Daniel Saunders

January 2021

How to Put the Cart Behind the Horse in the Cultural Evolution of Gender

In *The Origins of Unfairness*, Cailin O'Connor outlines a novel theory of the origins of gender (O'Connor, 2019). She draws on the tools of evolutionary game theory to show how gender might have emerged as a device for solving certain classes of coordination problems. Some tasks are best completed through a specialized division of labour. But without social roles, it can be difficult to coordinate on the issue of who should perform which tasks. Who should fish and who should make pottery? Sexual differences are one salient feature in early human societies that could provide a basis for the division of labour. Once endowed with social significance, sexual difference can transform into the autonomous cultural and normative force of modern systems of gender.

Her models are illuminating but have a difficulty. She assumes that agents engage in gendered social learning as the mechanism by which successful strategies spread through a population. But this seems to put the explanatory cart before the horse – how did early humans have a well-developed system of gendered social learning before the gendered division of labour? If we want to explain the origins of gender, one should not help themself to gender-like behaviour. One possibility is that gendered social learning and the division of labour incrementally co-evolved. But no formal model of such a process is currently available.

This paper closes that explanatory gap. It develops a pair of agent-based models for exploring how various learning mechanisms interact with gendered behaviours. It finds that adding fairly minimal assumptions makes it unnecessary to stipulate that agents engage in gendered social learning. A partial gendered division of labour can emerge merely by introducing distinctive agent identities. The partial division of labour, in turn, produces an environment in which gendered social learning can evolve endogenously. The growth of gendered social learning subsequently strengthens the division of labour.

This paper is organized into three sections. The first motivates O'Connor's broad theoretical picture and diagnoses a problem of explanatory circularity. The second describes a model that produces a partial gendered division of labour via a mechanism dubbed 'strategic inertia.' The third describes a model that represents how gendered strategic behaviour and gendered social learning may co-evolve.

1 – Gender as an evolved coordination device

This section describes a game-theoretic model of the evolution of gender that has been developed in previous work (O'Connor, 2019). This paper is not a broad defense of O'Connor's view. The aim is narrower. But motivating the model makes it easier to understand this paper's contribution.

1.1 – Gender & social convention

Gender organizes our social behaviour based on real or perceived sexual differences. A variety of activities – ways of talking, labouring, having sex, dressing, expressing emotions, positioning one's body, etc – are coded as feminine or masculine. Cultural norms indicate that people are supposed to engage in gender-specific activities. In some cases, gender norms can enable smooth social coordination[1]. When two people approach a door at the same time, there is ambiguity about who should open the door for whom. If both people reach for the handle, they might awkwardly collide. If neither reaches for the door, they will

[1] They can also, of course, generate social friction. Systematic gender-based oppression is a key source of antagonism in societies. This raises a puzzle about why gender persists. The explanation developed in this paper assumes gender provides some benefits that lead to its cultural reproduction. Whether the benefits outweigh the costs is an important question for normative philosophy, but it is not a question pursued here.

awkwardly stand outside. It is a slight inconvenience to open the door, but both people prefer that someone open the door most of all. The norm of men opening the door for women removes ambiguity.

One way of modeling these facts in game-theoretic terms is to treat them as part of a convention. Game theorists understand conventions as behavioural regularities in a group, where those behaviours constitute an equilibrium in repeated coordination problems[2] (O'Connor, 2019, ch 1). Gendered behaviours are a kind of convention that facilitate coordination in the division of labour. Centering the division of labour in the explanation of the evolution of gender has considerable empirical motivation. Anthropological research finds a fairly stable, cross-cultural connection between gender and the division of labour (Bird & Codding, 2015; Murdock & Provost, 1973; Wood & Eagly, 2012). Many societies, perhaps all of them, utilize gender as a social category to assign a range of tasks to individuals. Exactly which tasks are assigned to which gender varies significantly cross-culturally. But the role of gender in dividing labour remains invariant.

O'Connor encapsulates those observations with the game in Table 1.

		Player 2	
		A	B
Player 1	A	0, 0	1, 1
	B	1, 1	0, 0

Table 1 – Division of labour game

In this game, we can think of the strategy pairs (A, B) and (B, A) as representing a division of labour. Each agent does a different task. Both benefit from the dividing labour. Under the division of labour, each individual benefits due to specialization. If the men in a society pursue fishing, they will grow more skilled in it over time. Likewise, for gender roles that assign pottery to women. The payoffs do not stem from merely accomplishing the task. Instead, the payoffs represent the returns to specialization relative to what they would accomplish without a division of labour. In principle, everyone could perform a little bit of every task. Men could fish in the morning and do pottery in the afternoon. But a society that managed labour in this way would miss out on the benefits of specialization.

Despite their mutual interests, there is a good chance the agents fail to coordinate. Neither agent knows which equilibrium the other agent will aim for. Worse, even if they agreed on a particular equilibrium, they would still need to figure out who occupies the positions of player 1 and player 2. Absent some additional piece of information, they have no way of knowing[3]. Each agent has just as much reason to think they are player 1 as player 2. O'Connor suggests that sexual difference is a salient piece of information that breaks the informational symmetry. If a male and female pair are faced with a division of labour problem,

[2] David Lewis offered an early analysis of convention in terms of games (Lewis, 1969). While the definition in this paper keeps with the spirit of Lewis, it relaxes his requirement that the convention is common knowledge among its adherents and sustained through mutual expectations. Evolutionary game theorists have found it useful to work with a more inclusive account of conventions that applies to non-human animals and simple artificial agents who have fairly low levels of rationality (O'Connor, 2019; Skyrms, 1996). The agents in this paper do not have explicit representations of common knowledge or their expectations of others' behaviour.

[3] A natural question is why agents could not just talk to decide who should do what. First, a conversational solution is more difficult in cases where coordination has to be rapid, sustained over distances, or with performed with novel partners. Second, although the evidence is still uncertain, it appears that the gendered division of labour may be significantly older than complex languages that are capable of formulating plans and negotiating roles (Sterelny, 2012).

they can assign player positions by sex and develop a convention wherein player 1 performs action A and player 2 performs action B.

This provides a general, functional account of what gender is. But interest is in its evolutionary origins. How could an early human population with only sexual differences develop a gendered division of labour?

1.2 – Evolving gender in models

To explore this, O'Connor turns to evolutionary game theory. Suppose there is a population of agents that play the coordination game in Table 1 over several rounds. The initial population is assigned a distribution of strategies. More successful strategies grow over time in proportion to their success. The growth of successful strategies represents the mechanisms of cultural transmission. We tend to mimic successful people. If some members of the group are more successful at producing food, others will copy their strategies. The process by which successful strategies spread is governed by an equation known as the replicator dynamics[4]. At some point, the population stops evolving – it has reached an evolutionarily stable state. The simplest strategies are to take the same action every round. These are referred to as unconditional strategies:

- Always perform A
- Always perform B

The strategies we are most interested in are ones that condition the agent's action on the other player's sex. These will be referred to as type-conditional strategies:

- Perform A when playing against males; B when playing against females
- Perform B when playing against males; A when playing against females

By exploring a range of evolutionary game models, we can gain an understanding of which conventions a population is likely to settle on through simple processes of cultural evolution. Two iterations of models are important for our purposes.

First, imagine a population that starts with a random distribution of unconditional strategies. O'Connor shows this population will evolve until it achieves a 50-50 split in strategies. This is straightforward. 'Always A' does well when most people are playing 'always B.' If there is a surplus of 'always B' players, then 'always A' will grow until they achieve parity.

The average payoff for both strategies at equilibrium is 0.5. Half the time agents play against agents with the same strategy and get nothing. The other half the time, they play against agents with complementary strategies and receive a payoff of one. This corresponds to the situation where the population has no gendered social conventions.

[4] There are a few ways of writing replicator dynamic equations. The simplest is the discrete time replicator dynamics. It assumes the proportion of the population playing a strategy on the next round is a function of its present population proportion and a ratio of the strategy's expected utility relative to a weighted average of the expected utility of all strategies in the population. It has the form

$$x_i' = x_i * \left(\frac{f_i(x)}{\sum_{j=1}^{n} f_j(x) * x_j}\right)$$

Where x_i is the population proportion of some strategy i, x_i' is the population proportion on the next round, $f_i(x)$ is the expected utility and the denominator is a sum of the expected utility multiplied by the population proportion for all strategies.

Second, now imagine we have two sub-populations, males and females. Agents can interact with anyone, but successful strategies are spread internally to each sub-population. There is also a larger pool of strategies; agents are randomly assigned any of the four strategies described above. This makes it possible for agents to evolve a gendered social convention.

The result is that the average payoff increases to 0.75. The population will converge to either one of two states. In the first state, males will always play A against females and females will reciprocate by playing B. When agents play against the opposite sex, they always receive a payoff of one. Internal to each sub-population, they still only coordinate half the time. They do not have sexual differences to break role symmetry, so each sub-population is in a state akin to the first model. In the second state, they simply switch actions; males always play B against females and females play A against males. The other dynamics remain the same in the second state. This shows that there is something evolutionarily attractive about gender. It allows agents to improve their performance in coordinating over a division of labour.

1.3 – The circularity problem

In the move from the first model to the second, O'Connor added the assumption that strategies only transmit within each sexual sub-population, never across sub-populations. The assumption is equivalent to assuming the relevant mechanism of cultural transmission is entirely internal to each sex. Or simply, social learning is gendered. There is good empirical support for this assumption. A large body of experimental literature finds humans prefer to imitate the behaviours of others who shared their sex (Bussey & Bandura, 1984; Losin, Iacoboni, Martin, & Dapretto, 2012; Perry & Bussey, 1979; Shutts, Banaji, & Spelke, 2010).

Despite the empirical support, assuming gendered social learning introduces a degree of explanatory circularity into these models. To form stable gendered conventions, we need agents to engage in a gendered kind of social learning. Early human societies would need gender, or something gender-like, in place to organize the learning experience. But the model is supposed to explain how early humans got gender in the first place.

O'Connor acknowledges the difficulty:

> There is a worry here, which is that same-gender imitation, and other mechanisms for enforcing proper-gendered learning, are surely at least in part a response to the existence of gender roles and norms. In other words, while the models assume this sort of learning in order to get gendered division of labor, perhaps that is putting the cart before the horse. (O'Connor, 2019, pg 66)

But she leaves it as a task for future research,

> A full account of how groups manage to get all these features in place at once is beyond this book. (pg 67)

The remainder of this paper develops an account of how it is possible for gendered strategic behaviour and gendered social learning to co-evolve.

2 – From equations to agents

This section and the next describe ways of extending O'Connor's model to either mitigate or eliminate the circularity problem. Each section relies on a technique known as agent-based modeling. It would be useful to contrast the equation-based approach described in the previous section with the agent-based approach. In replicator dynamic models, agent identity is unimportant. What matters instead is simply the proportion of the population playing the various strategies. Equation-based models effectively abstract over individual agents to represent the population-level strategic dynamics more easily. In an agent-based model (ABM), each agent is a discrete entity. Instead of describing the behaviour of the system through equations, agents are programmed to follow rules that govern their interactions[5].

2.1 – An agent-based model

Suppose there is a population of 100 agents. Half are male and half are female. They are randomly assigned a strategy at the beginning of each simulation. The strategy specifies which actions they should play during interactions with other agents. The model evolves over a series of rounds.

Each round the agents take these actions:

1. Play - Agents pick a random partner and play the coordination game. If they play complementary strategies, they get a payoff of 1. Otherwise, they get a payoff of 0.
2. Learn - Agents pick a new partner. If that partner received a higher payoff, the learning agent switches to playing their partner's strategy.
3. Mutate – Agents have a small probability of randomly switching to a new strategy. The mutation rate is controlled by a parameter and is set to zero for most trials unless indicated.

Generally, the population evolves toward a dynamic, stochastic equilibrium. The distribution of strategies fluctuates around, but does not settle on, some fixed point.

It is easy to show that the agent-based model achieves similar results to the replicator dynamic models. The average payoff provides a useful indication of how well the agents are coordinating. In the first condition, only the unconditional strategies are present in the population and the average payoff fluctuates around 0.5, as O'Connor found.

Box 1

Dark blue - Always perform action A

Light blue - Always perform action B

Orange – Perform action A vs males and B vs females

Red – Perform action B vs males and A vs females

In the second condition, the learning rule is modified so agents only learn from partners of the same sex. The average payoff rises to 0.75. Figure 1 visualizes the distribution of strategies across the population for a single round. Circles represent females, squares represent males. The next several sections present several data visualizations which employ the same color scheme. Box 1 describes which colours correspond to which strategies.

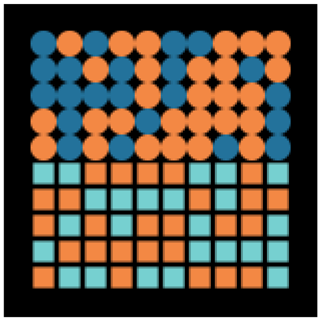

Figure 1

Notice that only circles are dark blue while only squares are light blue. This indicates strategic assortment by sex. This is what we would expect given the functional theory of gender described in section

[5] For a helpful discussion comparing equation-based and agent-based modeling strategies in the cultural evolution context, see (Alexander, 2007, chapter two). For a more general introduction to ABMs, see (Wilensky & Rand, 2015)

one. Each sex has a characteristic action they perform when interacting with the other sex. When interacting with agents of the same sex, they face uncertainty.

2.2 – The inertial mechanism

The first approach to the circularity problem is motivated by a simple question. What would happen if agents could employ conditional and unconditional strategies but did not engage in gender social learning? The mixture of strategies permits agents to engage in gendered interactions but learning dynamics do not force strategic assortment by sex. In O'Connor's first model, agents only use unconditional strategies, and the replication of strategies is not influenced by sex. In her second model, she introduced the full set of strategies but also adds gendered social learning. This section explores the space in between these two models.

If one runs the model with all four strategies but turns off gendered social learning, the average payoff in the population rises to ≈ 0.625. This is halfway between the average payoff of the first model and the second. Figure 2 depicts the average payoff over time. One interesting feature is that the average payoff starts around 0.5 but then grows until it stabilizes around the 0.625 mark. This indicates that the population is, in some sense, learning how to coordinate better than the population with only two strategies.

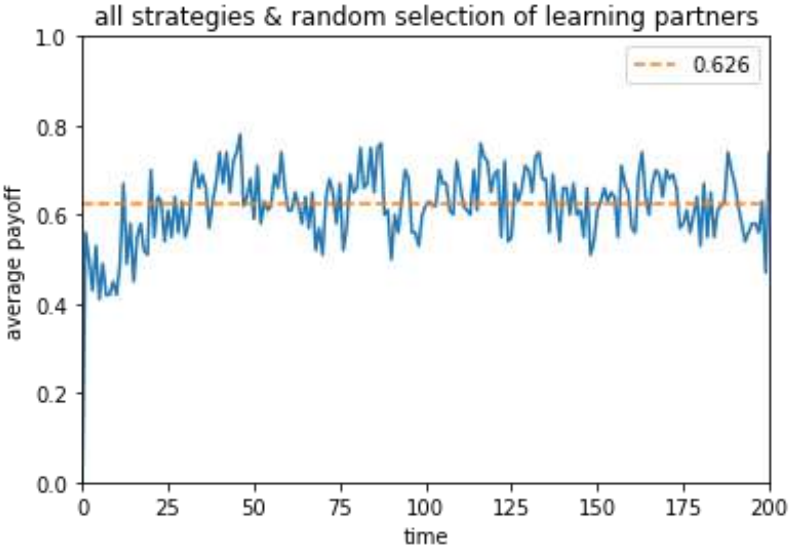

Figure 2

This result is rather surprising. If the strategies are evenly distributed by sex and pairing is random, the expected payoff for all the strategies is 0.5. The type-conditional strategies do not offer any inherent advantage over the unconditional ones. Their expected utility is only higher than 0.5 when a player's sex is predictive of what action they will take. The mere inclusion of type-conditional strategies should not make a difference unless strategies are unevenly sorted by sex. Figure 3 depicts the distribution of strategies in the population for a single round. It indicates that uneven assortment is precisely what is happening. Recall that circles represent females and squares represent males. There are three things to notice. First, about half the agents in the population are playing the red type-conditional strategy. Second, within each sex group, there is one popular unconditional strategy. For females, that is light blue. For males, that is dark blue. Third, there is one strategy that is rare for each sex. Only a handful of males play light blue, and a minority of females play dark blue.

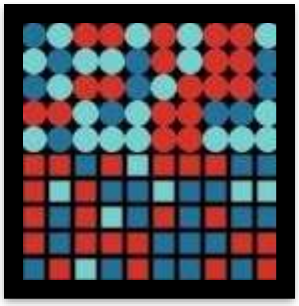

Figure 3

Figures 4 and 5 depict the frequency of each strategy in the population for, grouped by sex. They illustrate that the snapshot above is consistent with the behaviour of the model over time.

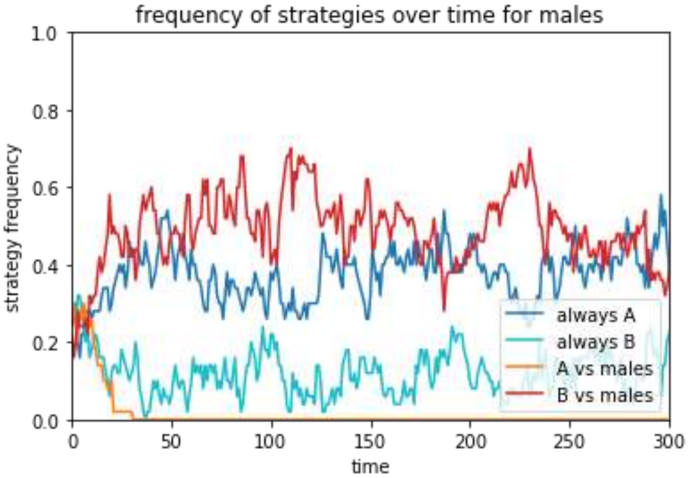

Figure 4

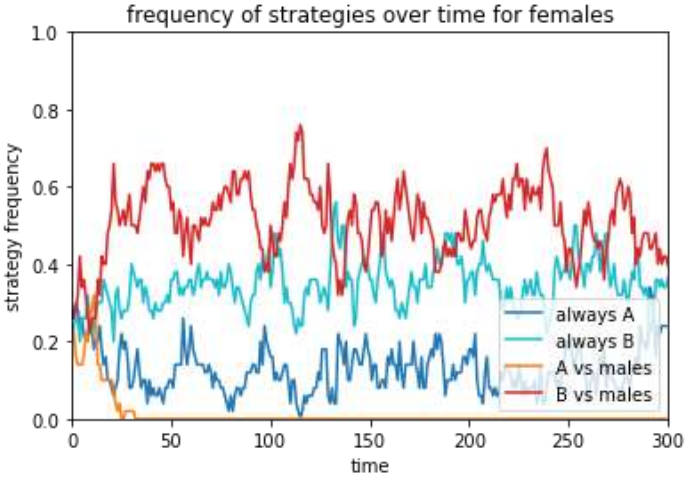

Figure 5

In this run of the model, one of the type-conditional strategies (red) quickly grows in both groups. Whatever unconditional strategy pairs well with the popular conditional strategy follows. Dark blue males play well with red agents of either sex. They also play well against the large number of light blue females. But there are still some dark blue females who do not pair well with the dark blue males. Repeated runs of the model produce similar behaviour except the strategies may be inverted: orange grows quickly, males gravitate toward light blue and females gravitate toward dark blue.

This much explains why the average payoff rises to ≈ 0.625. Strategies are partially assorted by sex. But it remains to be explained what produces this assortment in the absence of gendered social learning. The mechanism that drives these dynamics is subtle. It arises from a feature one might call 'strategic inertia.' More successful agents have little reason to change their strategy. If an agent received a payoff on a given round, they will decline to switch strategies during learning. The result is that the population can stumble into a beneficial distribution of strategies by chance and then hold onto to it via strategic inertia. In the iteration graphically depicted above, dark blue males win more often than light blue males. As such, the light blue males will be in the market for a new strategy more often. This decreases the number of light blue males as many agents are switching away. Sometimes, the light blue males will switch to dark blue or red. When they do, they will tend to keep those strategies for longer than they kept the light blue strategy.

Strategic inertial is less effective in driving assortment than gendered social learning. Under conditions of equilibria, agents employing gendered social learning will always pick a strategy that plays well against the opposite sex. Males always pick dark blue or red because they never look toward the females for strategies. But strategic inertia does not have this restrictive feature. Males are just as likely to pick up the popular female strategy. It will just not work out very well for them.

This section underscores one of the advantages of agent-based models over equation-based models. The replicator dynamics are well suited to modeling properties that are heritable. They assume that each strategy receives a certain expected payoff, and the strategies grow in proportion to that payoff. But when the payoff depends on the interaction effect between a heritable trait (strategy) and a non-heritable trait (sex), they struggle. O'Connor avoids this problem by representing sex groups as reproductively closed sub-populations. But if one wanted to relax that assumption, there is no sensible way to do that with the replicator dynamics. Introducing discrete agents solves this problem. Agents can have an immutable sex property that interacts with a mutable strategic property.

2.3 – Implications

These results show that gendered strategic behaviour can emerge in the absence of gendered social learning. The agents in this model, for the most part, perform gender-specific actions when interacting with the other sex. Men usually hunt and women usually gather in this artificial world. Whether this solves or only mitigates the circularity problem depends on how you specify the explanandum. Some gendered conventions are very strict; in Catholic societies, only men can perform the role of a priest. Other conventions are more permissive. Today, in North America, it is not uncommon for women to hold the door open for men. The degree of strictness varies across time and place. This model shows that, when it comes to explaining at least the permissive conventions, it is unnecessary to assume to gendered social learning.

The next section explores a model which can explain the formation and maintenance of the strict gendered conventions. Gendered social learning can evolve alongside gendered behaviours. The results in this section and the next are connected – the inertial mechanism is what generates a selectively advantageous environment for gendered social learners.

3 – The co-evolution of gendered learning and strategic interaction

The goal of this section is to explore the conditions under which gendered social learning can evolve. The basic setup is to introduce a variety of learning styles into the population and a mechanism for them to be selected for. This model retains all the assumptions of the previous one except agents employ one of three learning styles:

- Learn from anyone
- Learn from only agents of the same sex
- Learn from only agents of the opposite sex

Agents begin with a random learning style but can update their learning style in the same way they used to update their strategies – by selecting another agent at random, checking to see their payoff, and copying their learning style if they have a better payoff. The motivating assumption is that agents who tend to be more successful in strategic play will also tend to be better at discovering the right strategy. If successful agents find better strategies, others should copy their learning style[6].

This model has agents take an additional two actions:

4. Update learning style – Agents pick a new random partner. If the partner received a larger payoff on this round, switch to their learning style.
5. Mutate learning style – Agents have a small probability of randomly switching to a new learning style. The parameter is controlled in the same way.

3.1 – core results

1000 simulations were run in which the agents started with a random distribution of learning styles and strategies. If the population converged to uniform gendered social learning and the average payoff rose to 0.75, the run was considered a success. The success condition represents that state that O'Connor was able to achieve by assuming gendered social learning. If the model ran for 1000 rounds without achieving the success condition, it was a failure. 885 simulations were successful. This is an important result for solving the circularity problem. It demonstrates that gendered social learning is a strong attractor for populations playing coordination games where they can also employ gendered strategies. If a population could achieve some benefit from allocating strategies by gender, then populations will also be inclined towards gendered social learning. The presence of one encourages the growth of the other, in a mutually reinforcing cycle. In the other trials, random learning takes over the population and they remain stuck there until the trial expires. These simulations assume no mutations of either kind.

3.2 – Robustness by mutation rates

The robustness of the model was explored by varying two parameters: the learning style mutation rate and the strategic mutation rate. The behaviour of the model is sensitive to the rate of mutations in learning style (LS). A series of experiments explored this behaviour. The setup was the same; the model was run until it met the success condition, or it went through 1000 rounds[7]. For these initial results, the strategic

[6] There are two ways of interpreting this portion of the model. First, there could be genetic variants that code for gendered social learning and are governed by biological evolution. Second, there could be a kind of second-order social learning. Agents observe one another and imitate their learning styles. If early humans were capable of discerning who were the most successful learners in the group, then imitators could discover gendered social learning this way. There is now a range of evidence suggesting such second-order social learning is widespread in humans (Mesoudi, Chang, Dall, & Thornton, 2016).

[7] Limiting the experiment to 1000 rounds effectively *underestimates* how easy it is for gendered social learning to arise. If one allows the model to run indefinitely and set the mutation rate to a low, positive value, it will eventually discover the advantage of gendered social learning. But this can take a long time. The time limit is necessary to make running large-

mutation rate was set to 0. Table 2 displays the results of those experiments based on 1000 simulations at each mutation rate.

Table 2

mutation rate – LS	0%	0.01%	0.05%	0.10%	0.20%	0.30%	0.40%	0.50%	1%
% successful	88.5%	88.0%	92.7%	94.4%	93.5%	86.3%	70.6%	46.9%	3.4%

The general behaviour displayed in the table is that adding some level of mutation positively contributes to success but adding more causes a sharp decline. At an LS mutation rate of 1%, only 3.4% of trials were successful. One might interpret this as indicating a general lack of robustness. A more moderate interpretation is appropriate. Inspecting the behaviour of simulations under this condition reveals the level of gendered learning can fluctuate dramatically but generally stays between 100% and 50% of the population. This indicates that gendered social learning is still evolutionarily attractive, despite a disruptive mutation rate. Figures 6, 7, and 8 depict a typical trial run for each of three LS mutation rates. The graphs display the percentage of agents using gendered social learning as a function of time.

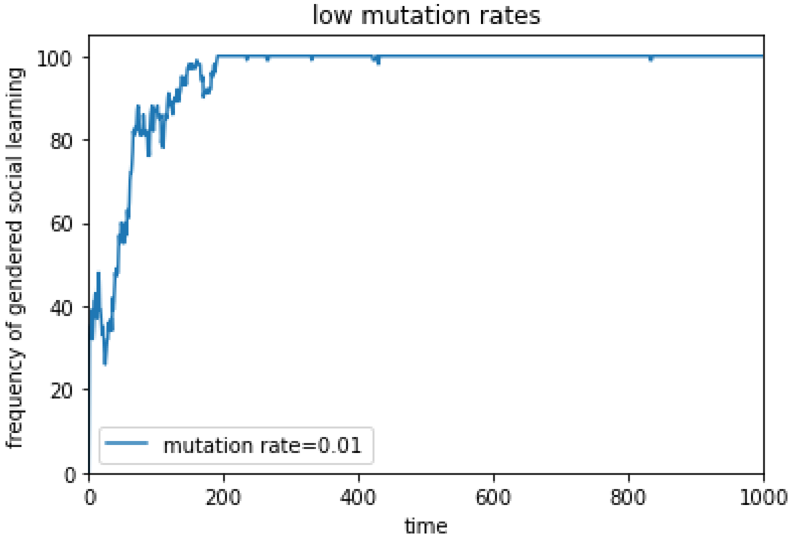

Figure 6

scale simulation experiments tractable. But if one thinks a longer time limit is a more realistic representation of the evolutionary environment, they should be more confident in O'Connor's conclusion than these results suggest.

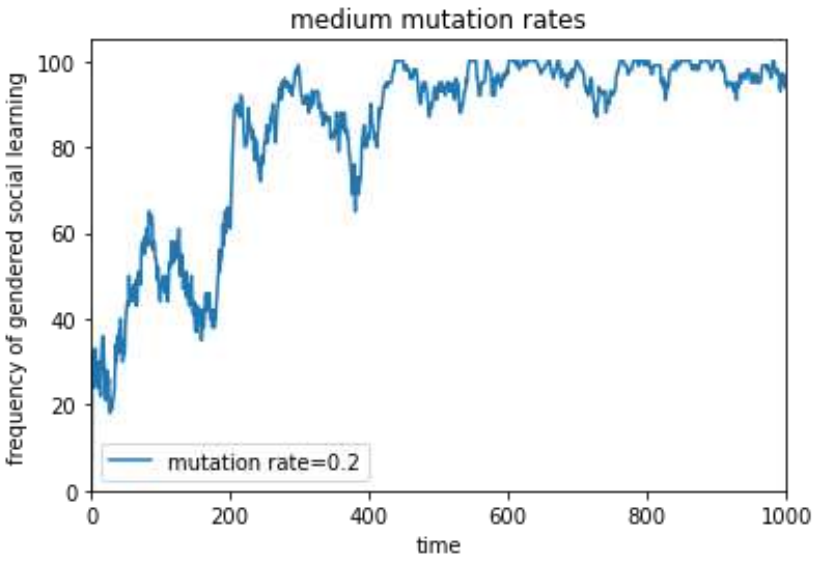

Figure 7

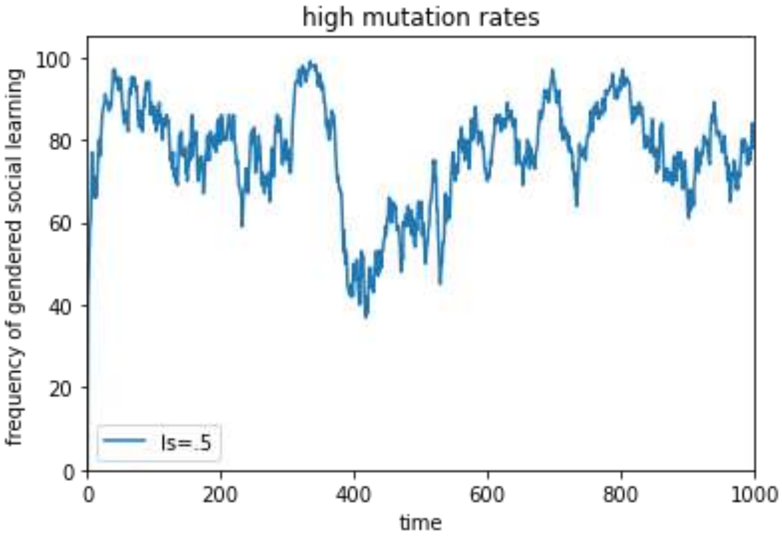

Figure 8

There are two basic insights from these data. First, the force that pulls the population toward gendered social learning is gentle. Small changes to the population can disrupt the process. If agents begin changing their learning strategy randomly, it introduces enough noise into the process that uniformity is hard to achieve. Second, despite the fragility of the process and the lack of uniformity, high rates of gendered social learning are still achieved under high mutation rates. This is broadly supportive of O'Connor's larger theory. Its also worth noting that, even in the absence of uniform social learning behaviour, the average payoff in the population still approaches 0.75.

The effect of the strategic mutation rate was also studied. The results here are more straightforward. The strategic mutation rate has a negligible effect on the rate at which the simulations achieve uniformity. I ran an experiment that tested 100 runs of the simulation at each combination of the two mutation rates. The results are presented in Table 3.

Table 3

	mutation rates – learning style								
mutation rates - strategy	0	0.01	0.05	0.1	0.2	0.3	0.4	0.5	1
0	88.5%	88%	92.7%	94.4%	93.5%	86.3%	70.6%	46.9%	3.4%
0.01	93%	87%	90%	96%	94%	85%	67%	45%	2%
0.05	85%	88%	91%	88%	97%	88%	76%	60%	1%
0.1	92%	88%	98%	96%	96%	91%	75%	47%	3%
0.2	86%	89%	91%	97%	95%	82%	58%	50%	0%
0.3	85%	85%	94%	96%	96%	88%	71%	48%	2%
0.4	86%	91%	91%	95%	97%	87%	72%	61%	1%
0.5	93%	92%	91%	98%	96%	91%	74%	46%	1%
1	84%	84%	93%	96%	97%	83%	76%	52%	0%

There is some variation running vertically in the columns, but it appears largely random. Running a larger volume of simulations shrinks the amount of variation. Table 4 displays the results of an experiment in which the learning style mutation rate was held constant at 0.01% but ran 1000 simulations at each strategic mutation rate.

Table 4

	Mutation rate - strategy								
%	0%	0.01%	0.05%	0.1%	0.2%	0.3%	0.4%	0.5%	1%
successful	89.2%	88.7%	87.0%	89.1%	88.9%	87.2%	89.4%	89.6%	86.6%

Notice that there is little variation. The model is successful at about 88.5% ± 1.5% for any mutation rate. The general conclusion is that the behaviour of the model is highly robust with respect to the strategic mutation rate.

3.3 - Growing gender from scratch

The above results go some of the way to solving the circularity problem. But they do not go all the way. In those trials, the population starts with a large proportion of type-conditioners in the mix. There could be a kind of critical mass phenomenon; gendered social learning is only culturally selected for in populations that have sufficiently high levels of gendered strategic interaction. The next set of results rules out that possibility.

The model was simulated with populations that initially only contain unconditional strategies and randomly distributed preferences for learning. Type-conditional strategies only enter the population via spontaneous mutations. The main effect of this change is that the model takes far longer to discover the advantage of gendered social learning. But over the long run, it is still consistently successful. How long it takes depends heavily on the strategic mutation rate.

Designing experiments that are both informative and feasible is more difficult for this setup. A time limit of 1000 rounds is usually too short to observe the convergence, but longer time limits massively increase computing time for performing large samples of trials. Fortunately, they are unnecessary. Once a large enough group of type-conditioners have entered the population, the behaviour of the model becomes identical with previous sections. Instead of studying the model statistically, it is more informative to discuss the mechanisms that allow for the growth of type-conditioning strategies.

In the absence of type-conditioners, there is no advantage to using one type of social learning over another. Importantly, when agents are only using unconditional strategies, nothing penalizes any type of learning. The population drifts between uniformity over the three learning strategies. Depending on which learning strategy is dominant, three different things can happen. First, if gendered social learning happens to be dominant, the conditions are favourable for the invasion of type-conditioners. The population will tend to segregate strategies by sex and the presence of type-conditioners will raise the average payoff of those who play their sex-specific strategy. Second, if indiscriminate social learning happens to be dominant, type-conditioners can still invade but they offer only a limited advantage to the population. Sometimes type-conditioning spreads to a sizable group of the population but it does not drive the payoffs up to 0.75. Third, if anti-gendered social learning is dominant, type-conditioners cannot successfully invade. Anti-gendered social learning effectively sorts the unconditional strategies evenly between the two sexes. This means type-conditional strategies offer no advantage to their users or other members of the population. Invaders must wait until the learning environment randomly drifts into one of the other two states. But that random transformation is inevitable given sufficient time. The general conclusion is that changing the initial distribution of strategies only changes how it takes for gender to emerge, not the core dynamics that drive that emergence.

Conclusion

This paper has developed two models which either mitigate or solve the circularity problem in O'Connor's explanation of the evolution of gender. The inertial mechanism can generate a modest gendered division of labour. This generates an environment that makes gendered social learning selectively advantageous. Agents who learn from same-sex partners will almost never select a bad strategy once the population is equilibrium. The synergy between the inertial mechanism and the co-evolutionary mechanism help to explain why the effects examined in section three can take so long. The inertial population will randomly drift between different configurations of strategies and learning styles. At some point, the population discovers a partial gendered division of labour via the inertial mechanism. If the learning conditions are right, they will reinforce that division. But sometimes the population is employing opposite-sex and random learning, diminishing the effect of the inertial mechanism.

The explanatory contribution of these models goes beyond a technical issue in O'Connor's book. They also provide a framework for understanding why there is variance in the strictness of gendered conventions. Building gendered social learning into the assumptions of the model produces gendered conventions that are much stronger than many observed in real life. Many gendered conventions are sustained with only partial compliance. Each of the two mechanisms described in this paper underscores the fragility of gendered conventions. The inertial mechanism produces a stable and consistent level of nonconformist behaviour. The co-evolutionary mechanism produces uniform gendered social learning only when it is undisturbed by mutations. The presence of mutations drives the population into a state with small but regular deviations from gendered conventions. These results suggest that there is likely a spectrum of gendered conventions and learning mechanisms that could evolve with various levels of strictness. Predicting whether strict or permissive conventions are likely to evolve in each context requires further investigation.

References

Alexander, J. M. (2007). *The Structural Evolution of Morality*. Cambridge, UK: Cambridge University Press.

Bird, R. B., & Codding, B. F. (2015). The Sexual Division of Labor. *Emerging Trends in the Social and Behavioral Sciences*, 1–16. https://doi.org/10.1002/9781118900772.etrds0300

Bussey, K., & Bandura, A. (1984). Influence of gender constancy and social power on sex-linked modeling. *Journal of Personality and Social Psychology*, *47*(6), 1292–1302. https://doi.org/10.1037/0022-3514.47.6.1292

Lewis, D. (1969). *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Losin, E. A. R., Iacoboni, M., Martin, A., & Dapretto, M. (2012). Own-gender imitation activates the brain's reward circuitry. *Social Cognitive and Affective Neuroscience*, *7*(7), 804–810. https://doi.org/10.1093/scan/nsr055

Mesoudi, A., Chang, L., Dall, S. R. X., & Thornton, A. (2016). The Evolution of Individual and Cultural Variation in Social Learning. *Trends in Ecology and Evolution*, *31*(3), 215–225. https://doi.org/10.1016/j.tree.2015.12.012

Murdock, G. P., & Provost, C. (1973). Factors in the Division of Labor by Sex : A Cross-Cultural Analysis. *Ethnology*, *12*(2), 203–225.

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

Perry, D. G., & Bussey, K. (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

Shutts, K., Banaji, M. R., & Spelke, E. S. (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

Skyrms, B. (1996). *Evolution of the Social Contract*. Cambridge, UK: Cambridge University Press.

Sterelny, K. (2012). *The Evolved Apprentice: How Evolution Made Humans Unique*. Cambridge, MA: MIT Press.

Wilensky, U., & Rand, W. (2015). *An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with Netlogo*. Cambridge, MA: MIT Press.

Wood, W., & Eagly, A. H. (2012). Biosocial Construction of Sex Differences and Similarities in Behavior. In *Advances in Experimental Social Psychology* (1st ed., Vol. 46). https://doi.org/10.1016/B978-0-12-394281-4.00002-7