When is Similarity-biased Social Learning Adaptively Advantageous?

Daniel Saunders

1 Introduction

Young children occupy a world surrounded by potential role models. The wealth of available information creates a problem. Different role models do different things. Children need strategies to separate the good role models from the bad ones. For many problems, the strategy is simple - imitate whoever is most successful. Extensive empirical evidence and theoretical modeling supports the idea that children use *success-biased* social learning (Mcelreath et al. 2008; Hoppitt and Laland 2013; J. Henrich and McElreath 2003; N. Henrich and Henrich 2007). But some learning problems are too hard for success-biased learning. What enables success for some people is dependent on what others in the community expect of them. Crucially, different expectations are placed on different people. The success of a behavioral trait depends upon a complex and invisible network of social roles, norms, and expectations. Moreover, children will be involuntarily socialized into a variety of roles which will impose constraints on their behavior, and acting inconsistently with one's social role can result in punishment or lost opportunities. Children need a special learning strategy to acquire just the right behaviors to act consistently with what people will come to expect of them.

Cultural evolution theorists have suggested that, in addition to success-biased learning, we also engage in *similarity-biased* social learning (Wood, Kendal, and Flynn 2013; J. Henrich 2016; N. Henrich and Henrich 2007; J. Henrich and McElreath 2003). We pay special attention to and tend to imitate people who are already similar to us. Of course, every pair of two people are similar and dissimilar in numerous ways. Similarity-biased learning must be confined to certain features which play an important role in expectations, norms, and roles. Similarities associated with gender, ethnicity, and race are leading candidate features for organizing learning. On this theory, we should expect men to primarily learn from men, women to learn from women, and so on.

Evolutionary explanations have been offered for similarity-biased learning (Kinzler, Corriveau, and Harris 2011; Wood, Kendal, and Flynn 2013; J. Henrich 2016; N. Henrich and Henrich 2007). The details of the explanation differ depending on whether the target is gender, race, or ethnicity. Perhaps the clearest

articulation of an evolutionary story in any of these cases comes from Natalie Henrich and Joseph Henrich's book *Why Humans Cooperate*:

> Individuals have to figure out what the right norm is for getting along in their social groups, keeping in mind that different social groups culturally evolve different norms. By "right," we mean the norm that allows the individual to maintain their reputation, avoid punishment for norm violations, and coordinate their behavior with other members of their social group. …
>
> Human psychology evolved to seek out "indicator traits" (language, dress, etc.) that match its own because people who have the same markers tend to also have the "right" norms. Using such markers, individuals can bias both their learning from, and interaction with, those individuals who share their same culturally transmitted indicator traits. (N. Henrich and Henrich 2007)

The suggestion is that ethnicity-biased social learning is adaptive because it enables reliable transmission of social roles across generations. This explanation runs into a variety of difficulties.

Some difficulties are theoretical. If someone spends most of their time with people from the same ethnic group as me, they do not need a special psychological bias to pay attention to them. Co-ethnics are the only people around them. Moreover, in the case where people from multiple ethnic groups are in close contact, people from other ethnic groups might have a new, useful innovation. It would be disadvantageous to ignore them (Howard et al. 2015).

Other difficulties are empirical. Experiments over the last 40 years have consistently demonstrated that people exhibit gender-biased social learning (Bussey and Bandura 1984; Shutts, Banaji, and Spelke 2010; Losin et al. 2012; Perry and Bussey 1979). But the experimental literature on racial- or ethnicity-biased learning contains no similarly robust findings. Some experiments find support (Kinzler, Corriveau, and Harris 2011; Kinzler, Dupoux, and Spelke 2007; Shutts et al. 2009). Others turn up null results (Howard et al. 2015; Krieger et al. 2016; Shutts, Banaji, and Spelke 2010). One study finds the effect in some conditions but not others (Buttelmann et al. 2013). If the evolutionary argument for similarity-biased learning is strong, one should not expect the two experimental literatures to generate such a different patterns of results.

Unlike other learning strategies, there has been little effort to formally model the conditions under which similarity-biased learning is evolutionarily adaptive. Instead, the literature contains a variety of informal arguments like the block quote above. The informal arguments provide little guidance on how to assess the difficulties. Given the complexity of human societies, it is difficult to verbally reason through the consequences of multiple, interacting factors (Muthukrishna and Henrich 2019; Mayo-Wilson and Zollman 2021; Smaldino 2020). This paper rectifies the situation by developing a series of agent-based models to study the evolution of similarity-biased social learning. The general insight is that there is

no evolutionary story to be told that is simultaneously successful and generally applicable. Whether we should expect similarity-biased social learning to evolve strongly depends on assumptions about the adaptive function of social roles, the initial conditions, a variety of parameter settings, and the population structure. Making small changes to these assumptions can collapse the explanation. The results suggest we should be very cautious about claims suggesting there is a universal, evolved tendency towards similarity-biased learning.

Developing a better understanding of similarity-biased social learning supports an existing research program in philosophy. Philosophers have turned to bargaining models to generate possible explanations for the emergence of inequality (O'Connor 2017, 2019; O'Connor and Bruner 2019, 2017; Bruner 2019; O'Connor, Bright, and Bruner 2019; Mohseni, O'Connor, and Rubin 2019; Amadae and Watts 2022; Bright et al. 2022; Popa and Muehlenbernd 2022). Under a variety of conditions, representing inequalities in power or demography, it is possible for two populations to develop a stable pattern of exploitative relations. Many of the cited models assume that people engage in gender-biased or racially-biased forms of social learning. However, it is often unclear what justifies this assumption, except passing appeals to the experimental results cited above or evolutionary reasoning. The end of this paper sketches some suggestions for how research into bargaining-driven explanations of inequality could improve going forward.

2 Model construction

Imagine two ethnic groups settle into a valley. Each group has a distinctive style of greeting. The first group bows while the second group hugs. Most of the time, individuals just interact with members of their own group. In those cases, greetings are smooth and uneventful. Occasionally, people from different ethnic groups will encounter each other. It is deeply awkward to move in for a hug while the other person bows. This is a case of failed coordination - a custom that is supposed to be nothing more than a prelude to a conversation has become a source of mutual embarrassment. The greeting problem is an instance of a general class of coordination games. Coordination games have this characteristic payoff table:

	A	B
A	1,1	0,0
B	0,0	1,1

The basic problem in coordination games is knowing what your partner will do. If players have no extra information beyond the table, they are left guessing as to whether their partner will play A or play B. If they guess wrong, they fail to coordinate. Philosophers have long noted that social conventions can provide

the missing information by specifying one equilibrium to be the expected one (Lewis 1969; Schelling 1960; Skyrms 1996, 1990). Cultural evolution theorists have followed suit and argued that we culturally evolved ethnic markers (distinctive styles of fashions, makeup, tattoos, etc) to help solve coordination problems. Ethnic identities function to settle coordination problems in advance by specifying the particular operative convention in this community (McElreath, Boyd, and Richerson 2003). Most obviously, ethnicity helps settle whether the conventional greeting is bowing or hugging. A huge variety of other coordination problems are also settled via ethnically-specific conventions: what language to speak, what dialectic to use, what slang to use, what currency to offer, who's family should offer the marriage dowry, what taboos to avoid, and so on. If altruistic punishment theories of cooperation are to be believed, the class of coordination problems is far larger, as social dilemmas are transformed into coordination problems by the threat of punishment (N. Henrich and Henrich 2007; Boyd et al. 2003).

This game provides a foundation for building a model of the conditions under which similarity-biased learning can evolve. Suppose that the two imaginary ethnic groups are composed of a large, finite number of players[1]. Each player is equipped with a strategy. In the first version of the model, players have only two *behavioral strategies* for the coordination game. They can either play A or they can play B. These correspond to always bowing or always shaking hands. Section 4.4 explores a version of the model in which players can adopt conditional strategies, where players vary their greeting depending on the ethnic marker of their partner. For now, the model assumes players only have access to simple strategies.

The model evolves over a series of rounds. Each round, the players pick a partner. With some probability, that partner is from their group. Otherwise, it is a random partner. Then they play the coordination game with their strategy and receive a payoff. The probability of in-group pairing represents the fact people tend to interact with their own ethnic group more often. This might be due to a psychological preference for in-group members (McElreath, Boyd, and Richerson 2003; N. Henrich and Henrich 2007). But the probability could also represent the fact that people who live in the same place will encounter each other more often and ethnic groups tend to live in the same place.

Each individual must reliably identify which strategy they should adopt. They can pursue two *learning strategies*. They might only learn from people of their own ethnic group. This represents a type of similarity-biased learning. Or, they might learn from whoever they encounter. Regardless of how they select partners, players imitate using "pairwise difference imitation" (Izquierdo, Izquierdo,

[1]Concentrating on ethnicity first provides a conceptually tractable entry point into thinking about similarity-biased social learning. However, this paper is interested in a broader phenomenon. For example, similarity-biased social learning also manifests in gender. Section 5 argues that the model can be abstracted away from interpretation in terms of ethnicity and generalized to any social roles which facilitate coordination.

and Sandholm 2019). In each round, each player selects a partner, calculates the difference between their own payoff and their partner's payoff, and imitates their partner's behavioral strategy with a probability proportional to the difference. This learning dynamic closely approximates the behavior of the famous replicator dynamics but in an agent-based context (Izquierdo, Izquierdo, and Sandholm 2019).

To identify the correct learning strategy, players also employ a second-order adaptive process. Learning strategies that tend to lead people to successful strategies are more adaptive. On each round, players select a new random partner[2]. They apply the pairwise difference imitation rule again. The difference is that they copy the learning strategy instead of the behavioral strategy.

There are a variety of ways to interpret this second-order adaptive process. The first option is to treat similarity-biased learning as the outcome of natural selection on genes. This is the route the Henrichs take (N. Henrich and Henrich 2007). The second option is to treat it as kind of second-order social learning. People learn how to learn by observing others[3]. Although second-order social learning is an under-explored topic, some empirical evidence suggests that it is prominent in human societies (Mesoudi et al. 2016). Plausibly, some combination of natural and cultural selection are at work. In theoretical research on the evolution of learning strategies, the standard methodological choice is to remain agnostic about what precise mechanism implements the learning strategy (Hoppitt and Laland 2013). It simply does not matter at the level of representation found in these models. The model is merely committed to the claim that if there is a trait for similarity bias in social learning and selection could act on it, then we should expect it to grow in the population.

To summarize, the initial conditions of the model are:

- There are two groups of finite size.
- Players from each group begin with a distinctive behavioral strategy. All players from one group will begin with strategy A. Players from the other group will begin with strategy B.

The dynamics of the model are[4]:

- **Play**. Pick a partner. With some probability, the partner is a member of the same group. Otherwise, they are picked at random from the entire

[2]Normally, the second-order social learning process is indiscriminate. Players can learn their learning strategy from anyone. But some simulations were attempted where the second-order social learning process was tied to ethnic markers. Qualitatively similar results were found.

[3]This interpretation may seem implausible at first glance. Given that learning is not externally visible, it may be difficult to directly imitate the mechanisms by which others learn. However, consider the case where a teacher directs the students' attention to another exemplary student. If the students respond by trying to imitate the exemplar, they are using a kind of second-order social learning.

[4]The model was coded in Netlogo. The data analysis was done in Python. Code for both parts is available at https://github.com/daniel-saunders-phil/dowry-game

population. Play the coordination game and receive a payoff.

- **Learn**. Pick a partner. If the learner is using a similarity-biased learning strategy, they pick a partner that shares their group identity. Otherwise, they pick a random partner. They copy the partner's behavioral strategy with probability proportional to the difference in payoffs.
- **Learn learning strategy**. Pick a partner. Copy the partner's learning strategy with probability proportional to the difference in payoffs.

3 Core Results

The Henrichs suggested that similarity-biased learning strategies are adaptive because they help agents select the "right" behaviors, where "right" is defined by ethnically-specific conventions. The model provides a partial vindication of their argument. In some runs of the model, the population evolves toward uniform similarity-biased social learning. Every player only learns from others in their group. This stabilizes the ethnically-specific greeting conventions. One group always bows and the other group always shakes hands. Call this the *distinctive strategy outcome*.

In these runs of the model, similarity-biased learning is adaptive because it steers players away from picking up the other group's greeting. Suppose one individual meets someone from the other group. Given the initial strategy assignment where one group plays A and the other group plays B, this pair will likely fail to coordinate. The defeated player will then be in the market for a new strategy. If they lack the similarity bias, they may observe someone from the other group and adopt their strategy. But this would likely be a mistake. If most of their future interactions will be with in-group members who use a different greeting, switching to the other group's greeting will be disadvantageous in the long run. Thus, selection encourages the growth of the similarity bias to prevent just this scenario. This finding indicates that the model is a charitable, formal reconstruction of the Henrichs' argument.

However, another stable state is possible for the population. The groups could *assimilate*. All players can simply adopt the same greeting, regardless of the group they belong to. To appreciate how this outcome arises, reconsider the scenario sketched above. This time, imagine our failed coordinator adopts the other group's characteristic behavior during the same round that a large number of other people in their group make the same choice. If enough players flip strategies, then playing the most popular strategy across groups is more advantageous than playing the ethnically-specific strategy. This can create a cascade of selection pressure toward the most popular strategy. As more players adopt one greeting, it becomes increasingly important for others to adopt that greeting. In this case, similarity-biased social learning is disadvantageous. It leaves players slow to adopt the most popular strategy. Once a single coordination strategy is adopted uniformly, selection pressure stops. It does not matter how

players learn because everyone coordinates on every interaction and everyone has the same behavior. This is a surprising insight that pushes against simple evolutionary arguments for learning strategies, the kind of insight that only becomes visible in light of formal modeling.

These are the only two possible stable states for the population. Notably, the assimilation outcome generates a higher average payoff for individuals. In the distinctive strategy outcome, players will fail to coordinate whenever they play with people from the other group. The assimilation outcome does not have this feature. Everyone is doing the same thing so coordination failure is impossible. Given that the assimilation outcome has a higher payoff, one should expect it to be more evolutionary attractive, all other things being equal. To reliably generate the distinctive strategy outcome, it is necessary to make further assumptions about the population structure and parameter values. The question, then, is what features control whether assimilation is more or less likely than the distinctive strategy outcome. Understanding the answer to that question will provide a theoretical understanding of when similarity-biased social learning is adaptively advantageous.

The following section explores the sensitivity of the model across a range of parameter settings and structural assumptions. A general theme emerges. The distinctive strategy outcome tends to emerge whenever additional features are added to the model to slow the possibility of assimilation enough that the learning bias has time to emerge. However, these additional features raise two new problems. First, the more specific the modeling assumptions need to be, the less generally the model can be applied. We do not know the true parameter values in historical human societies. Nor should one expect any one population structure or combination of parameters to be the "correct" one for characterizing the broad swath of human societies. Second, the very same features that slow assimilation also often weaken the overall force of selection. This raises concerns about whether non-adaptive forces such as mutations, drift, or migration might swamp selection. If that is the case, this type of adaptive explanation has limited explanatory power for human learning strategies.

4 Sensitivity analysis

4.1 Random pairing rate

The model assumes that people have a higher probability of interacting with people from their own group. This is controlled by the random pairing parameter. When it is low, people pair almost exclusively with in-group members. When it is high, players only have a slight preference for in-group pairing. The random pairing parameter has a large impact on whether similarity-biased learning is adaptive. It is beneficial to learn the in-group behavior if most interactions happen within the group. Otherwise, it is more beneficial to seek out the most popular coordination behavior. Figure 1 depicts the results of a sensitivity analysis exploring how variation in the random pairing parameter impacts how often

the distinctive strategy state emerges.

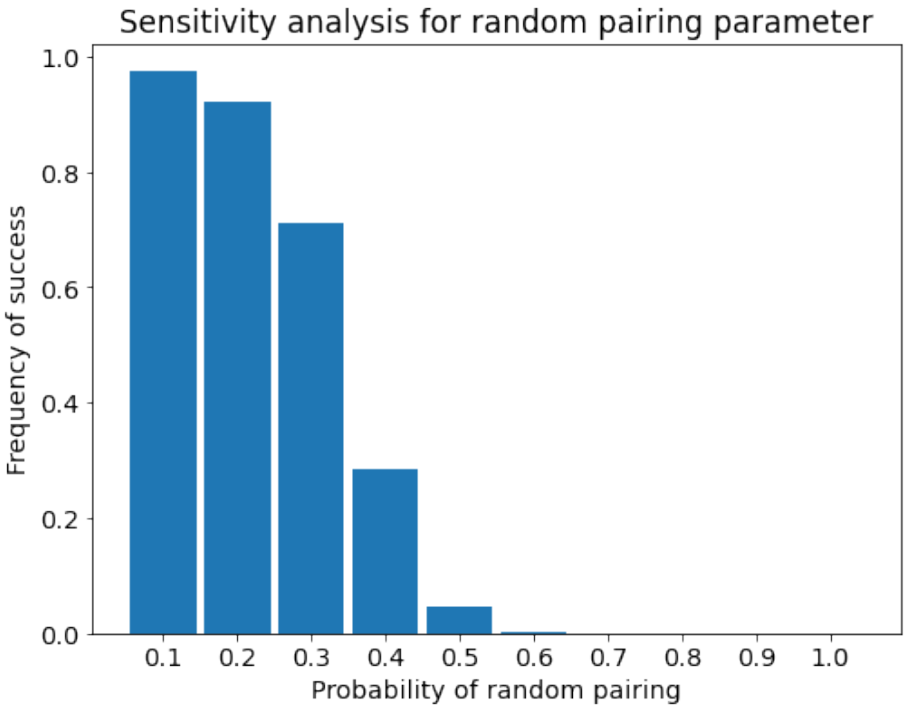

Figure 1: Bar heights represent the frequency of simulation experiments that ended in uniform similarity-biased social learning. 500 trials were conducted at each 0.1 increment of the random pairing parameter between 0 and 1. The results show similarity-biased social learning is more likely to evolve when most interactions take place within the group.

The probability that the learning bias emerges declines rapidly as pairing with the out-group becomes more common. This fact, taken alone, suggests some limitations on the adaptive explanation but is not too concerning. After all, it does make good sense to think people largely interact with their own group. However, frequent in-group interaction also dampens the strength of selection. If players coordinate most of the time, they have little need to explore new behavioral strategies. Learning, similarity-biased or not, will occur rarely in a population with little random pairing. One indication of the strength of the selection is the number of rounds it takes for the model to reach a stable state. When selection is weak, it takes much longer.

Inspection of figure 1 alongside figure 2 suggests a dilemma: when the distinctive strategy state is most likely, selection is at its weakest. But when selection is weak, other non-adaptive forces can play a determining role. Forces like migration and mutation can introduce random variation into the distribution of

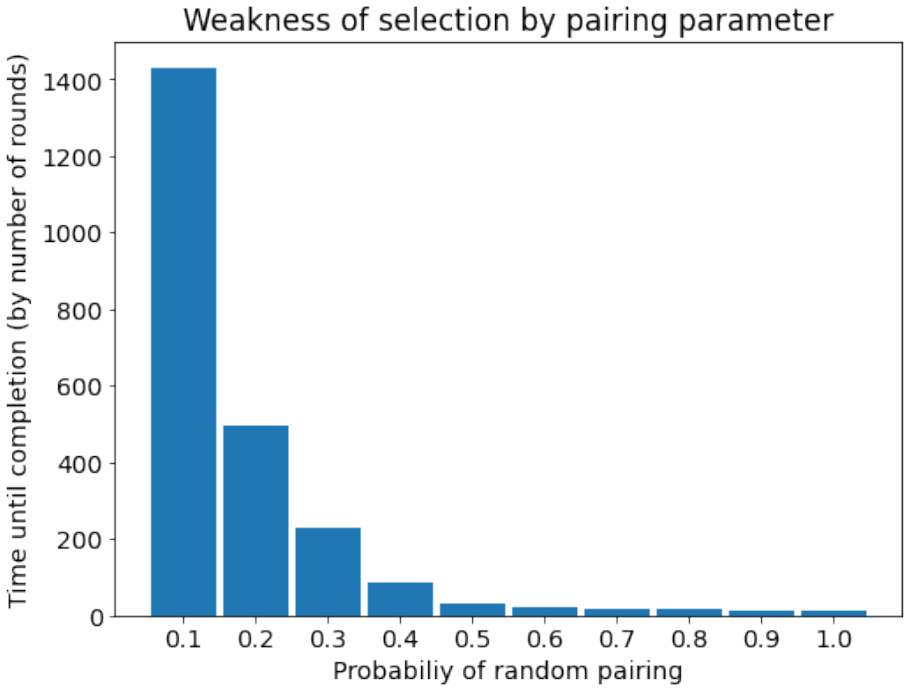

Figure 2: Bar heights represent the average number of rounds it takes the model to reach a stable state, either the distinctive strategy outcome or the assimilation outcome. These data were extracted from the same experiments depicted in figure 1. As random pairing decreases, the model takes exponentially more time to reach a stable state. Tall bars indicate weak selection.

strategies and pull the population toward assimilation. As an empirical generalization, in small, non-isolated, human populations the force of migration is typically stronger than the force of selection (McElreath and Boyd 2007; Boyd et al. 2003).

4.2 The initial distribution of learning strategies

When the population is initially generated, some frequency of the learning bias is already present. In the previous section, the model gave each player an equal probability of starting with the bias or not. On average, half the players start the simulation using the learning bias. It turns out that a high level of initial bias is necessary for the distinctive strategy outcome. If the population starts with low frequency of the learning bias, assimilation becomes more likely. Figure 3 depicts how often similarity-biased learning evolves across a range of initial frequencies.

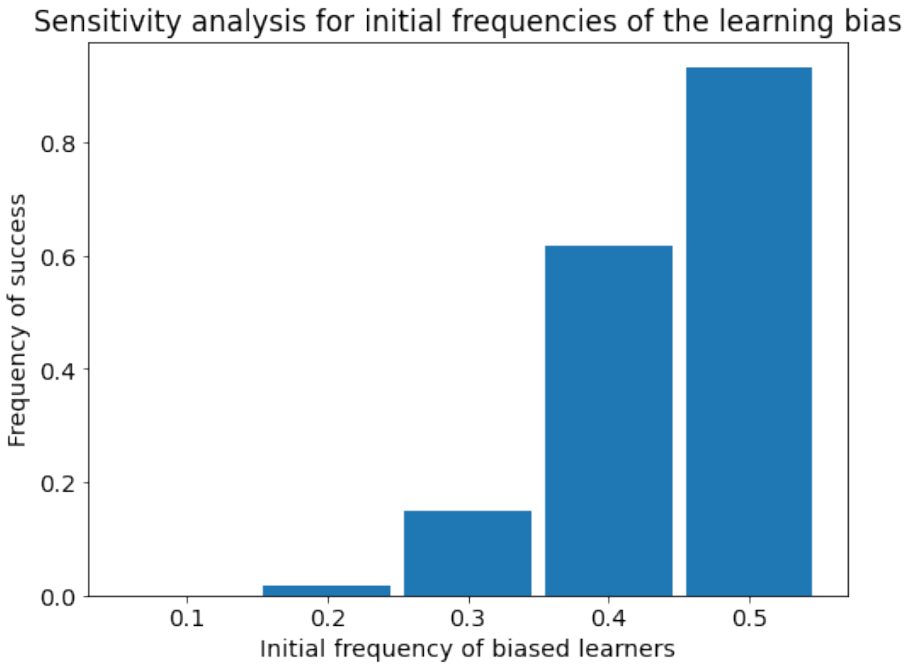

Figure 3: Bar heights represent the frequency of simulation experiments that ended in uniform similarity-biased social learning. 500 simulations were run at each parameter value. The random pairing parameter was set to 0.2. As the initial frequency of the learning bias decreases, so does the probability that learning bias will spread.

This effect arises because a high level of learning bias is necessary to prevent fast assimilation. If too many players begin copying behavioral strategies from

the other group, then that behavior may become popular in both groups. In turn, it becomes more advantageous to learn the most popular strategy rather than the ethnically-specific one. By contrast, if most people already learn from their in-group, then behaviors tend to cluster tightly with ethnic groups and it is more advantageous to conform to the ethnically-specific convention.

The strong sensitivity of the model to the initial conditions is troubling for evolutionary explanations. If the trait does not tend to evolve when it is rare, it is hard to understand how it ever became common in the first place.

4.3 The ratio of playing to learning

The model assumes that players play once and learn once in each round. This assumption is fairly arbitrary. Players could go through many interactions before they consider changing their behavioral strategy. Adjusting the ratio playing to learning could represent how stubborn or open-minded the players are. They might keep their strategy over the course of many interactions and only consider giving it up if it proves disadvantageous in the long run. The model was modified so players would execute n play actions per round and only one learn action. They would repeatedly select partners, interact, and receive payoffs. Both learning processes were modified so players compare the sum of payoffs across n interactions.

Unlike the previous sensitivity tests, this modification *improves* the probability that similarity-bias learning emerges. When players play more often, there is a higher penalty for picking up the other group's behavior. Given that most interactions still occur within the group, picking up the other group's behavior will lead to coordination failure in most interactions. Changing the ratio also slows down the rate of assimilation. Assimilation outcomes depend on the possibility of players flipping en mass. When out-group learning is so heavily penalized, few agents will be inclined to flip.

Figure 4 depicts how the results of the two previous sensitivity analyses change when $n = 10$. The figure demonstrates that increasing n can facilitate the evolution of similarity-biased learning at higher rates of random pairing. Increasing n can also push back against the problem of initial conditions but only to a modest degree.

The results point to a way forward for evolutionary explanations of similarity-bias learning. If we make the plausible assumption that people interact at a faster rate than they learn, it alleviates some problems the original model faced. However, one should be cautious about hanging their hat on any specific set of parameter values. It is not at all clear how we might decide what the right ratio of playing and learning is. Moreover, the bias is still not usually selected for when it is rare, leaving that challenge unsolved.

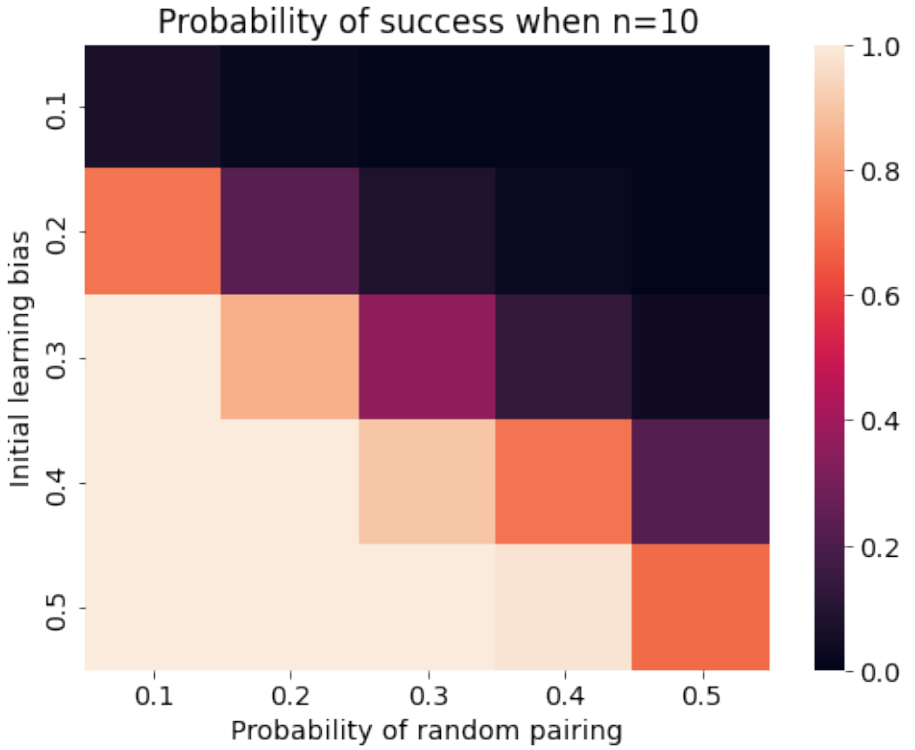

Figure 4: The heatmap represents the frequency of simulation experiments that ended in uniform similarity-biased social learning. Light colors represent high frequency while dark colors represent low frequency. 100 trials we conducted at each pair of parameter values. The figure suggests a high n improves the overall frequency of successful trials, although its effect is limited when the initial frequency of learning bias is close to 0.1.

4.4 Spatial structure

One lesson has become clear: when it is too easy to adopt the neighboring group's behaviors, assimilation is likely. If there was some additional feature of the model that could slow the assimilation process, it might give similarity-biased learning a better chance to survive. A spatial structure could serve just this role. Suppose the population is laid out on a grid and they can only play with and learn from the 8 players around them. This generates a kind of "ethnic-frontier" where players from each group can interact with the out-group. Players at the border have to progressively flip for assimilation to occur.

Introducing spatial structure induces two simultaneous effects. First, it decreases the rate of cross-group interaction. For most agents, they have no neighbors of the other ethnic group so they still pair with the in-group no matter the level of random-pairing. Only agents who live on the frontier have a chance of interacting across ethnic groups. Second, it spatially concentrates selection activity. If ethnicity-biased social learning has higher fitness, it has fitness because border agents benefit from it. Agents who are spatially removed from the frontier might adopt ethnicity-biased social learning only by observing its fitness-enhancing effects on others.

Spatial structure does not change the basic problem confronted by non-spatial models. It merely provides an alternative way of pushing the effective random-pairing parameter down. Pushing it down structurally certainly makes the evolution of ethnicity-biased social learning more likely across a wider range of pairing parameters. But it amplifies the problem of weak selection. Given that spatial structure limits selection activity, it also takes the model far longer to converge to a stable state.

4.5 Conditional strategies

In the original model, players can only adopt one strategy for coordinating. They cannot learn to adopt different strategies for different communities. It is as if learning to bow would cause you to forget about hugging. Of course, people can learn the customs necessary to live in two communities. Bilingual people are a good illustration. They can solve coordination problems in two different ways, depending on who they are speaking with. The model can be expanded to incorporate *conditional* strategies. Agents now modify their strategy based on whether they play with an in-group agent or an out-group agent.

When players can condition their strategies, it eliminates the need for assimilation. Each group can have their own particular convention they utilize when interacting with themselves while having a shared convention they utilize when interacting across the groups. The result is like a *lingua franca*. The scientific community has adopted English as the shared international language for publication. But scientists will continue to speak their native language with others from their community.

Crucially, the possibility of having two conventions (one for in-group interaction and one for out-group interaction) also eliminates selection for learning strategies. Every learning strategy has the same payoff once the interaction strategies reach equilibrium. Speaking English with scientists and English-speakers is just as good as speaking English with scientists and Spanish with Spanish-speakers. Everyone always successfully communicates. But selection only operates when there is variation in the fitness of traits. Modifying the model in this way is a non-starter for the original problem.

4.6 Summary of the sensitivity analysis

Formalizing the Henrichs' argument illuminates a variety of limitations. First, in all versions of the model explored, learning bias is selected against when it is rare. If the trait arose through selection, it must have spread through a group via mutation and drift when a group was isolated. Later on, if two groups come into contact, similarity-biased learning could be propelled throughout the population. This explanation is fragile. If things do not go just right, the learning bias is weeded out of the population.

Second, there is a general dilemma for this style of explanation. When selection pressure is strong, assimilation is likely. The assimilation state has the higher average payoff so uncontrolled selection will usually end there. One could introduce other factors that slow the assimilation process. But those same factors weaken selection. This raises problems with whether non-adaptive forces might overwhelm the force of selection. Spatial structure is a good illustration. Introducing spatial structure helps keep groups distinct while taking the burden off of social learning biases. But taking the burden off learning also takes away the incentive to improve learning strategies.

There are countless other ways to modify the population structure. The above argument provides some reason to be skeptical that there is an unexplored model which provides a clear and consistent narrative for the evolution of similarity-biased social learning, at least for coordination games. Modifications will either increase selection pressure or decrease it. Either direction raises its own challenges. At any rate, the more specific the population structure needs to be, the less generally applicable the model is. Ethnic groups have co-existed in a huge variety of ways across history. No one population structure should be taken to be the uniquely representative model.

5 Division of labor games

The original motivation for this paper was to understand similarity-biased social learning in general. Yet the proceeding discussion has focused exclusively on ethnicity-related interpretations of the bias. This limited focus provides a tractable entry point into thinking about the evolutionary dynamics around learning strategies. But it ignores the possibilities of gender-biased and racially-

biased social learning strategies. This section considers how one might generalize this model to thinking about other forms of similarity-biased learning.

There is nothing about the model that is intrinsically tied to ethnicity. The model is an abstract computational algorithm that can be interpreted in various ways. The same model could be interpreted to apply to gender or race. Whether this application is appropriate or not depends on whether coordination games represent an evolutionary function of gender or race. Suppose there was some strategic problem in which men need to coordinate with other men and people of the same sex tend to interact more often. Then it might be advantageous to have a special learning strategy that transmits strategies amongst the men in a community. This is one way of building out an evolutionary explanation for gender-biased social learning. Given the above discussion, this explanation requires a variety of other, strong assumptions and does not seem like a very promising path.

There is another story one might tell about gender. O'Connor argues that one evolutionary function of gender is to coordinate the division of labor (O'Connor 2019). A huge variety of human societies employ a gendered division of labor. Divisions of labor offer a variety of benefits in terms of specialization, risk management, and efficiency. But they require coordination. Societies need conventions to specify who should perform which tasks. O'Connor explores a series of evolutionary game models to show that populations can latch onto sexual difference to stabilize this kind of convention. In game-theoretic terms, division of labor problems can be represented as anti-coordination games:

	A	B
A	0,0	1,1
B	1,1	0,0

If each player knows the sex of their partner and they know the convention that women perform A while men perform B, they can reliably coordinate.

Moving from traditional coordination games to anti-coordination games provides a more promising explanation of similarity-biased social learning. Previous modeling shows gender-biased social learning can co-evolve with a gendered division of labor (Saunders 2022). These results are particularly robust - the explanation does not break down under reasonable adjustments to parameter values or structural assumptions. This suggests a fairly compact evolutionary explanation of gender-biased social learning: if gender functions to solve division of labor problems, then the learning bias comes for free.

This result also suggests a path forward for ethnicity-biased or racially-biased social learning. If a primary evolutionary function of ethnicity or race is to divide labor and people can adopt different behaviors based on who they are interacting with, then the gender-biased learning model can be exported to those

contexts. Suppose two ethnic groups live in close proximity. Each group collects a distinct kind of resource and they regularly trade. Under these circumstances, one should expect ethnicity-biased social learning to evolve. The same could be said for race. This kind of ethnic division of labor is not ubiquitous in the same way the gendered division of labor is but is still observed in some cases (Bunce and McElreath 2017). If we assume that ethnic divisions of labor are uncommon, at least relative to gendered divisions, it explains why the experimental literature on learning biases finds inconsistent results. There is a very general evolutionary story to tell for gender-biased social learning. Ethnicity-biased learning is possible under the logic of evolution. But the conditions that favor its evolution are only found in some societies, at some points in time.

It might be puzzling why switching from coordination to anti-coordination games avoids the myriad problems described in section 4. Anti-coordination games with conditional strategies only have one stable state. There is no analogy to assimilation in the division of labor. Players only receive benefits when they perform different actions. They need similarity-biased learning strategies to ensure that behaviors are distributed along group lines. So even when forces like mutation and drift are introduced into the model, the population will explore possible distributions of behavioral and learning strategies until it discovers a division of labor. Once this state is achieved, it is difficult to undo it.

6 Conclusion

The aim of this paper is not to accept or reject any one particular model. Rather, exploring a series of models can show when similarity-biased social learning is selective advantageous and when it is not. That insight can, in turn, reveal the strengths and weaknesses of various evolutionary stories found in the literature. It turns out that much depends on whether one thinks the primary function of social roles is to facilitate coordination or the division of labor. Coordination-driven explanations require a very specific population structure: groups must interact but not too much. Whether the goldilocks conditions required for this kind of evolutionary story are empirically realistic is a fairly speculative judgment, one best left to the reader. By contrast, stories that center the division of labor as the function of social roles fair much better. Given that assimilation outcomes are not possible in the division of labor, few strong assumptions about the population structure are necessary.

Philosophers studying game-theoretic models of the origins of inequality would benefit from these insights. As noted in the introduction, these models typically make use of strong similarity-biased social learning assumptions. These models are also intended to be fairly general, representing a possible mechanism for generating gendered, racial, or ethnic inequality in both historical and contemporary societies. Papers in this literature do not usually offer much explicit justification for their learning assumptions. O'Connor's book *The Origins of Unfairness*, has the most detailed discussion, writing "Henrich and Henrich

(2007) point out that selective processes should favor copying those of the same gender and same ethnicity, for the very reason that doing so improves uptake of appropriate social roles and behaviors."

The results of this paper encourage greater nuance here. The story we tell about learning strategies might need to be very different for gender and ethnicity. The experimental literature shows that gender-biased learning is far more easily generated in a lab than ethnicity. Similarly, if we assume, following Henrich, Henrich, McElreath, Boyd, and Richerson, that the primary function of ethnicity is to solve coordination problems, we should be cautious about the evolutionary explanation. This suggests that bargaining models of inequality are more easily applied to gender-based inequalities than ethnic or racial ones.

If, instead, we assume that both gender and ethnicity function to divide labor, then a consistent story can be told in both cases. It is unclear whether ethnicity performs a division of labor role in a very general way across human societies. If we want to apply evolutionary bargaining models to any real cases of racial- or ethnicity-based inequality, then we should make efforts to show a division of labor actually does characterize the historical relationship between the two groups.

Finally, there is the possibility of providing non-adaptive justifications for learning assumptions. For example, racial residential segregation or social network segregation might control the flow of information in a way that mirrors similarity-biased learning. Here, population structure does all the necessary work and psychological traits are largely irrelevant. Plausibly, people share stories about bargaining for wages or resources with their friends and friends tend to be of the same social group. Regardless of what set of assumptions one finds most plausible, modelers who study inequality should be more specific about what they are assuming about learning strategies and why.

References

Amadae, S. M., and Christopher J. Watts. 2022. "Red Queen and Red King Effects in cultural agent-based modeling: Hawk Dove Binary and Systemic Discrimination." *Journal of Mathematical Sociology* 00 (00): 1–28. https://doi.org/10.1080/0022250X.2021.2012668.

Boyd, Robert, Herbert Gintis, Samuel Bowles, and Peter J. Richerson. 2003. "The evolution of altruistic punishment." *Proceedings of the National Academy of Sciences of the United States of America* 100 (6): 3531–35. https://doi.org/10.1073/pnas.0630443100.

Bright, Liam Kofi, Nathan Gabriel, Cailin O'Connor, and Olufemi O. Taiwo. 2022. "On the Stabiliy of Racial Capitalism."

Bruner, Justin P. 2019. "Minority (dis)advantage in population games." *Synthese* 196 (1): 413–27. https://doi.org/10.1007/s11229-017-1487-8.

Bunce, John Andrew, and Richard McElreath. 2017. "Interethnic Interaction, Strategic Bargaining Power, and the Dynamics of Cultural Norms: A Field Study in an Amazonian Population." *Human Nature* 28 (4): 434–56. https://doi.org/10.1007/s12110-017-9297-8.

Bussey, Kay, and Albert Bandura. 1984. "Influence of gender constancy and social power on sex-linked modeling." *Journal of Personality and Social Psychology* 47 (6): 1292–302. https://doi.org/10.1037/0022-3514.47.6.1292.

Buttelmann, David, Norbert Zmyj, Moritz Daum, and Malinda Carpenter. 2013. "Selective Imitation of In-Group Over Out-Group Members in 14-Month-Old Infants." *Child Development* 84 (2): 422–28. https://doi.org/10.1111/j.1467-8624.2012.01860.x.

Henrich, Joseph. 2016. *The Secret of Our Success: How Culture is Driving Human Evolution Domesticating Our Species and Making Us Smarter*. Princeton: Princeton University Press.

Henrich, Joseph, and Richard McElreath. 2003. "The Evolution of Cultural Evolution." *Evolutionary Anthropology* 12 (3): 123–35. https://doi.org/10.1002/evan.10110.

Henrich, Natalie, and Joseph Henrich. 2007. *Why Humans Cooperate: A Cultural and Evolutionary Explanation*. Oxford, UK: Oxford University Press.

Hoppitt, William, and Kevin N. Laland. 2013. *Social Learning: An Introduction to Mechanisms, Methods and Models*. Princeton, NJ: Princeton University Press.

Howard, Lauren H., Annette M. E. Henderson, Cristina Carrazza, and Amanda L. Woodward. 2015. "Infants' and Young Children's Imitation of Linguistic In-Group and Out-Group Informants." *Child Development* 86 (1): 259–75. https://doi.org/10.1111/cdev.12299.

Izquierdo, Luis R., Segismundo S. Izquierdo, and William H. Sandholm. 2019. "An introduction to ABED: Agent-based simulation of evolutionary game theory." *Games and Economic Behavior* 118 (c): 434–46. https://doi.org/10.1016/j.geb.2019.09.014.

Kinzler, Katherine D., Kathleen H. Corriveau, and Paul L. Harris. 2011. "Children's selective trust in native-accented speakers." *Developmental Science* 14 (1): 106–11. https://doi.org/10.1111/j.1467-7687.2010.00965.x.

Kinzler, Katherine D., Emmanuel Dupoux, and Elizabeth S. Spelke. 2007. "The native language of social cognition." *Proceedings of the National Academy of Sciences of the United States of America* 104 (30): 12577–80. https://doi.org/10.1073/pnas.0705345104.

Krieger, Andrea A. R., Corina Möller, Norbert Zmyj, and Gisa Aschersleben. 2016. "Tom is not more likely to imitate Lisa Than Ying: The influence of a model's race indicated by physical appearance on children's imitation."

Frontiers in Psychology 7 (JUN): 1–8. https://doi.org/10.3389/fpsyg.2016.00972.

Lewis, David. 1969. *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Losin, Elizabeth A.Reynolds, Macro Iacoboni, Alia Martin, and Mirella Dapretto. 2012. "Own-gender imitation activates the brain's reward circuitry." *Social Cognitive and Affective Neuroscience* 7 (7): 804–10. https://doi.org/10.1093/scan/nsr055.

Mayo-Wilson, Conor, and Kevin J. S. Zollman. 2021. "The computational philosophy: simulation as a core philosophical method." *Synthese*. https://doi.org/10.1007/s11229-020-02950-3.

Mcelreath, Richard, Adrian V Bell, Charles Efferson, Mark Lubell, Peter J Richerson, and Timothy Waring. 2008. "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): 3515–28. https://doi.org/10.1098/rstb.2008.0131.

McElreath, Richard, and Robert Boyd. 2007. *Mathematical Models of Social Evolution: A Guide for the Perplexed*. Chicago: University of Chicago Press.

McElreath, Richard, Robert Boyd, and Peter J. Richerson. 2003. "Shared norms and the evolution of ethnic markers." *Current Anthropology* 44 (1): 122–29. https://doi.org/10.1086/345689.

Mesoudi, Alex, Lei Chang, Sasha R. X. Dall, and Alex Thornton. 2016. "The Evolution of Individual and Cultural Variation in Social Learning." *Trends in Ecology and Evolution* 31 (3): 215–25. https://doi.org/10.1016/j.tree.2015.12.012.

Mohseni, Aydin, Cailin O'Connor, and Hannah Rubin. 2019. "On the emergence of minority disadvantage: testing the cultural Red King hypothesis." *Synthese*. https://doi.org/10.1007/s11229-019-02424-1.

Muthukrishna, Michael, and Joseph Henrich. 2019. "A problem in theory." *Nature Human Behaviour* 3 (3): 221–29. https://doi.org/10.1038/s41562-018-0522-1.

O'Connor, Cailin. 2017. "The cultural red king effect." *Journal of Mathematical Sociology* 41 (3): 155–71. https://doi.org/10.1080/0022250X.2017.1335723.

———. 2019. *The Origins of Unfairness: Social Categories and Cultural Evolution*. Oxford: Oxford University Press.

O'Connor, Cailin, Liam Kofi Bright, and Justin P. Bruner. 2019. "The Emergence of Intersectional Disadvantage." *Social Epistemology* 33 (1): 23–41. https://doi.org/10.1080/02691728.2018.1555870.

O'Connor, Cailin, and Justin Bruner. 2017. "Power, Bargaining, and Collaboration." In *Scientific Collaboration and Collective Knowledge: New Essays*. October 2021. Oxford, UK: Oxford University Press. https://doi.org/10.1093/oso/9780190680534.001.0001.

———. 2019. "Dynamics and Diversity in Epistemic Communities." *Erkenntnis* 84 (1): 101–19. https://doi.org/10.1007/s10670-017-9950-y.

Perry, David G., and Kay Bussey. 1979. "The social learning theory of sex differences: Imitation is alive and well." *Journal of Personality and Social Psychology* 37 (10): 1699–1712. https://doi.org/10.1037/0022-3514.37.10.1699.

Popa, Mihaela, and Roland Muehlenbernd. 2022. "Fairness and Signaling in Bargaining Games."

Saunders, Daniel. 2022. "How to Put the Cart Behind the Horse in the Cultural Evolution of Gender." *Philosophy of the Social Sciences* 52 (1-2): 81–102. https://doi.org/10.1177/00483931211049770.

Schelling, Thomas. 1960. *The Strategy of Conflict*. 1st ed. Cambridge, MA: Harvard University Press.

Shutts, Kristin, Mahzarin R. Banaji, and Elizabeth S. Spelke. 2010. "Social categories guide young children's preferences for novel objects." *Developmental Science* 13 (4): 599–610. https://doi.org/10.1111/j.1467-7687.2009.00913.x.

Shutts, Kristin, Katherine D. Kinzler, Caitlin B. McKee, and Elizabeth S. Spelke. 2009. "Social Information Guides Infants' Selection of Foods." *Journal of Cognition and Development* 10 (1-2): 1–17. https://doi.org/10.1080/15248370902966636.

Skyrms, Brian. 1990. *The Dynamics of Rational Deliberation*. London, UK: Harvard University Press.

———. 1996. *Evolution of the Social Contract*. Cambridge, UK: Cambridge University Press.

Smaldino, Paul E. 2020. "How to translate a verbal theory into a formal model." *Social Psychology* 51 (4): 207–18. https://doi.org/10.1027/1864-9335/a000425.

Wood, Lara A, Rachel L Kendal, and Emma G Flynn. 2013. "Whom do children copy ? Model-based biases in social learning." *Developmental Review* 33 (4): 341–56. https://doi.org/10.1016/j.dr.2013.08.002.