbiased social learners across any degree of network efficiency and group sizes (Figure 5-2). Compared to unbiased imitation without an age filter, unbiased imitation with an age filter improves group performance and performance stability, especially in very efficient networks (e.g., 80%, 98% group imitation probability). The age filter shifts the group's attention to imitating agents with at least some knowledge about the environment, thereby reducing the likelihood of imitating low-quality strategy. In other words, the filter is one simple way to limit the spread of maladaptive culture. Previous work [38, 65, 14, 103] make "costly information hypothesis" and posit that the benefit of group reliance on social information depends on the relative costs of asocial and social learning; results from this chapter suggest that heavy reliance on social learning in fully connected networks can benefit groups in homogeneous environment even without cost assumptions. Unlike other studies that posit a decline in the benefit of social learning (in terms of group fitness) with an increase in the frequency of social learning [14, 103], this study assumes 1) that individual agents are learners throughout their lifespans, continuously exploring and exploiting without being stuck indefinitely at some locally maladaptive behavior, 2) that they are learning in a stable environment, and 3) that these agents are hybrid learners that oscillate between individual and social learning. Results in this chapter complement with [49], which hypothesizes that innovation is powered by cultural interconnectedness, facilitating small additions by multiple contributors and re-combinations of existing ideas and methods.

Even when the group has varied learning abilities (i.e. decision accuracy), these observations remain consistent: unbiased imitation after applying an age filter is more efficient for group learning than asocial learning with a longer learning period, and it generally becomes more effective as group imitation probability or network efficiency increases (See Chapter 5, Figure 5-3). Results are counter-intuitive, given that unbiased social learning is more costly for some individuals when vigilance and domain knowledge are unequal across the population; previous theoretical studies predict collective intelligence when individual learners make accurate decisions most of the times [32] and social information is more reliable [66, 65, 18]. Results suggest that nuanced details in a high-degree of unbiased imitation can help parcel out noise at an individual level and help group problem-solving.

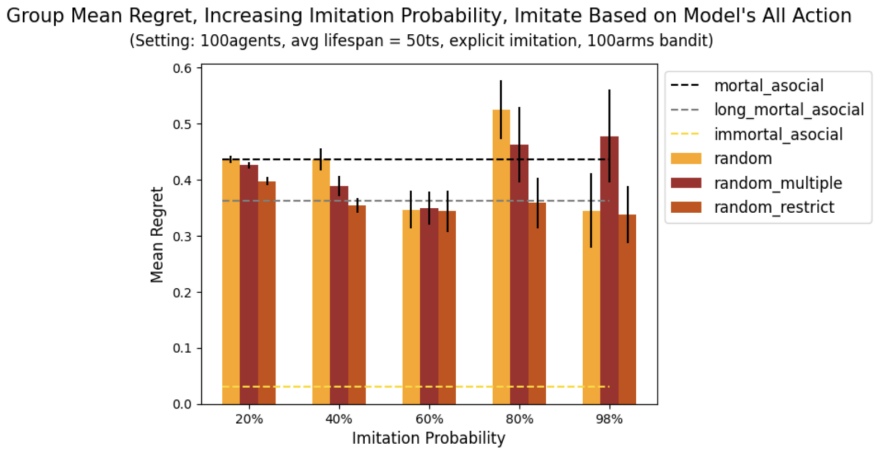

Figure 4-2: (Y-axis) **Group performance in terms of mean regret across the last 2500 timesteps** averaged across trials, each 5K steps long. (X-axis) Group imitation probability, or network efficiency. Greater group performance is represented by lower mean regret. The black dotted line is the performance of asocial learning agents with the same average lifespan of 50 timesteps. The gray dotted line is the performance of asocial learning agents with a longer average lifespan of 250 timesteps. The gold dotted line is the performance of immortal agents (the upper bound of asocial learning)

4.4.2 With an age filter, unbiased learners perform better when they imitate based on their model's complete action history than when they imitate based on their models' recent actions.

In the case of unbiased social learners with an age filter, imitating the most frequent action based on all action history is more effective for group learning than imitating based on recent actions with a time filter (Figure 4-3); the imitation strategy has the strongest comparative edge at high group imitation probabilities. When unbiased learners imitate based on the most frequent action across the model's complete action history, they essentially leverage the maximum amount of its knowledge transmitted and accumulated over time. This can better inform action decisions at an individual level, which could reduce errors cascading throughout the group when imitators aggregate information from random individuals who are not necessarily knowledgeable.

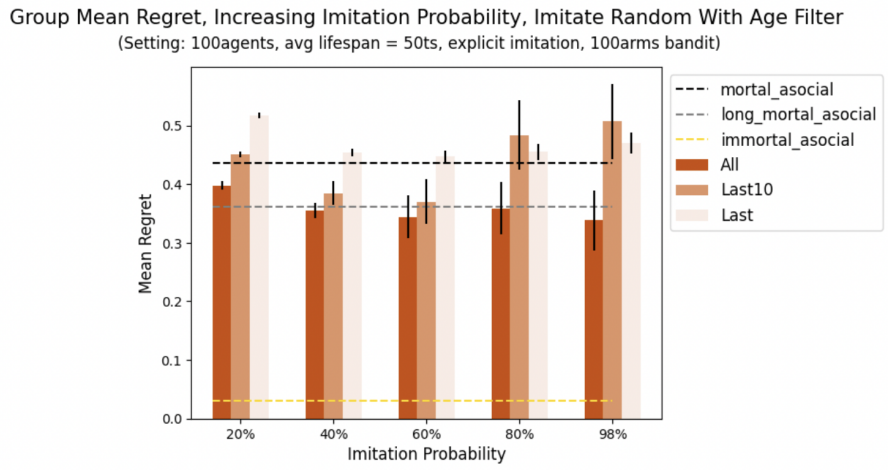

Figure 4-3: *Random imitation with age filter in 100 agents group* (Y-axis) **Group performance in terms of mean regret across the last 2500 timesteps** averaged across trials, each 5K steps long. (X-axis) Imitation probability. The darkest bar indicates the group imitating the most frequent action from a model's complete action history. The lighter bar right next to it indicates the group imitating the most frequent action pulled by the model from the last 10 timesteps. The lightest bar indicates the group pulling the model's last action. The black dotted line is the performance of asocial learning agents with the same average lifespan of 50 timesteps. The gray dotted line is the performance of asocial learning agents with a longer average lifespan of 250 timesteps. The gold dotted line is the performance of immortal agents (the upper bound of asocial learning)

Chapter 5

Comparing Prestige-based Social Learning Strategies

5.1 Overview

The purpose of this section is to conduct a comprehensive assessment of selective imitation in comparison to unbiased imitation. Each learning agent now has a social inductive bias toward prestigious individuals ranked based on some pre-designated metrics. In doing so, this section considers various definitions of prestige, aiming to explore the circumstances under which different forms of prestige bias are most likely to be fit across different group environments and modes of imitation. The section sheds light on the conditions in which selective imitation could be optimized for effective group learning. Results show that (selective) social learning can be a nuanced interaction, and how agents imitate matters; for agents to determine the most favorable course of action, they need a mental model to be tailored to the specific context and objectives of the learning environment.

The choice between imitating based on a model's complete action history or its recent actions for selective imitation in efficient networks depends on the characteristics of the prestige metric. The former strategy is more reliable when imitating the oldest agent, leveraging the maximum amount of experience available. The properties of a mean reward as a metric for skill can risk agents imitating the most skilled agent to exploit sub-optimal

actions; higher scores of the most skilled agent in terms of mean rewards can be due to luck, having done insufficient exploration to bootstrap and transmit the common knowledge. Age filter mitigates this risk, but an arbitrarily filtered pool of models may introduce new risks of limiting attention to models that are relatively lacking both in terms of quantity and quality of experience. Accordingly, when agents imitate based on models' complete action history, ranking agents based on the accumulated reward is a more efficient option than applying an age filter despite both being a combination of reward- and age-based wisdom; skill-biased imitation with an age filter is more effective when agents imitate based on models' recent actions. These observations demonstrate that the effects of combining prestige signals are not always additive in terms of group performance and vary by how agents imitate.

5.2 Experiments

As in Chapter 4, all social learning agents oscillate between imitation and Thompson sampling (asocial learning) based on some pre-determined group imitation probability (imitate_prob).

This chapter operationalizes different definitions of prestige status: age, skill, and accounting for both (age_skill and age_restricted_skill). Social learning groups with age bias would imitate agents with the longest timesteps alive. Social learning groups with skill bias would imitate agents with the highest mean reward. Accounting for both age and skill signals, age_skill bias considers agents with the highest cumulative reward. Social learning groups with age_restricted_skill bias would imitate agents with the highest mean reward that played for longer than 20 timesteps. Finally, selective social learners are compared against unbiased social learners from Chapter 4. A selective social learner will not imitate if the considered model has an equal or lower status than itself (f = timesteps alive for age bias, the mean reward for skill or restricted_skill bias). The rest of the environmental conditions remain the same as in Chapter 4. See Table 5.1 for parameters considered in our experiments.

Table 5.1: Experiment Parameters

Parameter	**Options**	**Description**
[WHO]	oldest, highest mean reward (skill), most skilled after age restriction (age_restricted_skill), highest cumulative reward (age_skill), random, random after age restriction	Model Selection
[HOW]	Imitate based on the model's complete action history, the model's last 10 actions, or its last action	Time-frame
	Explicit action-based imitation, Implicit belief-based imitation with or without filter	Imitation Type
[WHEN]	{0.2, 0.4, 0.6, 0.8, 0.98}	Imitation Probability (Network Efficiency)
	All agents observe true rewards, Beta(1,1), Beta(0.5,0.5)	Distribution of Population Learning Ability
	7 agents, 50 agents, 100 agents	Group Size
	100 arm bandit	Arm Size

5.2.1 Social Learning Policy

This chapter uses the same social learning algorithm from Chapter 4 [1], except now additional prestige signals are considered for status_type, and imitated models are selected based on who has the maximum score in terms of a chosen prestige metric. If there is more than one agent at the highest rank, the model is randomly selected from these agents. An agent only considers imitation if its current status (status_a) is not the same as the highest status among all agents. See Appendix 2 for its pseudo-algorithm.

Unlike Chapter 4, this chapter now considers two approaches to imitation and choosing an arm: explicit and implicit imitation. As in Chapter 4, explicit imitation involves selecting the most frequent action taken by the model during some period of time (See Equation 4.1). In contrast, implicit imitation involves forming temporary beliefs about the bandit based on the model's prior actions and then using Thompson sampling for decision-making. Implicit imitation can be categorized into two variations: one with a filter and another without (See Equation 5.1). In the case of implicit imitation with a filter, an agent not only assigns positive weights to the model's observed past actions but also puts greater negative weights to actions not taken by the model, with the weights being linearly correlated to the model's age.

[1] As in Chapter 4, asocial learners use Thompson Sampling (See Section 3.3.1) to select and pull an arm.

Forming Temporary Beliefs in Implicit Imitation Without Filter

For each arm i where i is in the set of actions A_t taken by the agent at time t, update the success count α_i as follows:

$$\alpha_i \leftarrow 1 + w \cdot n_i \tag{5.1}$$

where:

α_i is the updated success count for arm i in the Beta-Bernoulli bandit (See Section 4.2.1)

A_t represents the set of actions taken by the agent at time t.

w is the constant weight (model age) assigned to the model's observed past actions.

n_i is the number of times arm i has been chosen up during some time frame T.

and for each arm i, the posterior distribution is given by

$$\hat{p_{i,t}}|a_{1:t-1} \sim \text{Beta}(\alpha_i, 1) \tag{5.2}$$

and

$a_{1:t-1}$ is a complete history of actions pulled by model a

from which each imitating social learner conducts Thompson Sampling.

Forming Temporary Beliefs in Implicit Imitation With Filter

For each arm i where i is in the set of actions A_t taken by the agent at time t, update the success count α_i and the failure count β_i as follows:

$$\alpha_i \leftarrow 1 + w \cdot n_i \tag{5.3}$$

$$\beta_i \leftarrow \begin{cases} w_t & \text{if } n_i = 0 \\ 1 & \text{if } n_i \neq 0 \end{cases} \tag{5.4}$$

where:

α_i is the updated success count for arm i in the Beta-Bernoulli bandit (See Section 4.2.1)

β_i is the updated failure count for arm i in the Beta-Bernoulli bandit

A_t represents the set of actions taken by the agent at time t.

w is the constant weight (model age) assigned to the model's observed past actions at time t.

n_i is the number of times arm i has been chosen up to time t.

and for each arm i, the posterior distribution is given by

$$\hat{p_{i,t}} | a_{1:t-1} \sim \text{Beta}(\alpha_i, \beta_i) \tag{5.5}$$

and

$a_{1:t-1}$ is a complete history of actions pulled by model a

from which each imitating social learner conducts Thompson Sampling.

5.2.2 Evaluation

The same set of evaluation metrics from Chapter 4 is used. See Section 4.3.2.

5.3 Results and Discussion

5.3.1 Selective imitation among homogeneous agents is efficient for group learning, even at very high levels of imitation probability.

For the first few generations of mortal, adaptive imitators with limited learning periods can learn faster than immortal, asocial learners; the duration of its comparative edge depends on nuanced methods of imitation. Sometimes, these adaptive imitators can be unbiased.

After a few generations, mortal learners' performance stagnates, and immortal asocial learners with unlimited learning periods come to perform better over a long horizon (Figure 5-1). However, when the task complexity is low enough (e.g., low number of bandit arms), adaptive imitators continue to do comparably, or even better than immortal asocial learners over very long horizons (Appendix A-5).

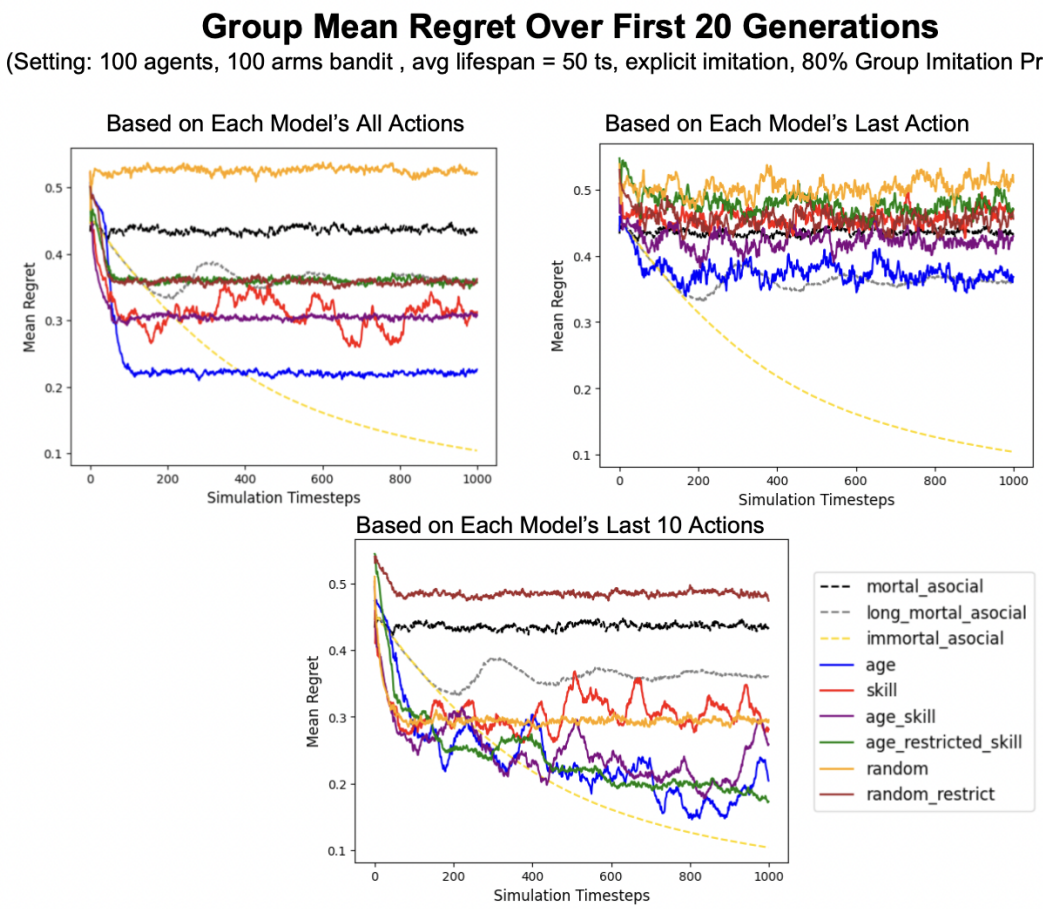

Figure 5-1: *Counter-clockwise:* Imitating Based on Models' Complete Action History, Actions From Last 10 Timesteps, and Last Action. (Y-axis) **Group performance in terms of mean regret, within 100 agent network with 80% group imitation probability, facing 100 armed bandit** (X-axis) **Across the first 1000 timesteps**, averaged across trials (each 5K steps long). The black dotted line is the performance of asocial learning agents with the same average lifespan of 50 timesteps. The gray dotted line is the performance of asocial learning agents with a longer average lifespan of 250 timesteps. The gold dotted line is the performance of immortal agents (the upper bound of asocial learning)

The rest of this section compares the average group performance across the last few generations within each simulation trial as in Section 4.4.1. Like unbiased imitators with age filters, selective imitators learn more efficiently than asocial learners with a more extended but still limited learning period. The relative advantage of selective imitation, even within efficient, highly connected networks, is consistently observed across group sizes and task complexity (Figure 5-2).

However, experimental results from this chapter show no strong and consistent effect of group imitation probability across all imitation strategies considered (Figures 5-1 and 5-5). More specifically, while the efficiency of selective imitation among homogeneous agents generally improves with increased network efficiency or group imitation probability, the threshold for reverse trend varies across imitation strategies. When models are chosen based on age filter and mean reward and agents imitate based on the model's complete action history, agents perform better in networks with moderate (e.g. 40%, 60% imitation probability) than in networks with very high (e.g. 98% imitation probability). Agents with this type of model bias have a more limited pool of models and hence have a limited diversity of social information relative to other social learning strategies (more discussed later in Section 5.3.2). Moderate imitation probability or network efficiency could facilitate building a more diverse knowledge repertoire per individual agent, which helps with group problem-solving [34]. These results suggest that network structure by itself is insufficient to predict group success in developing solutions.

Previous studies, including agent-based simulations and behavioral MTurk experiments, found seemingly contradictory effects of network structure on group problem-solving abilities. Experiments in this thesis stand in contrast to these studies that consider "rugged" landscape where solutions can get stuck at locally optimal but globally sub-optimal points; this thesis considers bandits where every action is technically accessible, whatever actions agents take previously. Some find evidence supporting the superiority of efficient networks [77] while some support the opposite as superior [71]. The takeaway from results in this chapter is most aligned with that of [10]. The study found that both levels of network connection can help groups facing identical types of problems in terms of "ruggedness"; the conformist strategy was effective in efficient networks, while it was best to follow the best member in less efficient networks. What matters is the building blocks of group learning or

social learning strategies at an individual level.

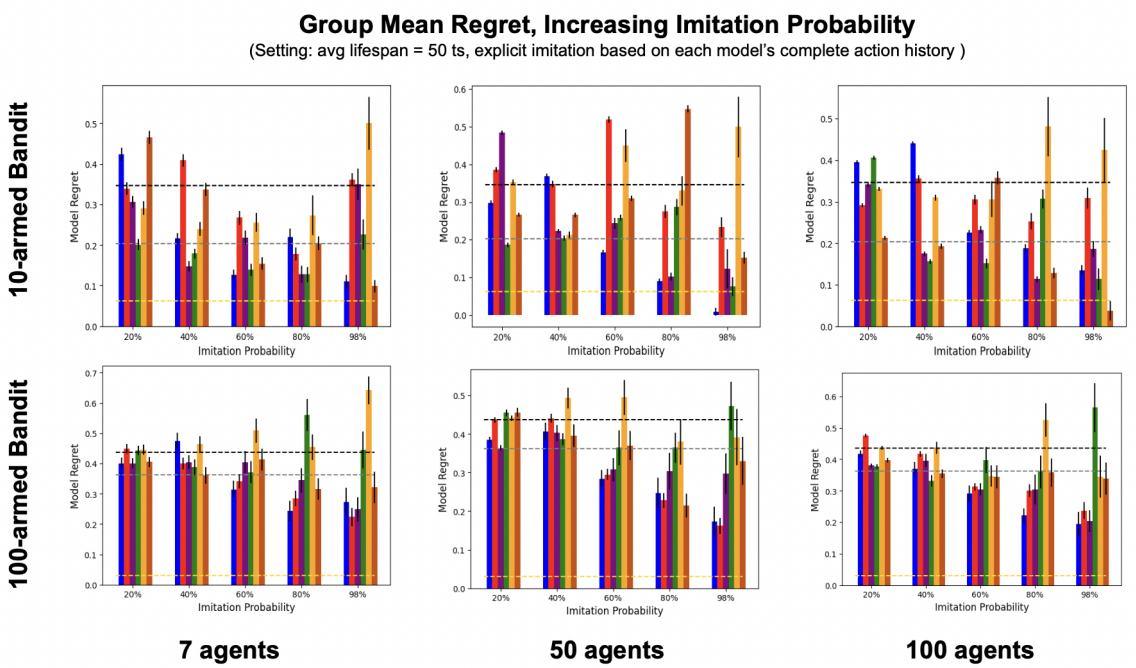

Figure 5-2: *Imitating Based on Models' Complete Action History Across group sizes (horizontal) and Task Complexity (vertical).* (Y-axis) **Group performance in terms of mean regret across the last 2500 timesteps** averaged across trials, each 5K steps long. (X-axis) Network with 20%, 80%, and 98 % group imitation probability where agents have varied learning abilities (i.e. decision accuracy). The black dotted line is the performance of asocial learning agents with the same average lifespan of 50 timesteps. The gray dotted line is the performance of asocial learning agents with a longer average lifespan of 250 timesteps. The gold dotted line is the performance of immortal agents (the upper bound of asocial learning)

Again, results remain consistent even when the group has varied learning abilities or when agents' quality of personal information varies as a result: Across different group imitation probabilities, choosing to selectively imitate one model per time step, guided by a prestige metric, proves to be a more efficient approach compared to engaging in asocial learning over an extended duration (see Figure 5-3). It seems even more true when all learners possess equivalent and accurate learning capabilities. Results suggest that model-based biases become relatively more important when skill differences exist among agents. Consistent with previous research [66, 98, 30], a minority of informed individuals could be sufficient to guide less-informed members and lead the group to high performance.

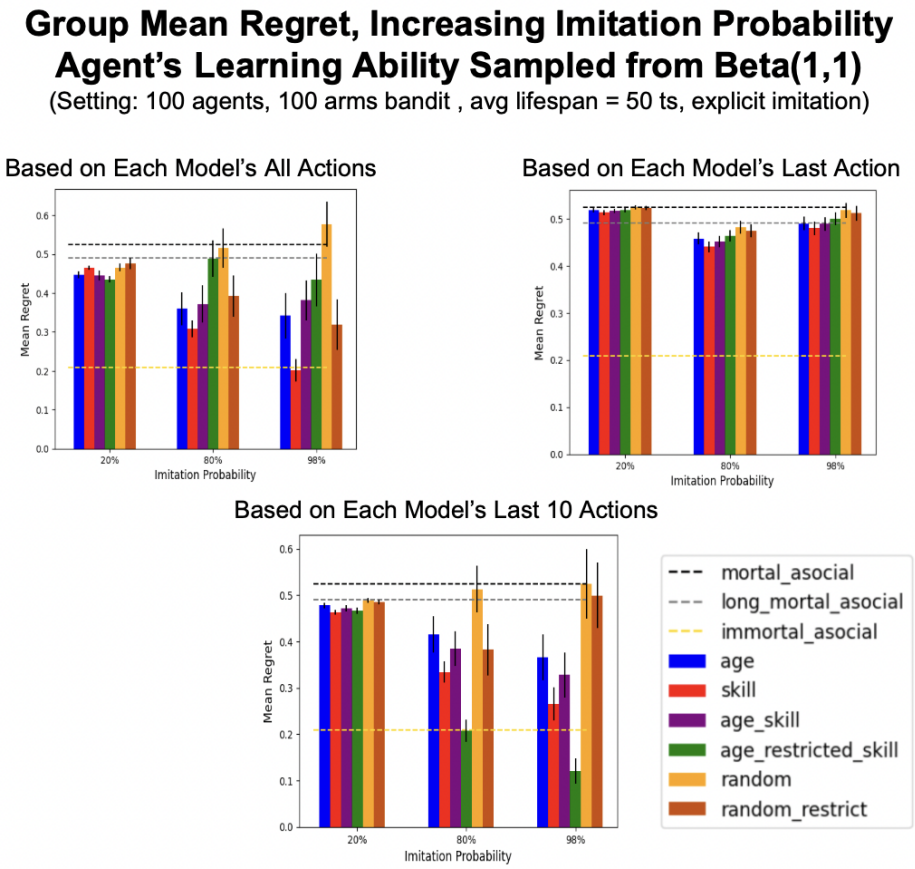

Figure 5-3: *Agents' learning abilities were sampled from Uniform Distribution (Beta(1,1)): Counter-clockwise: Imitating Based on Models' Complete Action History, Actions From Last 10 Timesteps, and Last Action.* (Y-axis) **Group performance in terms of mean regret across the last 2500 timesteps** averaged across trials, each 5K steps long. (X-axis) Network with 20%, 80%, and 98 % group imitation probability where agents have varied learning abilities (i.e. decision accuracy). The black dotted line is the performance of asocial learning agents with the same average lifespan of 50 timesteps. The gray dotted line is the performance of asocial learning agents with a longer average lifespan of 250 timesteps. The gold dotted line is the performance of immortal agents (the upper bound of asocial learning)

5.3.2 To effectively leverage the strength of a model for group learning in highly efficient networks, agents' imitation strategies need to be adaptive to model characteristics.

For any model bias, imitating the last action of the model is not effective for group learning across all experiments; this is not surprising – agents are learning from other learners who are also navigating the environment and balancing exploration and exploitation. Learners, especially prestige-biased imitators, should pay attention to a longer history of models' actions to perform well in highly efficient networks. Results shed light on the importance of looking into how agents imitate- agents need a mental model to understand and determine the most favorable course of action or decision in a given situation [41, 117].

Age-biased Imitation

When agents imitate models based on just age in Figure 5-4, imitating the most frequent action based on all action history is more effective than asocial learning. The imitation strategy is also more reliable than imitating based on recent actions (time filter) across group imitation probabilities. The strength of imitating the oldest agent is its amount of experience accumulated with age; imitating the most frequent action based on the model's complete action history fully leverages the size of models' knowledge pool at an individual level.

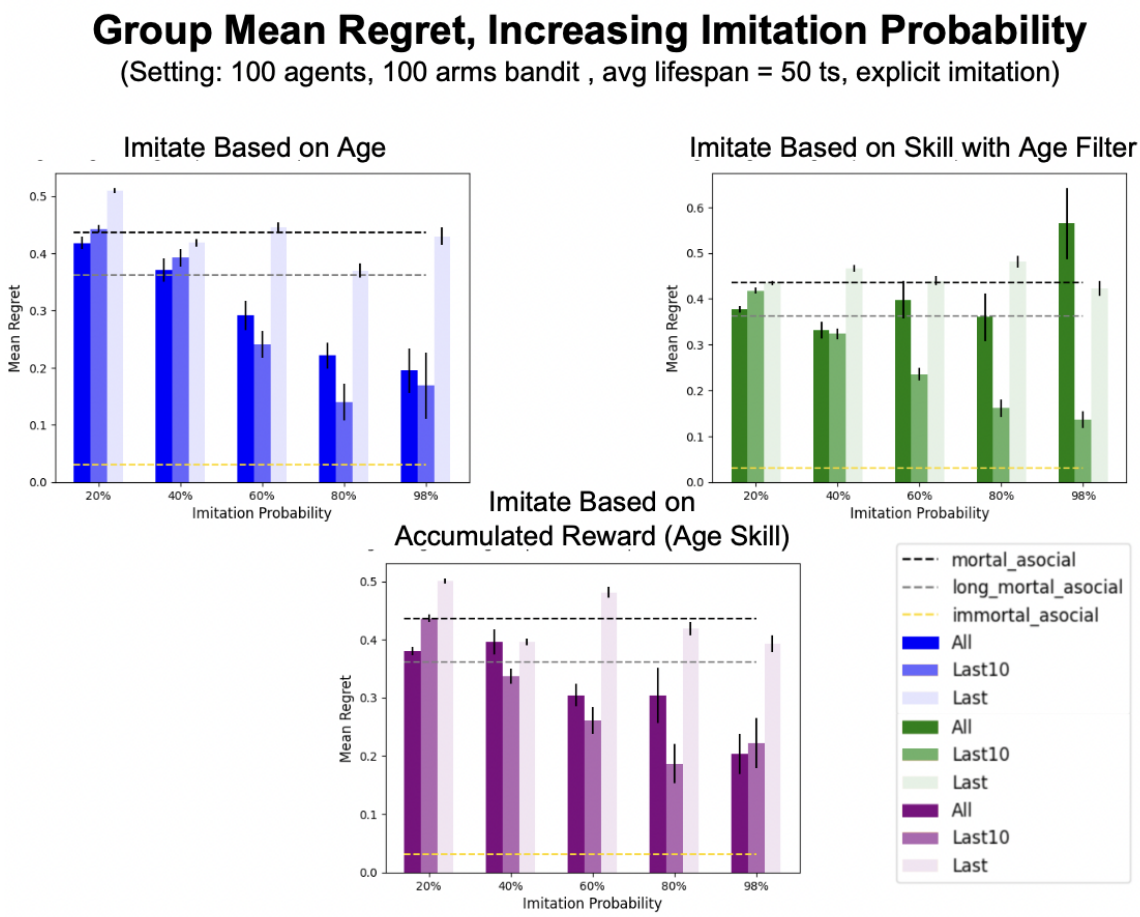

Figure 5-4: *Clockwise from the upper left corner: Explicitly imitating the oldest agent (Age), agent with the highest mean reward after applying age filter (Skill with Age Filter), and agent with the highest cumulative reward (Age Skill)* (Y-axis) **Group performance in terms of mean regret across the last 2500 timesteps** averaged across trials, each 5K steps long. (X-axis) Imitation probability. The darkest bar indicates the group imitating the most frequent action from a model's complete action history. The lighter bar right next to it indicates the group imitating the most frequent action pulled by the model from the last ten timesteps. The lightest bar indicates the group pulling the model's last action. The black dotted line is the performance of asocial learning agents with the same average lifespan of 50 timesteps. The gray dotted line is the performance of asocial learning agents with a longer average lifespan of 250 timesteps. The gold dotted line is the performance of immortal agents (the upper bound of asocial learning)

Skill-biased Imitation

When agents use skill as a signal to choose and imitate models, imitating the most frequent action based on the model's recent actions is more reliable and adaptive strategy than imitating based on the model's complete action history (Figures 5-4 and 5-5). This may imply that the strength of models chosen based on skill is rooted more in the quality of recent experience.

However, results (Appendix A-6 and A-7 on the left) provide evidence that just skill as a signal may distort models' actual knowledge level, inadvertently deceiving the group to learn from sub-optimal agents without knowing. With an adaptive imitation strategy, skill-biased imitators with age filters can learn more efficiently than those without. Here, the age filter prevents the group from imitating agents who have less experience and likely did not explore sufficiently but still earn high average rewards from luck.

To highlight once again, it is important to consider the interactive effect of who and how to imitate on group performance. Even with age filters, these learners can be less effective for group learning than unbiased imitators if they follow maladaptive imitation strategies (i.e., imitating the most frequent action based on the model's complete action history). In these groups, learners chosen as models are older than those chosen just based on average reward. Yet, the group is not necessarily imitating the oldest agent who has the maximum quantity of experience from which imitators can leverage the model's entire action history (Appendix A-7 on the left). Adding an age filter in skill bias also reduces the number of potential models to be imitated across time spans (Appendix A-7 on the right). In contrast, unbiased imitators (even with age filters) imitate from a larger sample of models, which means more variation in social information; as a result, groups with unbiased learners have a relative advantage in exploring a broader range of actions and expanding the knowledge pool at a collective level. Imitating based on models' complete action history may be adaptive only when agents imitate from models with absolute quantity or diversity of experience; agents with the highest mean reward do not exhibit such characteristics.

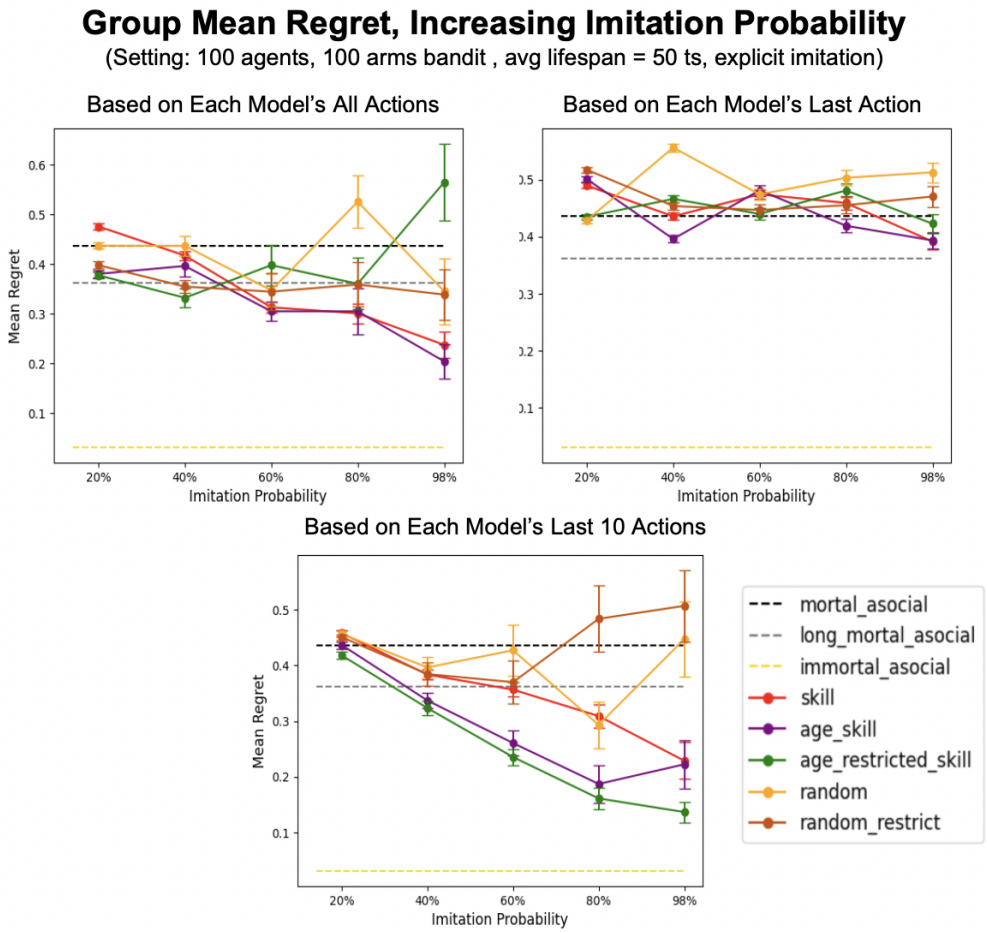

Figure 5-5: *Left to right: Explicit imitating groups based on the model's complete action history, action history from the last ten timesteps, and last action.* (Y-axis) **Group performance in terms of mean regret across the last 2500 timesteps** averaged across trials, each 5K steps long. (X-axis) Group imitation probability, or network efficiency. Greater group performance is represented by lower mean regret. The black dotted line is the performance of asocial learning agents with the same average lifespan of 50 timesteps. The gray dotted line is the performance of asocial learning agents with a longer average lifespan of 250 timesteps. The gold dotted line is the performance of immortal agents (the upper bound of asocial learning)

5.3.3 Imitation Strategy: Independent Belief Formation with Weighted Action Trajectories

Unlike explicitly imitating agents, implicitly imitating agents can perform comparably to, but generally not better than, asocial learners with a longer average lifespan. The observation implies that implicit, belief-based imitation is less effective for group learning than explicit imitation, which is more action-based. This could be because models imitated are learners themselves, and agents in these experiments do not have direct access to models' actual beliefs, as often seen in reality. Randomized probability matching with beliefs approximated based on models' actions could lead imitators to pull arms never sufficiently explored by the model, given the algorithm's tendency to balance exploration and exploitation. In contrast, explicit action-based imitation can provide more accurate information about what arms a model exploits.

Nonetheless, with nuanced changes in how agents imitate, implicit imitation can be more effective than asocial learning with the same average learning period; groups following age-based implicit imitation with filters are more efficient in learning than asocial learners with the same average lifespan. Groups with filters consider both actions (not) taken by the model when forming beliefs about the reward environment. Results (Figure 5-6) suggest inactivity provides useful additional information for effective imitation, assuming agents have access to the complete action space of the model.[2] However, individual models' amount of experience seems to be a crucial requirement for implicit imitation to be helpful in group learning. Models selected for having the maximum mean reward can fall short in this regard (as discussed in the previous section 5.3.2), making the given imitation strategy maladaptive for skill-biased learners.

[2]In more complex environments, it can quickly become impractical to consider every possible situation, and information about situations that did not occur may not be beneficial for learning.

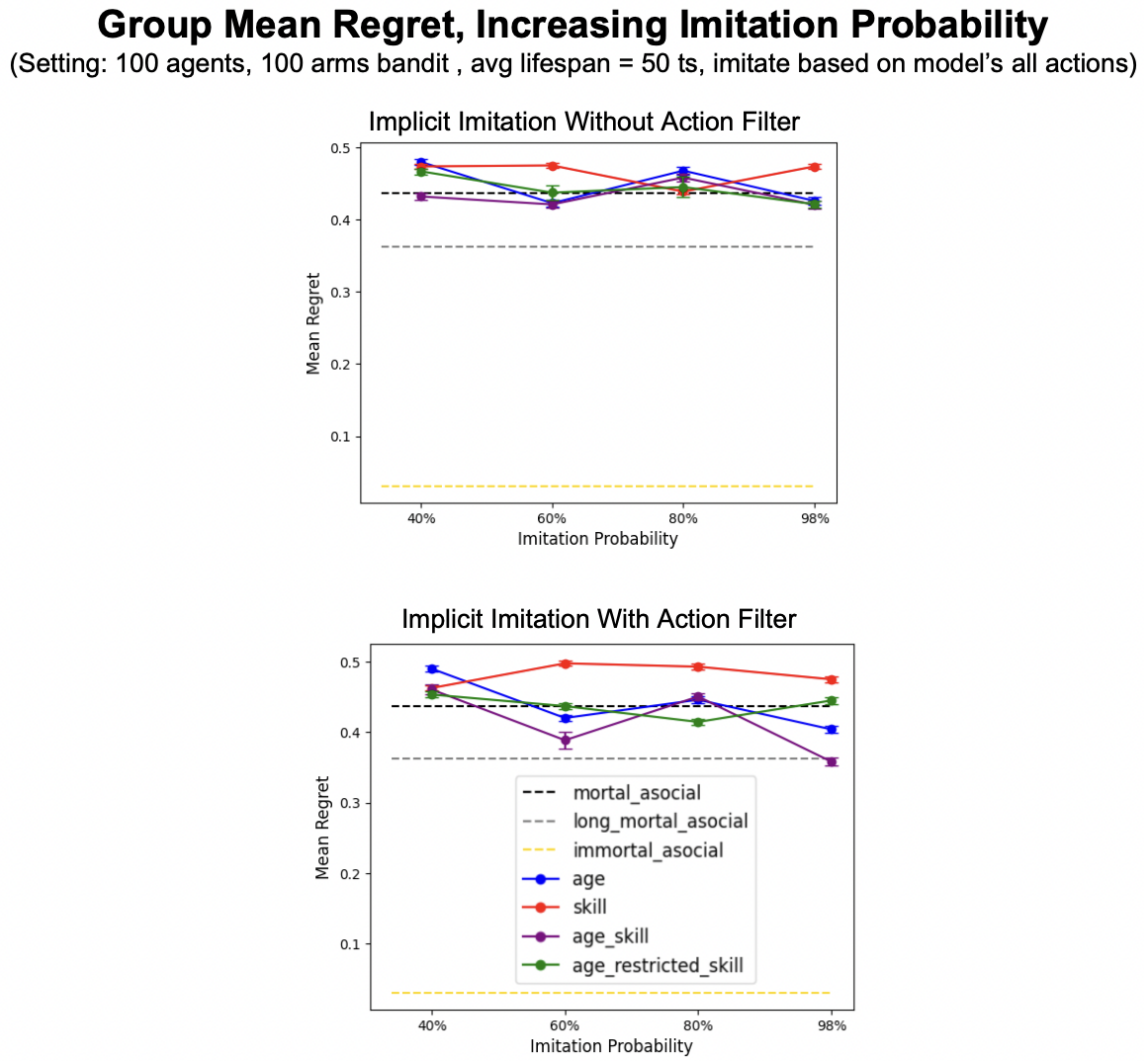

Figure 5-6: *Left to right: Implicitly imitating groups without (left) and with (right) filter.* (Y-axis) **Group performance in terms of mean regret across the last 2500 timesteps** averaged across trials, each 5K steps long. (X-axis) Group imitation probability, or network efficiency. Greater group performance is represented by lower mean regret. The black dotted line is the performance of asocial learning agents with the same average lifespan of 50 timesteps. The gray dotted line is the performance of asocial learning agents with a longer average lifespan of 250 timesteps. The gold dotted line is the performance of immortal agents (the upper bound of asocial learning)

Chapter 6

Multi-Agent Linear Thompson Sampling and Prestige-based Imitation in Small Groups

6.1 Overview

This chapter examines a more complex informational environment and offers preliminary findings that validate discussions from previous chapters; we now consider **Linear Thompson Sampling agents in multi-agent contextual bandits settings.** To continue exploring the effect of selective imitation across different learning environments, this chapter examines a small group of 7 agents navigating a 100-armed contextual bandit game. The task is now to approximate μ^* that determines rewards for each arm together with its context; μ^* could be considered as some shared norm, which agents try to learn.

The method of imitation is different from Chapters 4 and 5 in that imitating agents faithfully adopt an observed model's belief ($\widehat{\mu_i(t)}$), or estimate of μ^* about the linear contextual bandit. This method of belief-based imitation is more faithful and accurate than the implicit imitation method in Chapter 5 where models' beliefs are independently assumed by learners based on models' actions within some arbitrary time frame. Beliefs adopted as a method of imitation in this chapter are models' actual beliefs

about the model updated and accumulated across their entire history.

Results from this chapter again show that imitating others' behaviors and observing corresponding rewards is helpful in a more challenging learning environment, though who, when, and how they imitate matters altogether. Adaptive learning contexts, given access to model beliefs formed based on their entire action history, are similar to what was observed in previous chapters, where agents performed action-based imitation and chose the most frequent action across the model's full action history. The performance of agents imitating the oldest agent at each timestep is robust across imitation probabilities. Skill-biased social learners outperform unbiased counterparts until network efficiency reaches an extreme level. As in previous chapters, preliminary findings from this chapter also show that group performance is intricately linked to individual model performance; in highly efficient networks with a high group imitation probability, model dependency becomes more pronounced, which can be a critical risk to group performance. In summary, results from this chapter highlight that the effectiveness of belief-based imitation in the form of faithfully adopting other learners' beliefs about the environment ($\widehat{(\mu_i(t))}$) hinges on the quantity/diversity of model's experience, with the oldest agents holding a competitive edge in this regard.

6.2 Environment

6.2.1 Linear Contextual Bandit

In our contextual bandit environment, $\mu^* \in \mathbb{R}^d$ is shared across all agents where $d = 5$. As it maps the observed context of a chosen action to expected rewards, we consider the parameter of interest, μ^*, as an underlying function or preference vector within some information environment. μ^* could represent a social rule or ground truth in our society or a functional rule required to demonstrate a complex skill or use of specific technologies. For our project, we only consider stationary contextual bandits where μ^* remains fixed across timesteps.

At each time step t, each agent independently updates beliefs about μ^*, or $\widehat{\mu_i(t)}$); accordingly, each learning agent independently chooses an arm to pull. A group of agents share and selects from the same set of K arms represented by context vectors, $A_t = \{x_{1,t}, x_{2,t}, ..., x_{k,t}\}$ where $x_{k,t} \in \mathbb{R}^d$ and $K = 100$. The vector contains information about the context of action k. Contexts are fixed across time steps across timesteps.

A reward is the most true signal from the environment about agents' performance. $r_{k,t}$ is the reward associated with each arm k (or its context $x_{k,t}$) at timestep t. Each agent receives and seeks to maximize its total expected discounted future reward, operationalized as negative regret. The reward is not dependent on the actions of other agents.

We assume that the contexts are independent and identically distributed (i.i.d.) and that there is a linear structure in the function of rewards. We place a Uniform prior on μ^*, which is our parameter of interest. Context vectors and μ are then normalized. Reward likelihood is assumed to be Gaussian.

$$x_k \sim \mathcal{U}(-1_d, 1_d) \tag{6.1}$$

$$\mu^* \sim \mathcal{U}(0_d, 1_d) \tag{6.2}$$

$$r_{k,t}|\mu^*, X \sim N(<\mu^*, X>, \sigma^2) \tag{6.3}$$

Asocial learners use Linear Thompson Sampling (See Section 3.3.2) to select and pull an arm.

6.3 Experiments

As in Chapters 4 and 5, all social learning agents oscillate between imitation and Thompson sampling (asocial learning) based on some pre-determined group imitation probability (imitate_prob).

See Table 6.1 for parameters considered in our experiments.

6.3.1 Social Learning Policy

As in Chapter 5, imitated models are selected based on who has the maximum score in terms of a chosen prestige metric. If there is more than one agent at the highest rank, the model is chosen randomly from these agents. Social learners imitate with f = 20%, 40%,60%,80%,98% probability, which represent different levels of network efficiency. An agent will not imitate

Table 6.1: Experiment Parameters

Parameter	Options	Description
[WHO]	oldest, highest mean reward (skill), most skilled after age restriction (age_restricted_skill), highest cumulative reward (age_skill), random	Model Selection
[HOW]	Imitate based on the model's last 10 actions	Time Frame
	Explicit belief-based imitation per individual policy (i.e., Adopt $\widehat{\mu_i(t)}$ for timestep t)	Imitation Type
[WHEN]	{0.2, 0.4, 0.6, 0.8, 0.98}	Imitation Probability (Network Efficiency)
	7 agents	Group Size
	100 arm bandit	Arm Size

if its current prestige rank is the highest rank among all other agents; this condition is not applied in the case of random imitation. Social learners who imitate at timestep t do so by temporarily adopting and selecting an arm based on the model's belief $\widehat{\mu_i(t)}$. Social learners who are not imitating at timestep t choose an arm based on Linear Thompson sampling [114].

6.3.2 Evaluation

The same set of evaluation metrics from Chapters 4 and 5 is used. See Section 4.3.2.

6.4 Results and Discussion

6.4.1 Imitators perform better than asocial learners

Asocial agents with a longer average lifespan of 250 timesteps do not perform any better than those with a shorter average lifespan of 50 timesteps, meaning that information accumulation and exploration-exploitation problems are now more difficult in a contextual bandit setting, and they cannot simply be compensated by a longer learning period within a limited lifespan alone.

Imitation from one source is an effective mode for overcoming this problem though model biases and learning environments do matter. Given that models imitated in a group of unbiased learners are selected at random, model performance among unbiased learners could be considered as the average knowledge of an individual. As observed in Chapter 4, a high model reward (Bottom of Figure 6-1) in this environment implies that imitation can be

beneficial not only for cumulative shared welfare but also for individual agents.

6.4.2 Imitation-based group learning can be effective in efficient networks or very high levels of imitation probability, provided that agents have a model bias that is adaptive to their group imitation strategy.

Now facing a more complex task, groups generally perform better in settings with moderate group imitation probability, especially when imitated models are ranked based on skill. Such observation is consistent with [71, 10], which observed partially connected network structures [1] to be beneficial for groups facing complex problem-solving task in the long run. Network inefficiency forces exploration and reduces the likelihood of converging to sub-optimal solutions. Again, results suggest the importance of understanding the interactive effects of network efficiency with individual-level social learning strategies, which lead to different exploration-exploitation patterns.

Nonetheless, groups can still perform well in close to fully connected networks if individuals use adaptive imitation strategies. As with implicit imitation in Section 5.3.3, the amount of experience it has accumulated matters when an agent i updates its belief about the environment ($\widehat{(\mu_i(t)}$). Assuming that an agent observes accurate reward at each timestep, the more arm-reward pair it observes, the better it will approximate $\widehat{\mu_i(t)}$. Accordingly, the model's amount of experience matters for group learning in efficient networks where group imitation probabilities are high and model dependency is high.

Age-based Imitation

We observe that groups imitating the oldest agent perform best among social learners with different types of model biases; group mean regrets are lower across group imitation probabilities. As seen in section 5.3.2, age bias seems to be adaptive when agents imitate based on models' complete action histories, now represented by adopting models' beliefs (about μ^*) as it is continuously updated across the model's lifespan. Again, group regret inversely correlates with model reward. Oldest agents in age-biased homogeneous groups of learners demonstrate robust performance across changes in group imitation probabilities.

[1] Moderate network efficiency here is equivalent to moderate speed to which information about potential solutions diffuse throughout the network.

Skill-based Imitation

Skill-biased social learners perform better than unbiased social learners before their performance crashes when group imitation probabilities reach 80%-98% (i.e. when their network becomes highly efficient). Adopting the beliefs of learners with the highest mean reward is ineffective for learning at an individual level as well (Bottom of Figure 6-1). In the current learning context, the quality of the model's knowledge is dependent on its absolute quantity of (positive and negative) experience, hence skill-based imitation may be ineffective; based on our operationalization of skill, agents with the highest mean reward can be due to luck and lack an absolute amount of experience. As seen in Section 5.3.2, unbiased imitation has a relative advantage in this regard compared to skill-based imitation. Unbiased social learners will explore a broader range of models with varied age, or amounts of accumulated experience.

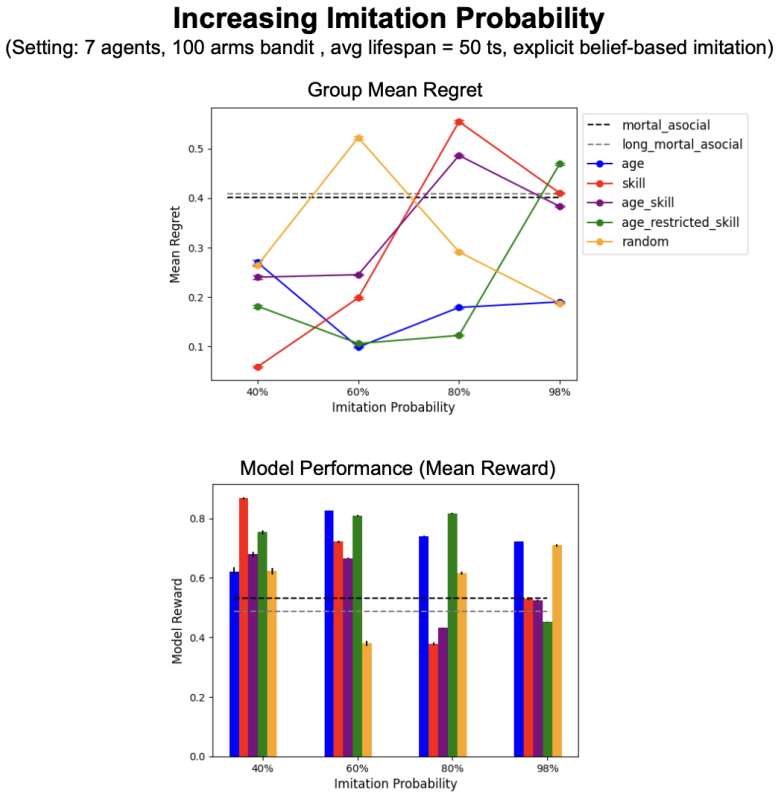

Figure 6-1: *Top:* (Y-axis) **Group performance in terms of mean regret across the last 2500 timesteps** averaged across trials, each 5K steps long. (X-axis) Group imitation probability, or network efficiency. Greater group performance is represented by lower mean regret. The black line is the performance of asocial learning agents with the same average lifespan of 50 timesteps. The gray line is the performance of asocial learning agents with a longer average lifespan of 250 timesteps.; *Bottom:* (Y-axis) **The performance of imitated models in terms of mean reward across the last 2500 timesteps** averaged across trials, each 5K steps long. Higher model rewards represent greater group performance. (X-axis) Group imitation probability, or network efficiency.

Chapter 7

Extended Discussion and Conclusion

7.1 Implications

Researchers can use multi-agent learning to formulate and test hypotheses when analyzing technology in relation to society. For example, the underlying objective of the U.S. Copyright law is to promote "the progress of science and useful arts". Science is essentially a collective search task, and innovation is "fundamentally a cultural evolutionary process" [49, 86]. [49] shares a theoretical model along with ethnohistorical and archeological cases to demonstrate that innovation can be created just with "incremental additions, recombinations, and lucky errors" without individuals contributing novel inventions; yet, he also posits that the flow of additions and recombinations can only be maximized when everyone openly shares their ideas and inventions (i.e., high degree of interconnectedness). Future work could potentially use multi-agent learning to explore how the efficient pooling of existing cultural knowledge by a central artificial agent like generative foundation models impacts innovation, measuring creative opportunity costs in terms of regret. The central artificial agent could generate or recommend an arm (representing content, ideas, or inspirations) from which other individual agents observe to make individual decisions. Experiments could then add and test the effect of additional socio-technical conditions that may influence group outcomes. Current tensions within the creative community on using AI as a tool versus human replacement provide another example demonstrating various modes of social learning. While experiments in previous chapters considered hybrid social learning (oscillating between individual learning

and imitation across timesteps), agents could perform non-hybrid social learning where they always adopt recommendations generated by the central agent instead of independently learning through experience. Experiments could also discount rewards to represent the lack of adequate compensation for artists' work from which these models have imitated and used as engines to power their generative capabilities. Following decreased incentives for artists to share and upload products of experiential learning online [1], lower imitation probability could represent the change in network efficiency or interconnectedness.

Nonetheless, it is important to exercise caution when utilizing multi-agent learning to inform high-impact decisions:

Simulation designs are contingent upon assumptions about the workings of technology and human behavior when interacting with technology. Before running simulations, further user research is needed to investigate how average individuals use or exploit the technology and its context. While experiments in this thesis considered a task with a single objective among homogeneous agents, humans and organizations are highly heterogeneous with varying needs; having algorithmic solutions that work on average does not mean they will be useful for sub-populations [45]. One could easily focus on the dominant user groups at the expense of overlooking marginalized populations. Faulty modelling premises that inform system design decisions can disproportionally affect humans [92, 29].

Furthermore, the type of multi-agent models should be carefully considered. While traditional agent-based models considered in this thesis operate on pre-defined, fixed behavioral rules to simulate learning agents, an increasing number of works are considering alternative computational tools; exploring whether or not large language models (LLMs) learn human-like traits and biases, these works propose LLMs as a tool to simulate (human) agents [5, 6, 95]. While LLMs produce text based on statistical patterns extracted from extensive human linguistic data, the proprietary "black box" nature of LLM poses difficulties for researchers in testing and assessing the underlying mechanisms behind collective outcomes and replicating results. Indiscriminate use of LLMs may lead to drawing conclusions about human society that are ungrounded in reality [106]. Again, a recent study [7] also suggests WEIRD-in, WEIRD-out problems in LLM outputs where LLM's average responses

[1] To avoid having their artworks trained by generative foundation models, artists are actively pulling down their works and even removing their accounts from online art platforms (Appendix A-1)

are biased towards behaviors in Western, Educated, Industrialized, Rich, and Democratic (WEIRD) countries.

Last but not least, care is required when one translates simulation results to support decisions aimed at achieving relevant but often abstract policy goals. There is yet an absence of comprehensive measures that define human creativity and innovation. While multi-agent learning can highlight potential collective benefits or harms of a design choice before deployment, researchers translating simulation results to high-impact decisions must be wary about the potential to change behaviors on the basis of policy considerations.

7.2 Limitations

Models are predictive; in essence, they predict emergent group outcomes grounded in assumptions about specific learning circumstances. They are also abstractions of reality; much predictive power can be lost by (unknowingly) simplifying important details and abstracting. While experiments in this thesis shed light on the importance of the types and frequencies of social learning strategies, they assume a single learning strategy per group instead of a population with mixed strategies. Despite yielding quantified variances in diverse learning parameters, simple simulations like the ones considered in this thesis will almost certainly fail to capture the complete reality of increasingly agentic social systems at scale [25]. Findings will need additional validation through empirical, large-scale human studies before implementing these social learning strategies as part of the system design to mediate social interactions and coordinate group behaviour. For instance, experiments from previous chapters operate under the assumption that all agents have direct access to the actions of their peers, and that they can replicate with a high degree of accuracy. This assumption does not always hold in reality; as demonstrated by the transmission experiments conducted by [112], even linguistic communication among human participants exhibits noisy transmission. Despite these limitations, however, observations from these simulation experiments can connect micro-level interactions to macro-level phenomena and generate relevant predictions; these predictions can 1) bring attention to potentially worrisome phenomena that warrant further examination and 2) motivate further allocation of resources to measure these risks that may manifest in real-world scenarios.

7.3 Future Work

Future work could consider reinforcement learning models in multi-agent contextual bandits settings and validate findings from previous chapters. The extended study can also explore the challenge posed by following fixed prestige cues and consider a learning environment where these explicit heuristics are absent. We have seen that action trajectories can be a powerful signal of agents' private rewards. This study could address how agents should learn who to imitate based on other agents' action trajectories. Future experiments involving reinforcement models could consider and compare the effects of different representations/observations, feedback (reward function) and strategies (network) on effective social inference and group learning. More exploration with engineering will be needed to think in terms of Markov games and address issues like scalability and computational inefficiency.

Furthermore, case studies could be conducted to illustrate when social learning works or fails in socio-technical systems. For instance, it would be valuable to explore the interplay between algorithmic amplification, which enhances the visibility of "prestigious" individuals, and social learning mechanisms, all within an unified paradigm. This investigation could delve into the dual nature of the prestige signal—examining its role as a filter for high-quality information while also considering its potential for amplifying bias. Multi-agent learning could help investigate and compare the impact of functional goal alignment (or misalignment) in socio-technical systems that leverage social learning; researchers could consider scenarios where social learning agents imitate the actions of a particular agent who shares similar contextual cues but have different, misaligned normative goals. The exploration could take place within the framework of a contextual bandit setting, shedding light on how differing objectives (denoted as μ^* in the case of contextual bandits) among agents might influence the dynamics of information sharing and learning.

Finally, this work focuses on learning among independent agents in a homogeneous, static environment with an interaction reward to predict how others would act; future work should study sequential social dilemmas and the potential role of social learning in group dynamics and norm acquisition. An example of a sequential social dilemma could be Harvest [73, 28], where individuals compete over finite resources that deplete over time. Sequential social dilemmas have spatio-temporal dynamics and are situated in strategic environments where

agents have interdependent reward structures and share tensions between a social/normative goal and self-interests. Norm acquisitions have two parts: 1) efficiently aggregating and transmitting important information across generations of agents and 2) cooperating with other agents with mismatched incentives. As Shona L. Brown puts it, maintaining a norm is like building a new equilibrium, and we need to reinforce it every day [36]. Agents must infer the cooperative intent among multiple agents in a shared environment. Running computational experiments in this environment would require more thought on how to model strategic imitation and dynamic rewards from social interactions that predict how others will act.

7.4 Final Thoughts

Multi-agent learning is a useful tool to investigate the role of social learning at an individual level in group learning dynamics.

Exploration of imitation's impact on group fitness in this thesis has shed light on the interactive effects of learning decisions over who (prestige metrics/biases), when (group imitation probabilities/network efficiency), and how to imitate in homogeneous group learning scenarios. Unbiased imitation can work well in homogeneous groups across different learning environments if a minimum age requirement is applied before model selection. Additionally, the choice between imitation strategies based on models' complete action histories or recent actions significantly influences group fitness outcomes, emphasizing the critical role of strategy selection. Surprisingly, group learning can be efficient in networks with exceptionally high levels of imitation probability, reaching up to 98%, if the group imitation strategy is adaptive to strengths and pitfalls of different model biases, or vice versa. Our findings underscore the significance of taking into account the social learning approaches at an individual level; social interventions within organizations may fail to achieve desired outcomes if these strategies and their interactions with the environment are not considered.

Careful experimentations with multi-agent learning can be helpful for policy and technology designers when they leverage social learning to build socio-technical systems. Designers mediating and coordinating social learning among users on social media platforms would want to run counterfactual experiments when they engineer explicit prestige signals that

shape group communication and learning. Designers building an algorithmic agent that is pre-trained with internet data (a repository of collective knowledge) need to consider the two-way social learning between the agent and humans; they should be wary of learning contexts and methods on both sides to mitigate ramifications of these interactions and ensure these technologies are helpful and harmless to our society at large.

In essence, this thesis illuminates the nuanced interplay between imitation, group dynamics, and learning efficiency, offering valuable insights across various domains where social learning plays a pivotal role.

Appendix A

Figures

A.1 Chapter 3

A.1.1 Changing Content Distribution Practices on ArtStation

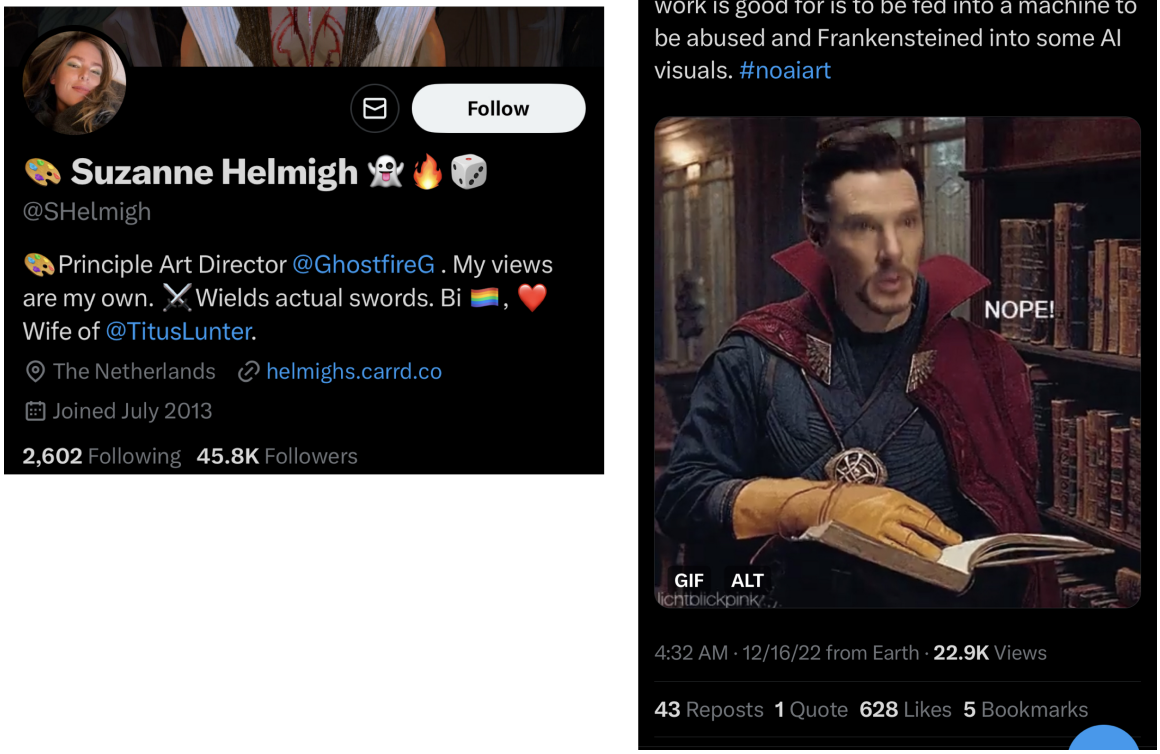

Figure A-1: Image: Tweets from @SHelmigh with 45.8K followers

A.1.2 Potentially Deadly (AI-generated) Foraging Books on Amazon

Figure A-2: Image: Tweet from The New York Mycological Society

A.1.3 Stack Overflow Information Pool

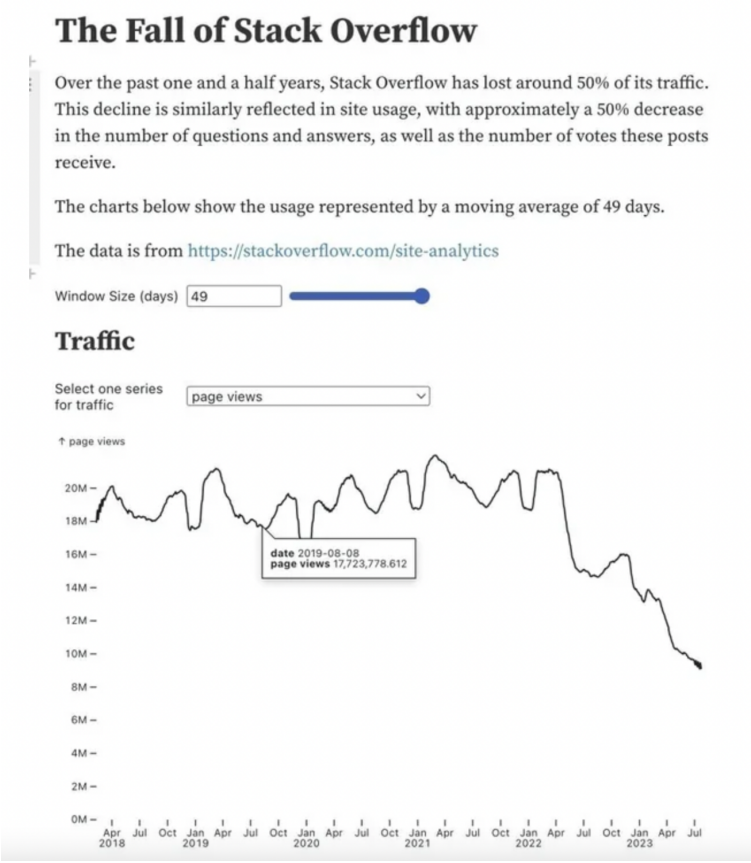

Figure A-3: From Reddit Post "ChatGPT was trained on Stackoverflow data and is now putting Stackoverflow out of business." ChatGPT is considered to be one of the causes for the decrease in traffic. One user shares "I mean, idk. Speaking only for myself, but like 90% of the things I'd previously go to stack overflow for I now ask GPT(-4) first, and much more often than not it's sufficient for solving whatever problem I've come up against."

A.2 Chapter 4

A.2.1 Model Performance Among Unbiased Learners

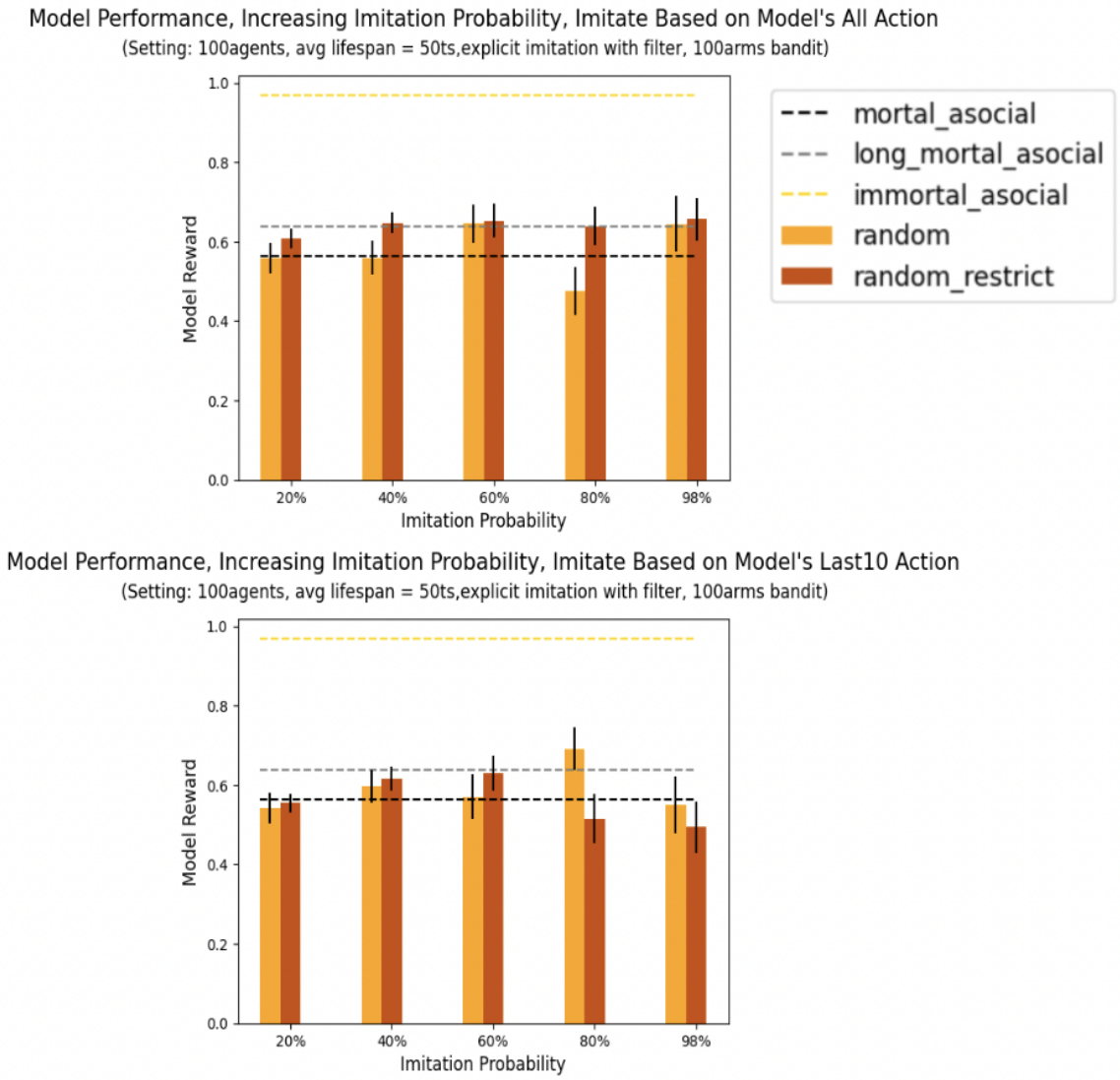

Figure A-4: **Model performance in terms of mean regret across the last 2500 timesteps** averaged across trials, each 5K steps long. (X-axis) Group imitation probability, or network efficiency. Greater Model performance is represented by higher mean reward of a model. The black dotted line is the performance of asocial learning agents with the same average lifespan of 50 timesteps. The gray dotted line is the performance of asocial learning agents with a longer average lifespan of 250 timesteps. The gold dotted line is the performance of immortal agents (the upper bound of asocial learning).

A.3 Chapter 5

In this pseudo-algorithm, "prestige" is defined as the skill level among agents who have been active for more than 20 timesteps:

Algorithm 2 Prestige-biased Social Learning among Multiple Agents; prestige: age_restricted_skill.

0: **Input:** num_agents N, status_type $= s$, imitate_freq f, learning_ability φ, avg_lifespan υ, num_timesteps T, num_arms K
0: **for** $a_i = 1$ **to** N **do**
0: **if** IMITATE = Ber(f) AND $t_a > 1$ AND Max(allAgentStatus) $\neq s_a$ **then**
0: Consider agents $(t_a)_j > 20$ to imitate
0: Select ImitatedModel = max(allAgentStatus)
0: Choose and pull an arm based on ImitatedModel's previous actions
0: **else**
0: Choose and pull an arm using Thompson Sampling
0: **end if**
0: **if** Ber(φ) = 1 AND corresponding reward > 0 **then**
0: Observe a reward of 0
0: **else**
0: Observe the true reward of the chosen arm
0: **end if**
0: Update belief about the chosen arm
0: **if** a_i reaches pre-defined lifespan **then**
0: Reset agent beliefs, status, and history
0: **end if**
0: Update agent status s_a
0: Update agent age t_a
0: **end for**=0

A.3.1 Comparing Efficiency of Social Learning versus Asocial Learning When Task Complexity is Lower

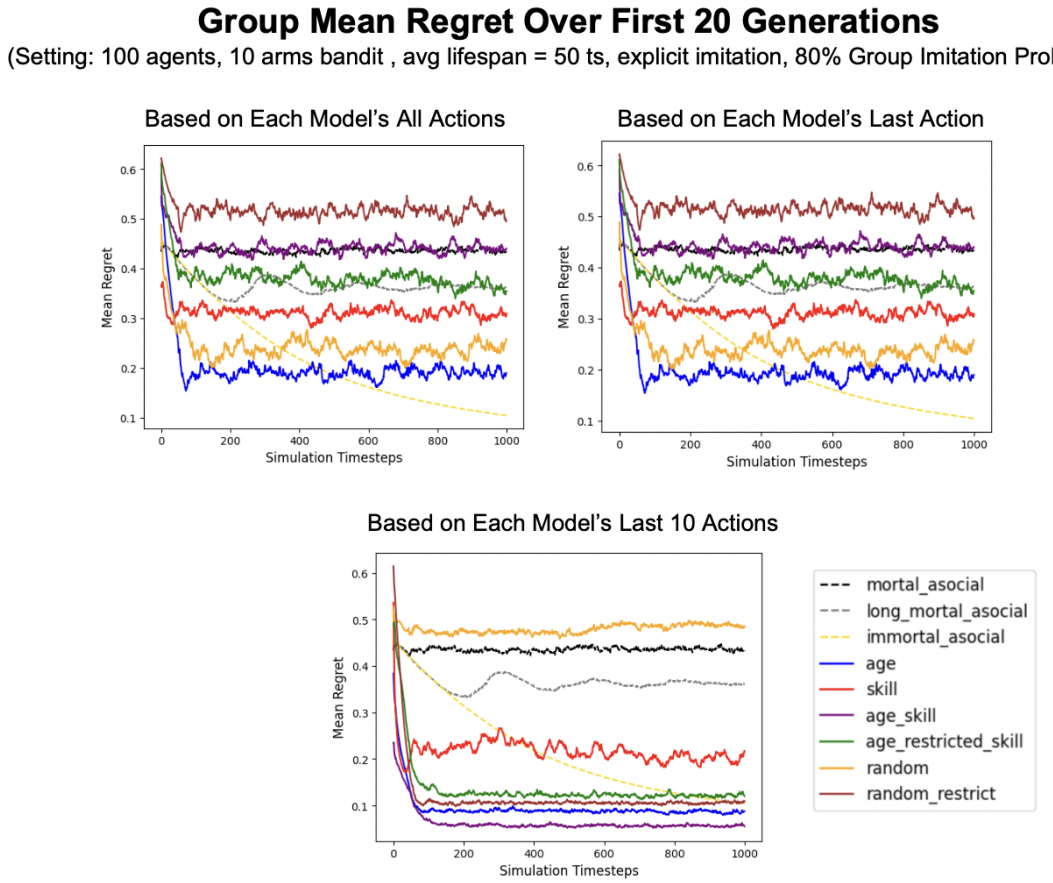

Figure A-5: *Counter-clockwise:* Imitating Based on Models' Complete Action History, Actions From Last 10 Timesteps, and Last Action. (Y-axis) **Group performance in terms of mean regret, within 100 agent network with 80% group imitation probability, facing 10 armed bandit** (X-axis) **Across the first 1000 timesteps**, averaged across trials (each 5K steps long). The black dotted line is the performance of asocial learning agents with the same average lifespan of 50 timesteps. The gray dotted line is the performance of asocial learning agents with a longer average lifespan of 250 timesteps. The gold dotted line is the performance of immortal agents (the upper bound of asocial learning)

A.3.2 Potential Pitfalls of Models Selected Based on Skill, or Mean Reward

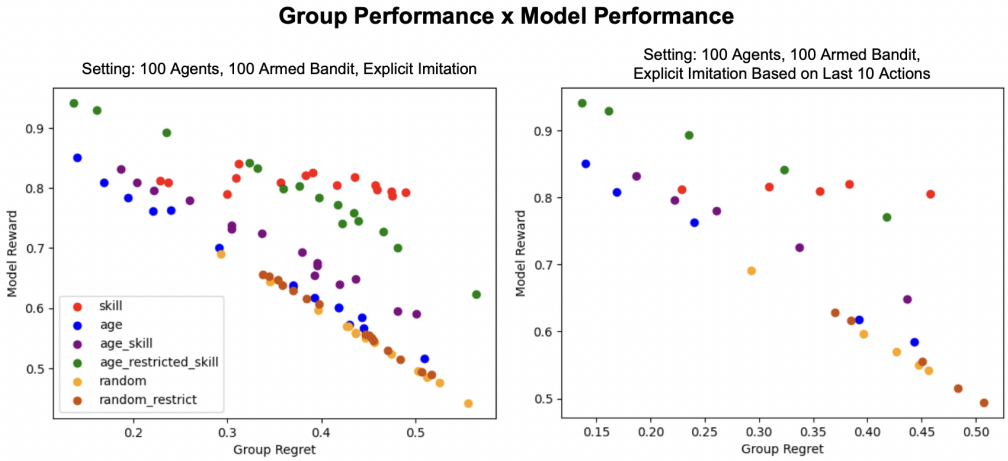

Figure A-6: **Model and group performance per model bias.** We observe a positive correlation between group performance and model performance. All learning biases have an inverse relationship between group regret and imitated model reward except bias towards the highest mean reward (skill). (Y-axis) Performance of imitated models in terms of mean reward across the last 2500 timesteps averaged across trials, each 5K steps long. (X-axis) Group performance in terms of mean regret across the last 2500 timesteps averaged across trials, each 5K steps long. Each point indicates an experiment exploring different learning parameters among 100 agents.

Social learning group performance, including random imitation, is generally positively correlated with model performance; low group performance can imply that eligible agents considered skillful after applying an age filter are not sufficiently knowledgeable. There was one type group that showed one exception to this trend: agents biased towards the highest mean reward (skill)

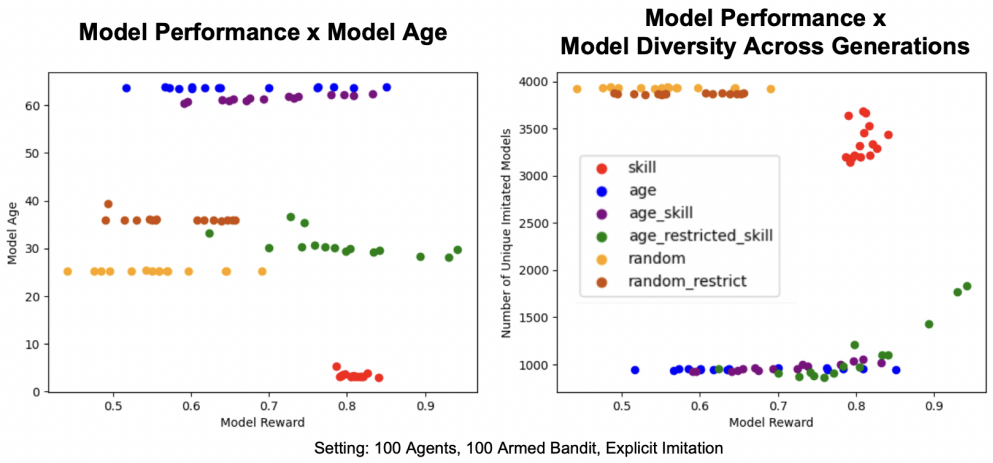

Figure A-7: *On the left:* **Correlation between Model Performance and Model Age.** (X-axis) Performance of imitated models in terms of mean reward across the last 2500 timesteps averaged across trials, each 5K steps long. (Y-axis) Model age in terms of timesteps, accumulated across its lifespan. Each point indicates an experiment exploring different learning parameters among 100 agents. *On the right:* **Correlation between Model Performance and Model Diversity.** (X-axis) Performance of imitated models in terms of mean reward across the last 2500 timesteps averaged across trials, each 5K steps long. (Y-axis) The Number of Unique Models Imitated across trials, each 5K steps long. Each point indicates an experiment exploring different learning parameters among 100 agents.

Bibliography

[1] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, page 1, New York, NY, USA, July 2004. Association for Computing Machinery.

[2] Alberto Acerbi and Joseph Stubbersfield. Large language models show human-like content biases in transmission chain experiments, July 2023.

[3] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 39.1–39.26, Edinburgh, Scotland, 25–27 Jun 2012. PMLR.

[4] Shipra Agrawal and Navin Goyal. Thompson Sampling for Contextual Bandits with Linear Payoffs, February 2014. arXiv:1209.3352 [cs, stat].

[5] Jacob Andreas. Language Models as Agent Models, December 2022. arXiv:2212.01681 [cs].

[6] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. Out of One, Many: Using Language Models to Simulate Human Samples. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 819–862, 2022. arXiv:2209.06899 [cs].

[7] Mohammad Atari, Mona J. Xue, Peter S. Park, Damián Ezequiel Blasi, and Joseph Henrich. Which Humans? preprint, PsyArXiv, September 2023.

[8] Chris L. Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B. Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):1–10, March 2017. Number: 4 Publisher: Nature Publishing Group.

[9] Aparna Balagopalan, David Madras, David H. Yang, Dylan Hadfield-Menell, Gillian K. Hadfield, and Marzyeh Ghassemi. Judging facts, judging norms: Training machine learning models to judge humans requires a modified approach to labeling data. *Science Advances*, 9(19):eabq0701, May 2023.

[10] Daniel Barkoczi and Mirta Galesic. Social learning strategies modify the effect of network structure on group performance. *Nature Communications*, 7(1):13109, October 2016. Number: 1 Publisher: Nature Publishing Group.

[11] C. J. Barnard and R. M. Sibly. Producers and scroungers: A general model and its application to captive flocks of house sparrows. *Animal Behaviour*, 29(2):543–550, May 1981.

[12] Alex Beutel, Paul Covington, Sagar Jain, Can Xu, Jia Li, Vince Gatto, and Ed H. Chi. Latent Cross: Making Use of Context in Recurrent Recommender Systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 46–54, Marina Del Rey CA USA, February 2018. ACM.

[13] Eric Bonabeau, Marco Dorigo, and Guy Theraulaz. *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press, October 1999.

[14] Robert Boyd and Peter J. Richerson. *Culture and the Evolutionary Process*. University of Chicago Press, Chicago, IL, June 1988.

[15] Robert Boyd and Peter J Richerson. Social learning as an adaptation. *Lectures on mathematics in the life sciences*, 20(1-26):26, 1989.

[16] Robert Boyd and Peter J. Richerson. Why does culture increase human adaptability? *Ethology and Sociobiology*, 16(2):125–143, March 1995.

[17] William J. Brady, Joshua Conrad Jackson, Björn Lindström, and M. J. Crockett. Algorithm-mediated social learning in online social networks. *Trends in Cognitive Sciences*, August 2023.

[18] C. O. Brand, S. Heap, T. J. H. Morgan, and A. Mesoudi. The emergence and adaptive use of prestige in an online social learning task. *Scientific Reports*, 10(1):12095, July 2020.

[19] L. Brinkmann, D. Gezerli, K. V. Kleist, T. F. Müller, I. Rahwan, and N. Pescetelli. Hybrid social learning in human-algorithm cultural transmission. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 380(2227):20200426, May 2022. Publisher: Royal Society.

[20] Axel Bruns. It's not the technology, stupid: How the 'echo chamber' and 'filter bubble' metaphors have failed us. 2019.

[21] Charlotte Canteloup, William Hoppitt, and Erica van de Waal. Wild primates copy higher-ranked individuals in a social transmission experiment. *Nature Communications*, 11(1):459, January 2020.

[22] Mauricio Cantor, Michael Chimento, Simeon Q. Smeele, Peng He, Danai Papageorgiou, Lucy M. Aplin, and Damien R. Farine. Social network architecture and the tempo of cumulative cultural evolution. *Proceedings. Biological Sciences*, 288(1946):20203107, March 2021.

[23] Micah Carroll, Anca Dragan, Stuart Russell, and Dylan Hadfield-Menell. Estimating and Penalizing Induced Preference Shifts in Recommender Systems, July 2022. arXiv:2204.11966 [cs].

[24] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback, September 2023. arXiv:2307.15217 [cs].

[25] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. Harms from Increasingly Agentic Algorithmic Systems. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 651–666, June 2023. arXiv:2302.10329 [cs].

[26] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[27] Eunseo Choi and Emőke-Ágnes Horvát. Airbnb's Reputation System and Gender Differences Among Guests: Evidence from Large-Scale Data Analysis and a Controlled Experiment. In Ingmar Weber, Kareem M. Darwish, Claudia Wagner, Emilio Zagheni, Laura Nelson, Samin Aref, and Fabian Flöck, editors, *Social Informatics*, Lecture Notes in Computer Science, pages 3–17, Cham, 2019. Springer International Publishing.

[28] Phillip J. K. Christoffersen, Andreas A. Haupt, and Dylan Hadfield-Menell. Get It in Writing: Formal Contracts Mitigate Social Dilemmas in Multi-Agent RL, August 2022. arXiv:2208.10469 [cs, econ].

[29] A. Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 864–876, Seoul Republic of Korea, June 2022. ACM.

[30] Iain D. Couzin, Jens Krause, Nigel R. Franks, and Simon A. Levin. Effective leadership and decision-making in animal groups on the move. *Nature*, 433(7025):513–516, February 2005.

[31] Henry K. Dambanemuya, Eunseo Choi, Darren Gergle, and Emőke-Ágnes Horvát. Hidden Influences of Crowd Behavior in Crowdfunding: An Experimental Study, June 2022. arXiv:2206.07210 [cs].

[32] Marquis de Condorcet. Essay on the application of analysis to the probability of majority decisions, 1785.

[33] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, January 2016. Publisher: Proceedings of the National Academy of Sciences.

[34] Maxime Derex and Robert Boyd. Partial connectivity increases cultural accumulation within groups. *Proceedings of the National Academy of Sciences*, 113(11):2982–2987, March 2016. Publisher: Proceedings of the National Academy of Sciences.

[35] William H. Dutton, Bianca Christin Reisdorf, Elizabeth Dubois, and Grant Blank. Social Shaping of the Politics of Internet Search and Networking: Moving Beyond Filter Bubbles, Echo Chambers, and Fake News. *SSRN Electronic Journal*, 2017.

[36] Kathleen M Eisenhardt and Shona L Brown. Competing on the edge: Strategy as structured chaos. *Long Range Planning*, 31(5):786–789, 1998.

[37] Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR, 17–23 Jul 2022.

[38] Jr Galef, BENNETT G. Why behaviour patterns that animals learn socially are locally adaptive. *Animal Behaviour*, 49(5):1325–1334, May 1995.

[39] Francis Galton. Vox Populi. *Nature*, 75(1949):450–451, March 1907. Number: 1949 Publisher: Nature Publishing Group.

[40] Samuel J. Gershman and Yael Niv. Novelty and Inductive Generalization in Human Reinforcement Learning. *Topics in Cognitive Science*, 7(3):391–415, July 2015.

[41] Alison Gopnik and Andrew N. Meltzoff. *Words, Thoughts, and Theories*. The MIT Press, 1998.

[42] Thore Graepel, Joaquin Quiñonero Candela, Thomas Borchert, and Ralf Herbrich. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 13–20, Madison, WI, USA, 2010. Omnipress.

[43] Ole-Christoffer Granmo and Stian Berg. Solving non-stationary bandit problems by random sampling from sibling kalman filters. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 2010.

[44] Wenshuo Guo, Kumar Krishna Agrawal, Aditya Grover, Vidya Muthukumar, and Ashwin Pananjady. Learning from an Exploring Demonstrator: Optimal Reward Estimation for Bandits, February 2022. arXiv:2106.14866 [cs, math, stat].

[45] Nika Haghtalab, Michael I. Jordan, and Eric Zhao. A Unifying Perspective on Multi-Calibration: Game Dynamics for Multi-Objective Learning, September 2023. arXiv:2302.10863 [cs].

[46] Robert D. Hawkins, Andrew M. Berdahl, Alex "Sandy" Pentland, Joshua B. Tenenbaum, Noah D. Goodman, and P. M. Krafft. Flexible social inference facilitates targeted social learning when rewards are not observable, August 2023. arXiv:2212.00869 [cs].

[47] Daniel Hawthorne-Madell and Noah D. Goodman. Reasoning about social sources to learn from actions and outcomes. *Decision*, 6(1):17–60, 2019. Place: US Publisher: Educational Publishing Foundation.

[48] Joseph Henrich. Demography and Cultural Evolution: How Adaptive Cultural Processes can Produce Maladaptive Losses: The Tasmanian Case. *American Antiquity*, 69(2):197–214, 2004. Publisher: Society for American Archaeology.

[49] Joseph Henrich. Why societies vary in their rates of innovation the evolution of innovation-enhancing institutions. 2007.

[50] Joseph Henrich and James Broesch. On the nature of cultural transmission networks: evidence from Fijian villages for adaptive learning biases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1567):1139–1148, April 2011.

[51] Joseph Henrich and Francisco J. Gil-White. The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior*, 22(3):165–196, 2001. Place: Netherlands Publisher: Elsevier Science.

[52] Joseph Henrich and Natalie Henrich. The evolution of cultural adaptations: Fijian food taboos protect against dangerous marine toxins. *Proceedings of the Royal Society B: Biological Sciences*, 277(1701):3715–3724, July 2010. Publisher: Royal Society.

[53] Joseph Patrick Henrich. *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton university press, Princeton, 2016.

[54] Cecilia Heyes. Blackboxing: social learning strategies and cultural evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1693):20150369, May 2016. Publisher: Royal Society.

[55] Jonathan Ho and Stefano Ermon. Generative Adversarial Imitation Learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[56] Lu Hong and Scott E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389, November 2004. Publisher: Proceedings of the National Academy of Sciences.

[57] Alexis Jacq, Matthieu Geist, Ana Paiva, and Olivier Pietquin. Learning from a Learner. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2990–2999. PMLR, May 2019. ISSN: 2640-3498.

[58] Natasha Jaques, Judy Hanwen Shen, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Shane Gu, and Rosalind Picard. Human-centric Dialog Training via Offline Reinforcement Learning, October 2020. arXiv:2010.05848 [cs].

[59] Julian Jara-Ettinger. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29:105–110, October 2019.

[60] Gawesh Jawaheer, Martin Szomszor, and Patty Kostkova. Comparison of implicit and explicit feedback from an online music recommendation service. In *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, pages 47–51, Barcelona Spain, September 2010. ACM.

[61] Ángel V. Jiménez and Alex Mesoudi. Prestige-biased social learning: current evidence and outstanding questions. *Palgrave Communications*, 5(1):1–12, February 2019. Number: 1 Publisher: Palgrave.

[62] Tatsuya Kameda and Daisuke Nakanishi. Does social/cultural learning increase human adaptability?: Rogers's question revisited. *Evolution and Human Behavior*, 24(4):242–260, July 2003.

[63] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson Sampling: An Asymptotically Optimal Finite Time Analysis, July 2012. arXiv:1205.4217 [cs, stat].

[64] Rachel L. Kendal, Neeltje J. Boogert, Luke Rendell, Kevin N. Laland, Mike Webster, and Patricia L. Jones. Social Learning Strategies: Bridge-Building between Fields. *Trends in Cognitive Sciences*, 22(7):651–665, July 2018.

[65] Rachel L. Kendal, Isabelle Coolen, Yfke van Bergen, and Kevin N. Laland. Trade-Offs in the Adaptive Use of Social and Asocial Learning. In *Advances in the Study of Behavior*, volume 35, pages 333–379. Academic Press, January 2005.

[66] Andrew J King and Guy Cowlishaw. When to use social information: the advantage of large group size in individual decision making. *Biology Letters*, 3(2):137–139, February 2007. Publisher: Royal Society.

[67] Max Kleiman-Weiner, Felix Sosa, Bill Thompson, Bas van Opheusden, Thomas L. Griffiths, Samuel Gershman, and Fiery Cushman. Downloading Culture.zip: Social learning by program induction. 2020.

[68] Eun-Young Ko, Eunseo Choi, Jeong-woo Jang, and Juho Kim. Capturing Diverse and Precise Reactions to a Comment with User-Generated Labels. In *Proceedings of the ACM Web Conference 2022*, pages 1731–1740, Virtual Event, Lyon France, April 2022. ACM.

[69] P. M. Krafft, Erez Shmueli, Thomas L. Griffiths, Joshua B. Tenenbaum, and Alex Sandy Pentland. Bayesian collective learning emerges from heuristic social learning. *Cognition*, 212:104469, July 2021.

[70] Kevin N. Laland. Social learning strategies. *Animal Learning & Behavior*, 32(1):4–14, February 2004.

[71] David Lazer and Allan Friedman. The Network Structure of Exploration and Exploitation. *Administrative Science Quarterly*, 52(4):667–694, December 2007. Publisher: SAGE Publications Inc.

[72] Eun-Ju Lee and Yoon Jae Jang. What Do Others' Reactions to News on Internet Portal Sites Tell Us? Effects of Presentation Format and Readers' Need for Cognition on Reality Perception. *Communication Research*, 37(6):825–846, December 2010.

[73] Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent Reinforcement Learning in Sequential Social Dilemmas. 2017. Publisher: arXiv Version Number: 1.

[74] Yunzhu Li, Jiaming Song, and Stefano Ermon. InfoGAIL: Interpretable Imitation Learning from Visual Demonstrations. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[75] Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020–9025, May 2011. Publisher: Proceedings of the National Academy of Sciences.

[76] Lauren E. Marsh, Danielle Ropar, and Antonia F. de C. Hamilton. The Social Modulation of Imitation Fidelity in School-Age Children. *PLOS ONE*, 9(1):e86127, January 2014. Publisher: Public Library of Science.

[77] Winter Mason and Duncan J. Watts. Collaborative learning in networks. *Proceedings of the National Academy of Sciences*, 109(3):764–769, January 2012. Publisher: Proceedings of the National Academy of Sciences.

[78] Winter A. Mason, Andy Jones, and Robert L. Goldstone. Propagation of innovations in networked groups. *Journal of Experimental Psychology: General*, 137(3):422–433, 2008. Place: US Publisher: American Psychological Association.

[79] Richard McElreath, Adrian V Bell, Charles Efferson, Mark Lubell, Peter J Richerson, and Timothy Waring. Beyond existence and aiming outside the laboratory: estimating frequency-dependent and pay-off-biased social learning strategies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1509):3515–3528, September 2008. Publisher: Royal Society.

[80] Nicola McGuigan, Jenny Makinson, and Andrew Whiten. From over-imitation to super-copying: Adults imitate causally irrelevant aspects of tool use with higher fidelity than young children. *British Journal of Psychology*, 102(1):1–18, 2011. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1348/000712610X493115.

[81] Kevin R. McKee, Joel Z. Leibo, Charlie Beattie, and Richard Everett. Quantifying the effects of environment and population diversity in multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 36(1):21, April 2022. arXiv:2102.08370 [cs].

[82] Alex Mesoudi. An experimental simulation of the "copy-successful-individuals" cultural learning strategy: adaptive landscapes, producer–scrounger dynamics, and informational access costs. *Evolution and Human Behavior*, 29(5):350–363, September 2008.

[83] Andrea Bamberg Migliano and Lucio Vinicius. The origins of human cumulative culture: from the foraging niche to collective intelligence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1843):20200317, December 2021. Publisher: Royal Society.

[84] Lucas Molleman, Pieter van den Berg, and Franz J. Weissing. Consistent individual differences in human social learning strategies. *Nature Communications*, 5(1):3570, April 2014. Number: 1 Publisher: Nature Publishing Group.

[85] T. J. H. Morgan, L. E. Rendell, M. Ehn, W. Hoppitt, and K. N. Laland. The evolutionary basis of human social learning. *Proceedings of the Royal Society B: Biological Sciences*, 279(1729):653–662, July 2011. Publisher: Royal Society.

[86] Michael Muthukrishna and Joseph Henrich. Innovation in the collective brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1690):20150192, March 2016. Publisher: Royal Society.

[87] Anis Najar, Emmanuelle Bonnet, Bahador Bahrami, and Stefano Palminteri. The actions of others act as a pseudo-reward to drive imitation in the context of social reinforcement learning. *PLoS biology*, 18(12):e3001028, December 2020.

[88] German Neubaum and Nicole C. Krämer. Monitoring the Opinion of the Crowd: Psychological Mechanisms Underlying Public Opinion Perceptions on Social Media. *Media Psychology*, 20(3):502–531, July 2017.

[89] Andrew Y. Ng and Stuart J. Russell. Algorithms for Inverse Reinforcement Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 663–670, San Francisco, CA, USA, June 2000. Morgan Kaufmann Publishers Inc.

[90] Mark Nielsen and Cornelia Blank. Imitation in young children: When who gets copied is more important than what gets copied. *Developmental Psychology*, 47(4):1050–1053, 2011. Place: US Publisher: American Psychological Association.

[91] Mark Nielsen, Keyan Tomaselli, Ilana Mushin, and Andrew Whiten. Exploring tool innovation: A comparison of Western and Bushman children. *Journal of Experimental Child Psychology*, 126:384–394, 2014. Place: Netherlands Publisher: Elsevier Science.

[92] Helen Nissenbaum. Accountability in a computerized society. *Science and Engineering Ethics*, 2(1):25–42, March 1996.

[93] Elisabeth Noelle-Neumann. The spiral of silence a theory of public opinion. *Journal of Communication*, 24:43–51, 1974.

[94] Eduardo B. Ottoni, Briseida Dogo de Resende, and Patrícia Izar. Watching the best nutcrackers: What capuchin monkeys (Cebus apella) know about others' tool-using skills. *Animal Cognition*, 8(4):215–219, 2005. Place: Germany Publisher: Springer.

[95] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18, Bend OR USA, October 2022. ACM.

[96] Hart E. Posen, Jeho Lee, and Sangyoon Yi. The power of imperfect imitation. *Strategic Management Journal*, 34(2):149–164, 2013. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/smj.2007.

[97] Markus Prior. Media and Political Polarization. *Annual Review of Political Science*, 16(1):101–127, May 2013.

[98] Stephan G. Reebs. Can a minority of informed leaders determine the foraging movements of a fish shoal? *Animal Behaviour*, 59(2):403–409, 2000.

[99] L. Rendell, R. Boyd, D. Cownden, M. Enquist, K. Eriksson, M. W. Feldman, L. Fogarty, S. Ghirlanda, T. Lillicrap, and K. N. Laland. Why Copy Others? Insights from the Social Learning Strategies Tournament. *Science (New York, N.Y.)*, 328(5975):208–213, April 2010.

[100] L. Rendell, R. Boyd, M. Enquist, M. W. Feldman, L. Fogarty, and K. N. Laland. How copying affects the amount, evenness and persistence of cultural knowledge: insights from the social learning strategies tournament. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1567):1118–1128, April 2011. Publisher: Royal Society.

[101] Luke Rendell, Laurel Fogarty, William J. E. Hoppitt, Thomas J. H. Morgan, Mike M. Webster, and Kevin N. Laland. Cognitive culture: theoretical and empirical insights into social learning strategies. *Trends in Cognitive Sciences*, 15(2):68–76, February 2011.

[102] Peter J. Richerson and Robert Boyd. *Not By Genes Alone: How Culture Transformed Human Evolution*. University of Chicago Press, Chicago, 2008. OCLC: 1059001661.

[103] Alan R. Rogers. Does Biology Constrain Culture. *American Anthropologist*, 90(4):819–831, 1988. Publisher: [American Anthropological Association, Wiley].

[104] Stuart Russell. Learning agents for uncertain environments (extended abstract). In *Proceedings of the eleventh annual conference on Computational learning theory*, COLT' 98, pages 101–103, New York, NY, USA, July 1998. Association for Computing Machinery.

[105] Steven L. Scott. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, November 2010.

[106] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The Curse of Recursion: Training on Generated Data Makes Models Forget. 2023. Publisher: arXiv Version Number: 2.

[107] Jonathan Stray. Designing recommender systems to depolarize. *First Monday*, May 2022.

[108] Jonathan Stray, Alon Halevy, Parisa Assar, Dylan Hadfield-Menell, Craig Boutilier, Amar Ashar, Lex Beattie, Michael Ekstrand, Claire Leibowicz, Connie Moon Sehat, Sara Johansen, Lianne Kerlin, David Vickrey, Spandana Singh, Sanne Vrijenhoek, Amy Zhang, McKane Andrus, Natali Helberger, Polina Proutskova, Tanushree Mitra, and Nina Vasan. Building human values into recommender systems: An interdisciplinary synthesis, 2022.

[109] Sally E. Street, Ana F. Navarrete, Simon M. Reader, and Kevin N. Laland. Coevolution of cultural intelligence, extended life history, sociality, and brain size in primates. *Proceedings of the National Academy of Sciences*, 114(30):7908–7914, July 2017. Publisher: Proceedings of the National Academy of Sciences.

[110] Natalie Jomini Stroud, Ashley Muddiman, and Joshua M Scacco. Like, recommend, or respect? Altering political behavior in news comment sections. *New Media & Society*, 19(11):1727–1743, November 2017.

[111] S S Sundar and C Nass. Conceptualizing sources in online news. *Journal of Communication*, 51(1):52–72, March 2001.

[112] B. Thompson, B. van Opheusden, T. Sumers, and T. L. Griffiths. Complex cognitive algorithms preserved by selective social learning in experimental populations. *Science*, 376(6588):95–98, April 2022. Publisher: American Association for the Advancement of Science.

[113] Bill Thompson and Thomas L. Griffiths. Human biases limit cumulative innovation. *Proceedings of the Royal Society B: Biological Sciences*, 288(1946):20202752, March 2021. Publisher: Royal Society.

[114] William R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3/4):285, December 1933.

[115] Wataru Toyokawa, Andrew Whalen, and Kevin N. Laland. Social learning strategies regulate the wisdom and madness of interactive crowds. *Nature Human Behaviour*, 3(2):183–193, February 2019. Number: 2 Publisher: Nature Publishing Group.

[116] Natalia Vélez and Hyowon Gweon. Learning from other minds: an optimistic critique of reinforcement learning models of social learning. *Current Opinion in Behavioral Sciences*, 38:110–115, April 2021.

[117] Henry M. Wellman. *Making Minds: How Theory of Mind Develops*. Oxford University Press, November 2014.

[118] Max Wolf, Ralf H. J. M. Kurvers, Ashley J. W. Ward, Stefan Krause, and Jens Krause. Accurate decisions in an uncertain world: collective cognition increases true positives while decreasing false positives. *Proceedings. Biological Sciences*, 280(1756):20122777, April 2013.

[119] Lara A. Wood, Rachel L. Kendal, and Emma G. Flynn. Whom do children copy? Model-based biases in social learning. *Developmental Review*, 33(4):341–356, 2013. Place: Netherlands Publisher: Elsevier Science.

[120] Yueh-Hua Wu, Nontawat Charoenphakdee, Han Bao, Voot Tangkaratt, and Masashi Sugiyama. Imitation Learning from Imperfect Demonstration. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6818–6827. PMLR, May 2019. ISSN: 2640-3498.

[121] Eunice Yiu, Eliza Kosoy, and Alison Gopnik. Imitation versus Innovation: What children can do that large language and language-and-vision models cannot (yet)?, May 2023. arXiv:2305.07666 [cs].

[122] Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. Recommendation as Instruction Following: A Large Language Model Empowered Recommendation Approach, May 2023. arXiv:2305.07001 [cs].

[123] Frederik J. Zuiderveen Borgesius, Damian Trilling, Judith Möller, Balázs Bodó, Claes H. De Vreese, and Natali Helberger. Should we worry about filter bubbles? *Internet Policy Review*, 5(1), March 2016.