PHPE 308M/PHIL 209F
Fairness

Eric Pacuit, University of Maryland

September 22, 2025

Nash Bargaining Game

Brian Skyrms (2012). Chapters 1 & 2 in *Evolution of the Social Contract*. Cambridge University Press.

J. McKenzie Alexander and B. Skyrms (1999). *Bargaining with Neighbors: Is Justice Contagious*. Journal of Philosophy, 96(11), pp. 588 - 598.

Player 1 \ Player 2	1/3	1/2	2/3
1/3	$\frac{1}{3}, \frac{1}{3}$	$\frac{1}{3}, \frac{1}{2}$	$\frac{1}{3}, \frac{2}{3}$
1/2	$\frac{1}{2}, \frac{1}{3}$	$\frac{1}{2}, \frac{1}{2}$	$0, 0$
2/3	$\frac{2}{3}, \frac{1}{3}$	$0, 0$	$0, 0$

Skyrms on the Ultimatum Game

Richard Thaler chose the ultimatum game as the subject for the initial article in a series on anomalies in economics—an anomaly being "an empirical result which requires implausible assumptions to explain within the rational choice paradigm."

Skyrms on the Ultimatum Game

Richard Thaler chose the ultimatum game as the subject for the initial article in a series on anomalies in economics—an anomaly being "an empirical result which requires implausible assumptions to explain within the rational choice paradigm." But we have a clear violation of the rational choice paradigm here only on the assumption that, for these subjects, utility = income.

Skyrms on the Ultimatum Game

Richard Thaler chose the ultimatum game as the subject for the initial article in a series on anomalies in economics—an anomaly being "an empirical result which requires implausible assumptions to explain within the rational choice paradigm." But we have a clear violation of the rational choice paradigm here only on the assumption that, for these subjects, utility = income. From the standpoint of rational choice theory, the subjects' utility functions are up to them. There is no principled reason why **norms of fairness** cannot be reflected in their utilities in such a way as to make their actions consistent with the theory of rational choice.

Skyrms on the Ultimatum Game

Richard Thaler chose the ultimatum game as the subject for the initial article in a series on anomalies in economics—an anomaly being "an empirical result which requires implausible assumptions to explain within the rational choice paradigm." But we have a clear violation of the rational choice paradigm here only on the assumption that, for these subjects, utility = income. From the standpoint of rational choice theory, the subjects' utility functions are up to them. There is no principled reason why **norms of fairness** cannot be reflected in their utilities in such a way as to make their actions consistent with the theory of rational choice.

Appeal to norms of fairness, however, hardly constitutes an explanation in itself. Why do we have such norms? Where do they come from? How could they evolve?

(Skryms, p. 29)

The projected evolutionary explanation seems to fall somewhat short. The best we might say on the basis of pure replicator dynamics is that **fixation of fair division is more likely than not, and that polymorphisms far from fair division are quite unlikely**.

The projected evolutionary explanation seems to fall somewhat short. The best we might say on the basis of pure replicator dynamics is that **fixation of fair division is more likely than not, and that polymorphisms far from fair division are quite unlikely**.

Two solutions

- ✓ **Inject some probability**: Every once and a while a member of the population just picks a strategy at random and tries it out perhaps as an experiment, perhaps just as a mistake.
- **Add correlation of players with the same strategy**: There is a higher probability of playing the game with players of the same strategy.

Correlated Strategies

Let us...replace the assumption of random encounters with one of positive correlation between like strategies.

Correlated Strategies

Let us...replace the assumption of random encounters with one of positive correlation between like strategies.

It is evident that in the extreme case of perfect correlation, stable polymorphisms are no longer possible. Strategies that demand more than $1/2$ meet each other and get nothing. Strategies that demand less than $1/2$ meet each other and get what they demand. The fittest strategy is that which demands exactly $1/2$ of the cake....

Correlated Strategies

Let us...replace the assumption of random encounters with one of positive correlation between like strategies.

It is evident that in the extreme case of perfect correlation, stable polymorphisms are no longer possible. Strategies that demand more than $1/2$ meet each other and get nothing. Strategies that demand less than $1/2$ meet each other and get what they demand. The fittest strategy is that which demands exactly $1/2$ of the cake....

In the real world, both random meeting and perfect correlation are likely to be unrealistic assumptions. The real cases of interest lie in between.

(Skyrms, p. 18)

Example: Correlating Strategies

Let $0 \leq \epsilon \leq 1$ be a **level of correlation**.

- $Pr_t(Modest \mid Modest)$ is the proportion playing *Modest* against *Modest*
  $Pr_t(Modest \mid Modest) = Pr_t(Modest) + \epsilon * (1 - Pr_t(Modest))$

Example: Correlating Strategies

Let $0 \leq \epsilon \leq 1$ be a **level of correlation**.

- $Pr_t(Modest \mid Modest)$ is the proportion playing *Modest* against *Modest*
  $Pr_t(Modest \mid Modest) = Pr_t(Modest) + \epsilon * (1 - Pr_t(Modest))$

- $Pr_t(Fair \mid Modest)$ is the proportion playing *Fair* against *Modest*
  $Pr_t(Fair \mid Modest) = Pr_t(Fair) - \epsilon * (Pr_t(Fair))$

Example: Correlating Strategies

Let $0 \leq \epsilon \leq 1$ be a **level of correlation**.

- $Pr_t(Modest \mid Modest)$ is the proportion playing *Modest* against *Modest*
  $Pr_t(Modest \mid Modest) = Pr_t(Modest) + \epsilon * (1 - Pr_t(Modest))$

- $Pr_t(Fair \mid Modest)$ is the proportion playing *Fair* against *Modest*
  $Pr_t(Fair \mid Modest) = Pr_t(Fair) - \epsilon * (Pr_t(Fair))$

- $Pr_t(Greedy \mid Modest)$ is the proportion playing *Greedy* against *Modest*
  $Pr_t(Greedy \mid Modest) = Pr_t(Greedy) - \epsilon * (Pr_t(Greedy))$

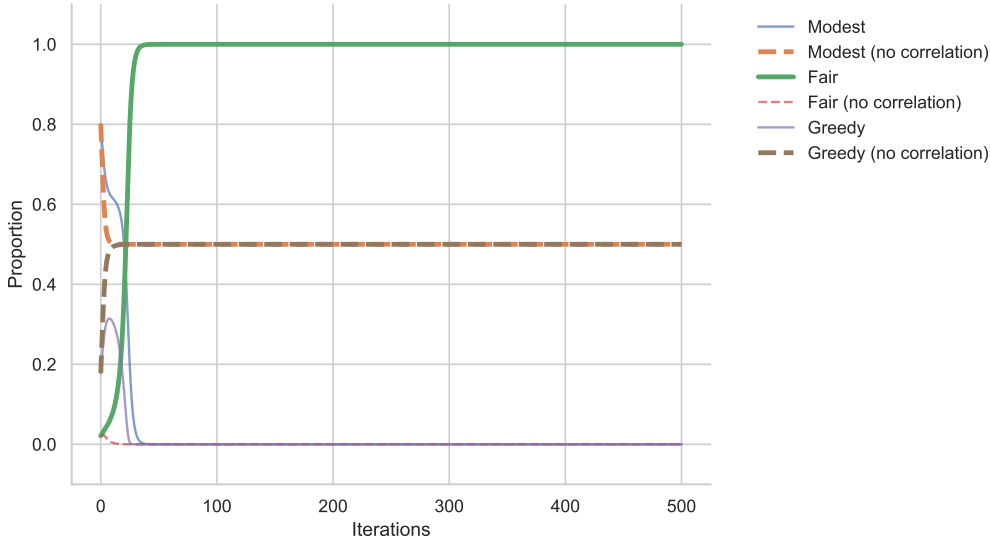
Example: Correlating Strategies

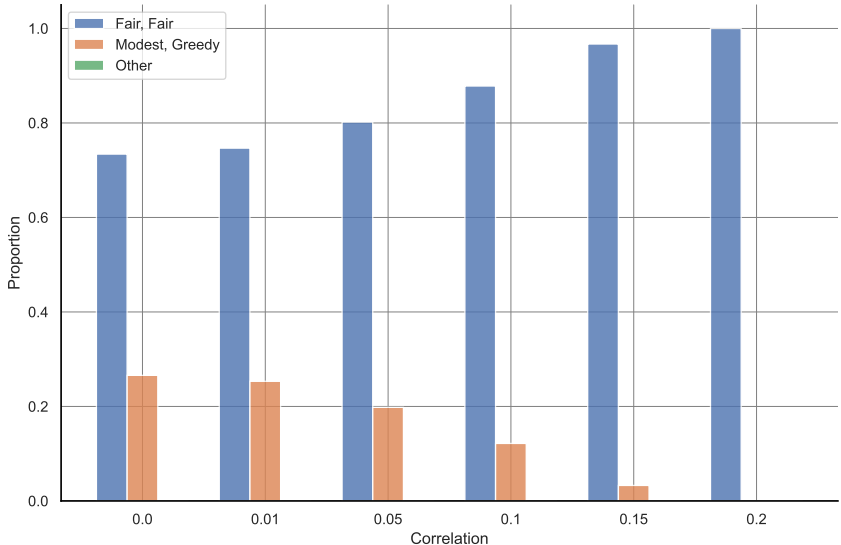
$$f\_Modest_t = Pr_t(Modest) * \frac{1}{3} + Pr_t(Fair) * \frac{1}{3} + Pr_t(Greedy) * \frac{1}{3}$$

Example: Correlating Strategies

$$f\_Modest_t = Pr_t(Modest) * \frac{1}{3} + Pr_t(Fair) * \frac{1}{3} + Pr_t(Greedy) * \frac{1}{3}$$

$$Pr_t(Modest \mid Modest) * \frac{1}{3} + Pr_t(Fair \mid Modest) * \frac{1}{3} + Pr_t(Greedy \mid Modest) * \frac{1}{3}$$


Modest
Modest (no correlation)
Fair
Fair (no correlation)
Greedy
Greedy (no correlation)
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
Iterations


Fair, Fair
Modest, Greedy
Other
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
0.0
0.01
0.05
0.1
0.15
0.2
Correlation

What is the justification for adding a correlation factor, though? Once Skyrms relaxes the requirement of random interactions in the population, and allows some degree of assortative interactions, we need to hear a justification for assuming that the likely departure from random interactions will be toward correlation in particular. Why think that individuals are especially likely to meet others playing the same strategy as they play? (D'Arms, Batterman, and Gorny, p. 92)

Justin D'Arms, Robert Batterman, and Krzyzstof Gorny (1998). *Game Theoretic Explanations and the Evolution of Justice*. Philosophy of Science, 65(1), pp. 76-102.

Local Interaction

J. McKenzie Alexander and B. Skyrms (1999). *Bargaining with Neighbors: Is Justice Contagious*. Journal of Philosophy, 96(11), pp. 588 - 598.

J. McKenzie Alexander (2000). *Evolutionary Explanations of Distributive Justice*. Philosophy of Science, 67(3), pp. 490 - 516.

The dynamics is driven by imitation. Individuals imitate the most successful person in the neighborhood. A generation an iteration of the discrete dynamics has two stages:

1. Each individual plays the Nash bargaining game with each of her neighbors using her current strategy. Summing the payoffs gives her current success level.
2. Each player looks around her neighborhood and changes her current strategy by imitating her most successful neighbor, providing that her most successful neighbor is more successful than she is; otherwise, she does not switch strategies. (Ties are broken by a coin flip.)

Neighborhoods

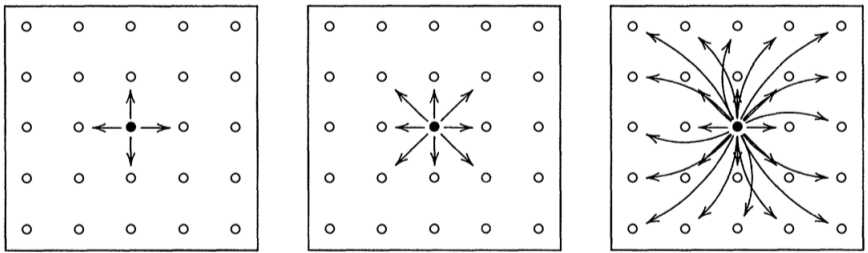


Figure 1. Three common neighborhoods defined on a square lattice.

Dynamics

1. **Imitate the best neighbor**: Each player looks at her neighbors and adopts the strategy of the neighbor who did the best, where "best" means "earned the highest score."

Dynamics

1. **Imitate the best neighbor**: Each player looks at her neighbors and adopts the strategy of the neighbor who did the best, where "best" means "earned the highest score."

2. **Imitate with probability proportional to success**: Assigns to every neighbor q who did better than the player p a nonzero probability that p will adopt q's strategy.

Dynamics

1. **Imitate the best neighbor**: Each player looks at her neighbors and adopts the strategy of the neighbor who did the best, where "best" means "earned the highest score."

2. **Imitate with probability proportional to success**: Assigns to every neighbor q who did better than the player p a nonzero probability that p will adopt q's strategy.

3. **Imitate best average payoff**: Calculate the average payoff of each strategy in their neighborhood and select the one with the highest value.

Dynamics

1. **Imitate the best neighbor**: Each player looks at her neighbors and adopts the strategy of the neighbor who did the best, where "best" means "earned the highest score."

2. **Imitate with probability proportional to success**: Assigns to every neighbor q who did better than the player p a nonzero probability that p will adopt q's strategy.

3. **Imitate best average payoff**: Calculate the average payoff of each strategy in their neighborhood and select the one with the highest value.

The general question of how one's choice of the update rule affects the limit form of the model remains an open and difficult problem.

	Bargaining with Neighbors		Bargaining with Strangers	
	A	*B*	*C*	*D*
0-10	0	0	0	0
1-9	0	0	0	0
2-8	0	0	54	57
3-7	0	0	550	556
4-6	26	26	2560	2418
fair	9972	9973	6833	6964

Table 2: Convergence results for five series of 10,000 trials

Sometimes we bargain with neighbors, sometimes with strangers. The dynamics of the two sorts of interaction are quite different.

Sometimes we bargain with neighbors, sometimes with strangers. The dynamics of the two sorts of interaction are quite different.

Bargaining with neighbors almost always converges to fair division and convergence is remarkably rapid.

Sometimes we bargain with neighbors, sometimes with strangers. The dynamics of the two sorts of interaction are quite different.

Bargaining with neighbors almost always converges to fair division and convergence is remarkably rapid.

Both bargaining with strangers and bargaining with neighbors are artificial abstractions. In initial phases of human cultural evolution, bargaining with neighbors may be a closer approximation to the actual situation than bargaining with strangers. The dynamics of bargaining with neighbors strengthens the evolutionary explanation of the norm of fair division.

Many roads lead to the egalitarian norm.

Many roads lead to the egalitarian norm. In a finite population, in a finite time, **where there is some random element in evolution**, some reasonable amount of divisibility of the good and **some correlation**, we can say that it is likely that something close to share and share alike should evolve in dividing-the-cake situations.

Many roads lead to the egalitarian norm. In a finite population, in a finite time, **where there is some random element in evolution**, some reasonable amount of divisibility of the good and **some correlation**, we can say that it is likely that something close to share and share alike should evolve in dividing-the-cake situations. If the equal split is a convention in such situations, it is no surprise that greedy players should be despised or ostracized, since they spoil things for those with whom they interact. This is, perhaps, a beginning of an explanation of the origin of our concept of justice. (Skyrms, p. 21-22)

Unfairness

Cailin O'Connor (2022). *Why Natural Social Contracts are Not Fair*. forthcoming in *New Social Contract Theory*.


the
origins of
unfairness
cailin
o'connor
social
categories
and
cultural
evolution

[T]hese models show that fair conventions of behavior do tend to emerge naturally from an uncoordinated "state of nature". They support the idea that natural social contracts tend to favor equality.

[T]hese models show that fair conventions of behavior do tend to emerge naturally from an uncoordinated "state of nature". They support the idea that natural social contracts tend to favor equality.

Of course, when we look at real world conventions and norms regarding the division of resources, fairness is not typically the rule....despite the high ideals and optimism of traditional social contract theorists, the real world is rife with inequity....

[T]hese models show that fair conventions of behavior do tend to emerge naturally from an uncoordinated "state of nature". They support the idea that natural social contracts tend to favor equality.

Of course, when we look at real world conventions and norms regarding the division of resources, fairness is not typically the rule....despite the high ideals and optimism of traditional social contract theorists, the real world is rife with inequity....**How do we square these observations with the modeling literature showing that fairness emerges naturally via cultural evolution?**

The answer is that we need to add **social categories** to these models. A social category is a recognizable group within a society. Most important to us here are primary categories, which Ridgeway (2011) describes as the small number of social categories most generally used for coordinating behavior. Across societies, these always include gender and age, and often also include race, religion, caste, or class.

Tags

Our model will involve a population with two groups (representing social categories) that each have a different arbitrary *tag*. The tags might be “green” and “yellow”, for example, or “star-belly” and “plain belly”.

Tags

Our model will involve a population with two groups (representing social categories) that each have a different arbitrary *tag*. The tags might be "green" and "yellow", for example, or "star-belly" and "plain belly".

Agents in this model play the bargaining game...but in doing so may condition their strategy on the tag of their partner.

Tags

Our model will involve a population with two groups (representing social categories) that each have a different arbitrary *tag*. The tags might be "green" and "yellow", for example, or "star-belly" and "plain belly".

Agents in this model play the bargaining game...but in doing so may condition their strategy on the tag of their partner.

For example, an agent in the green group might play Medium against other greens, and Low against yellows. We can label this two part strategy, listing the in-group strategy first, as follows: $\langle Medium, Low \rangle$. For now, we can also assume that agents learn from in-group members only. I.e., a yellow will only copy the strategies of other yellows.