# *Evolution and the Social Contract*

*BRIAN SKYRMS*

The Tanner Lectures on Human Values

Delivered at

The University of Michigan
November 2, 2007

BRIAN SKYRMS is Distinguished Professor of Logic and Philosophy of Science and of Economics at the University of California, Irvine, and Professor of Philosophy and Religion at Stanford University. He graduated from Lehigh University with degrees in Economics and Philosophy and received his Ph.D. in Philosophy from the University of Pittsburgh. He is a Fellow of the American Academy of Arts and Sciences, the National Academy of Sciences, and the American Association for the Advancement of Science. His publications include *The Dynamics of Rational Deliberation* (1990), *Evolution of the Social Contract* (1996), and *The Stag Hunt and the Evolution of Social Structure* (2004).

# DEWEY AND DARWIN

Almost one hundred years ago John Dewey wrote an essay titled "The Influence of Darwin on Philosophy." At that time, he believed that it was really too early to tell what the influence of Darwin would be: "The exact bearings upon philosophy of the new logical outlook are, of course, as yet, uncertain and inchoate." But he was sure that it would not be in providing new answers to traditional philosophical questions. Rather, it would raise new questions and open up new lines of thought. Toward the old questions of philosophy, Dewey took a radical stance: "Old questions are solved by disappearing . . . while new questions . . . take their place."

I don't claim that the old philosophical questions will disappear, but my focus is the new ones. Evolutionary analysis of the social contract—be it cultural or biological evolution—*does not tell you what to do.* Rather, it attempts to investigate how social conventions and norms evolve—how social contracts that we observe could have evolved and what alternative contracts are possible.

The tools for a Darwinian analysis of the social contract are those of evolutionary game theory. From the theory of games it takes the use of simple stylized models of crucial aspects of human interaction; from evolution it takes the use of adaptive dynamics. The dynamics need not have its basis in genetics; it may as well be a dynamic model of cultural evolution or of social learning (Björnerstedt and Weibull 1996; Schlag 1998; Samuelson 1997; Weibull 1995). No part of what follows implies any kind of genetic determinism or innateness hypothesis. Problems of cooperation may be solved by genetic evolution in some species and by cultural evolution in others.

Here are three features of the evolutionary approach to bear in mind in the ensuing discussion:

1. Different social contracts have evolved in different circumstances.
2. Existing social contracts are not altogether admirable.
3. We can try to change the social contract.

## Correlation and the Evolution of Cooperation

Cooperation may be easy or hard, or somewhere in between. Here is a game-theory model of an easy problem of cooperation; I call it Prisoner's Delight:

PRISONER'S DELIGHT

| | *Cooperate* | *Defect* |
|---|---|---|
| Cooperate | 3 | 1 |
| Defect | 2 | 0 |

The entries in the table represent the payoffs of Row's strategy when played against Column's strategy. If your partner cooperates, you are better off if you do as well, for a payoff of 3 rather than 2. If your partner defects, you are still better off cooperating for 1 rather than 0, although your partner does even better with 2. So no matter what your partner does, you are better off cooperating. Your partner is in the same situation, and reasons likewise. It is easy to cooperate.

We can change the story a little bit to fit a different kind of situation. Perhaps if your partner defects, your attempt at cooperation is counterproductive. You are better off defecting. This change gives us the game known as the Stag Hunt:

STAG HUNT

| | *Cooperate* | *Defect* |
|---|---|---|
| Cooperate | 3 | 1 |
| Defect | 2 | 2 |

In the Stag Hunt, what is best for you depends on what your partner does. If you both cooperate, you are both doing the best thing given your partner's action—likewise if you both defect. Cooperation is more difficult. It is an equilibrium, but not the only one.

Another modification of the situation calls for a different game. Suppose that defecting against a cooperator actually pays off. Then we have the Prisoner's Dilemma:

PRISONER'S DILEMMA

| | *Cooperate* | *Defect* |
|---|---|---|
| Cooperate | 3 | 1 |
| Defect | 4 | 2 |

Your optimal act again no longer depends on what your partner does. Now it is to defect. Cooperation is hard.

Each of these games is a reasonable model of some social interactions. Two men sit in a rowboat, one behind the other. Each has a set of oars.

They have been out fishing, and a hot dinner awaits them across the lake. If one doesn't row for some reason, the other will row to get them there; if one does row, the other prefers to row also to get home faster. This is *Prisoner's Delight.* Cooperation is easy. Now change the picture. They sit side by side, and each has one oar. One man rowing alone just makes the boat go in circles. This is a *Stag Hunt.*[1] Back to the first rowboat with two sets of oars, but take away the hot dinner on the opposite shore and suppose that the men are very tired. They could camp on this shore for the night, although they would prefer to get back to the opposite shore. But either prefers not to row no matter what the other does. This is a *Prisoner's Dilemma.*

Consider any reasonable adaptive dynamics operating on a large population of individuals paired at random to play one of these games. Then in Prisoner's Delight cooperation takes over the population, in Prisoner's Dilemma defection goes to fixation, and in Stag Hunt one or the other may prevail depending on the initial composition of the population.

There is an enormous literature devoted to explaining the evolution of cooperation in the Prisoner's Dilemma, while the other two games are relatively neglected. Everyone wants to crack the hardest problem—the *evolution of altruism.*[2]

All these accounts of cooperation in Prisoner's Dilemma either (1) use an interaction that is not really a Prisoner's Dilemma or (2) use pairings to play the game that are not random. It must be so.[3] Suppose the interaction is a Prisoner's Dilemma, and pairings are random in a large population. Then cooperators and defectors each have the same proportion of partners of each type. Defectors must on average do better than cooperators. Replicator dynamics increases the proportion of the population of the type that does better, and that's all there is to it.

But if nature somehow arranges for positive correlation—for cooperators to meet cooperators and defectors to meet defectors more often than they would with random matching—then it is possible for cooperators to do better than defectors. The point is obvious if we consider perfect correlation. Then the relevant comparison in payoffs is not vertical but diagonal:

[1] This is David Hume's famous example from the *Treatise of Human Nature.* For more game theory in Hume, see Vanderschraaf 1998.

[2] Binmore (1994) devotes a chapter titled "Squaring the Circle in the Social Sciences" to attempts to justify cooperation in the one-shot Prisoner's Dilemma.

[3] This is the "Iron Rule of Selfishness" of Bergstrom (2002). See also Eshel and Cavalli-Sforza 1982.

PRISONER'S DILEMMA

| | *Cooperate* | *Defect* |
|---|---|---|
| Cooperate | **3** | 1 |
| Defect | 4 | **2** |

Every explanation of the evolution of cooperation in real Prisoner's Dilemmas—kin selection, group selection, repeated games, spatial interaction, static and dynamic interaction networks, and all the rest—works by providing a mechanism that induces correlation in plays of the Prisoner's Dilemma. This was clear to William Hamilton and to George Price back in the 1960s (Hamilton 1964, 1995; Price 1970; Eshel and Cavalli-Sforza 1982; Frank 1995). (A version of Hamilton's rule for kin selection can be derived just from the positive correlation.)

But to say that a mechanism can sometimes generate enough positive correlation to maintain cooperation in the Prisoner's Dilemma is not to say that it can always do so. In some circumstances correlation may fall short. Thus, in each of these accounts, an examination of specific correlation mechanisms is of interest. And often a scenario can be analyzed *both* as Prisoner's Dilemma with correlation *and* as a larger game in which the plays of Prisoner's Dilemma are embedded. I will illustrate with two examples.

Axelrod (1984) directs our attention to the *shadow of the future.* Cooperation may be maintained not by immediate payoffs but by the consequences of current actions on future cooperative behavior of partners. In this he follows Thomas Hobbes and David Hume:

> HOBBES: He, therefore, that breaketh his Covenant, and consequently declareth that he think that he may with reason do so, cannot be received into any society that unite themselves for Peace and Defense, but by the error of them that receive him.
>
> HUME: Hence I learn to do a service to another, without bearing him any real kindness; because I foresee, that he will return my service, in expectation of another of the same kind, and in order to maintain the same correspondence of good offices with me and with others.

Axelrod, following John Nash and other founding fathers of game theory,[4] analyzes the shadow of the future using the theory of indefinitely

[4] The use of discounted repeated games to explain cooperation in Prisoner's Dilemma is already to be found in Luce and Raiffa 1957, with no claim of originality. John Nash is reported to have invented the explanation in conversation.

repeated games. Suppose that the probability that the Prisoner's Dilemma will be repeated another time is constant. This (somewhat far-fetched) idealization of geometrical discounting of the future allows us to sum an infinite series and compute the expected payoffs of strategies in the large repeated game. For simplicity, we consider only two strategies in the repeated game, *Always Defect* and *Tit for Tat.* A Tit for Tat player initially cooperates and then does what was done to him in the preceding round, and Always Defect is self-explanatory. Two players remain matched for the whole repeated game, but this restrictive assumption can be relaxed in more complicated "community enforcement" models (Sugden 1986; Kandori 1992; Milgrom, North, and Weingast 1990; Nowak and Sigmund 1998).

Since the payoffs on each individual play are those of Prisoner's Dilemma, the strategies in the repeated games must induce a correlation between individual plays of Cooperate and Defect if cooperation is not to be driven to extinction. The presence of Tit for Tat players in the population is this correlation device. They *always* cooperate with each other and quickly learn to defect against defectors.

What is the larger game in which plays of Prisoner's Dilemma are embedded? Using the version of Prisoner's Dilemma given before and probability of another trial as six-tenths, we get:

| | *Tit for Tat* | *Always Defect* |
|---|---|---|
| Tit for Tat | 7.5 | 4 |
| All Defect | 7 | 5 |

This is a Stag Hunt. There are two stable equilibria, one where everyone always plays Tit for Tat and one where everyone always plays Always Defect. Which one you get depends on initial population proportions. In our example an initial population split equally evolves to universal defection.

Sober and Wilson (1998) direct our attention to group selection. Consider the haystack model of Maynard Smith (1964). In the fall, farmers cut hay and field mice randomly colonize the haystacks. In the haystacks they play the Prisoner's Dilemma and reproduce according to the payoffs. In the spring the haystacks are torn down, the mice scatter, and the cycle is continued. If there are enough generations of mice in the life of a haystack, then it is possible for cooperators to do better on average than defectors. This is because differential reproduction within haystacks creates positive correlation in the population. In haystacks colonized by cooperators and defectors, defectors take over. Then cooperative haystacks outreproduce noncooperative ones.

What is the larger game within which the plays of the Prisoner's Dilemma are embedded? Following Ted Bergstrom (2002), we consider the game played by founding members of a haystack. Their payoff is the number of descendants in the spring, at the end of the life of a haystack. Analysis of this founders game shows that it too is a Stag Hunt. There are two equilibria, one with all cooperators and one with all defectors. Which you get depends on where you start.

Each of these models explains how a population of cooperators can be at a stable equilibrium, but neither explains the origin of cooperation. That is because noncooperation is also a stable state, and the transition from noncooperation to cooperation is left a mystery.

We are left with the question of how it is possible to evolve from the noncooperative equilibrium to the cooperative one in interactions with the structure of the Stag Hunt. Axelrod and Hamilton (1981) raise this question and take kin selection and cooperation in family groups as an origin of cooperative behavior. Beyond this there are a few other good answers available. I will focus on three of them.

The first, due to Arthur Robson (1990), is the use of a signal as a *secret handshake*. Consider a population of defectors in the Stag Hunt game. Suppose that a mutant (or innovator) arises who can send a signal, cooperate with those who send the same signal, and defect against those who don't. The new type behaves like a native against the natives and like a cooperator against itself, and so can (slowly) invade. (The signal needn't really be a secret, but it shouldn't be in current use by defectors for other purposes.) Once cooperators are established, it does not matter if the signaling system somehow falls apart, because in Stag Hunt—unlike Prisoner's Dilemma—no one can do better against a population of cooperators.

The second involves a special kind of local interaction with neighbors. Instead of the random encounters of the usual model, individuals interact with their neighbors on a spatial grid (or some other spatial structure).[5] They play a Stag Hunt game with each neighbor, and cultural evolution

[5] There are pioneering papers by Pollock (1989), Nowak and May (1992), and Hegselmann (1996). These are models where the interaction is Prisoner's Dilemma, or in the case of Nowak and May at a bifurcation between Prisoner's Dilemma and Hawk-Dove. Ellison (1993, 2002) discusses local interaction models of the Stag Hunt in which the <Defect, Defect> equilibrium is risk-dominant. In his models it is the noncooperators who can invade and quickly take over. The differences between Ellison's models and ones in which cooperators can invade and take over are discussed in Skyrms 2004.

proceeds by imitation of the most successful (on average) strategy in the neighborhood. For biological evolution, there is an alternative interpretation in which success translates into reproduction in the neighborhood. We thus have both neighborhood interaction and neighborhood imitation.

Eshel, Sansone, and Shaked (1999) point out that these two neighborhoods need not be the same. For a biological example, take a plant that interacts locally but disperses its seeds widely. On the cultural side we can consider cases where the flow of information allows an individual to observe success beyond the confines of immediate interactions. This is the crucial modification to local interaction models that allows robust evolution of cooperation: the imitation neighborhood must be sufficiently larger than the interaction neighborhood. Then a small clump of contiguous cooperators will grow and eventually take over the population. That is because defectors can see the success of the internal cooperators who interact only with cooperators, and imitate them.

Kevin Zollman (2005) shows that the secret handshake and local interaction work especially well together. The signal used as a secret handshake now needs only to be a *local* secret to work. It could be used elsewhere in the population to mean all sorts of other things. This makes the hypothesis of the existence of an unused signal much more plausible. The local secret handshake can then facilitate the initial formation of a large-enough clump of contiguous cooperators to allow cooperation to spread by imitation.

The third answer involves dynamic networks. Instead of constraining individuals to interact with neighbors on a fixed structure, we can allow the interaction structure to evolve as a result of individuals' choices (Skyrms and Pemantle 2000; Bonacich and Liggett 2003; Pemantle and Skyrms 2004a, 2004b; Liggett and Rolles 2004; Santos, Pacheco, and Lenaerts 2006; Pacheco, Traulsen, and Nowak 2006; Skyrms 2007; Skyrms and Pemantle forthcoming). Cooperators want to interact with each other. In Stag Hunt—unlike Prisoner's Dilemma—noncooperators do not much care. Cooperators and defectors may not wear their strategies on their lapels, but even if it is not so easy to tell cooperators from noncooperators, it is not hard to learn. Robin Pemantle and I show how even naive reinforcement learners will form cooperative associations in Stag Hunt provided the social-network structure is sufficiently fluid (Skyrms and Pemantle 2000; Pemantle and Skyrms 2004a, 2004b). The conclusion is robust to various variations in the learning dynamics (Skyrms 2004, 2007). These cooperative associations could then be the focal points for imitation to spread cooperation, as in the foregoing discussion.[6]

Theory is borne out in laboratory studies of human interactions. There is experimental evidence that there are many different types of individuals (Page, Putterman, and Unel 2005; Burlando and Guala 2005; Fischbacher and Gächter 2006) and that given the opportunity and the requisite information, cooperators will learn to associate with one another to their own benefit.[7]

It is evident that the three preceding accounts of transition from the noncooperative equilibrium of the Stag Hunt to the cooperative one also rely on the establishment of correlated interactions. They share two distinctive features that are missing in many accounts of cooperation:

1. Sufficient positive correlation can be established by a few cooperators in a large population of noncooperators.
2. Once the cooperative equilibrium is reached, it can be maintained even if the correlation fades away.

### Negative Correlation and Spite

If social structure can create positive correlation of encounters, it can also create negative correlation. Negative correlation can overturn the conclusions of conventional game theory just as radically as positive correlation. Consider the effect of perfect negative correlation—of always meeting the other type—on our three games. We are now comparing the other diagonal:

PRISONER'S DILEMMA

| | *Cooperate* | *Defect* |
|---|---|---|
| Cooperate | 3 | **1** |
| Defect | **4** | 2 |

STAG HUNT

| | *Cooperate* | *Defect* |
|---|---|---|
| Cooperate | 3 | **1** |
| Defect | **2** | 2 |

[6] For other models of correlation induced by partner choice, see Wright 1921, where the literature begins; Hamilton 1971; Feldman and Thomas 1987; Kitcher 1993; Oechssler 1997; Dieckmann 1999; and Ely 2002.

[7] For an experiment in a public goods–provision game with voluntary association, see Page, Putterman, and Unel 2005.

PRISONER'S DELIGHT

| | *Cooperate* | *Defect* |
|---|---|---|
| Cooperate | 3 | **1** |
| Defect | **2** | 0 |

Not only is defection restored in the Prisoner's Dilemma and favored in the Stag Hunt, but it is also imposed on Prisoner's Delight. In this last case an individual hurts himself by defecting, but hurts his partner more. This is a case of spiteful behavior. Hamilton and Price also showed how negative correlation is the key to the evolution of spite.

Both spite and altruism on their face appear to violate the rational-choice paradigm. Both have an evolutionary explanation in terms of correlated interactions. Yet spite rarely receives the attention given to altruism. A search on Google Scholar for "Evolution of Altruism" gets 1,570 hits, while one for "Evolution of Spite" gets 32. One cannot help asking whether this is due to a Pollyanna bias. It's enjoyable to write about the sunny side of human nature. But the world is full of spiteful behavior: feuds, vendettas, senseless wars. It is as important to study it as to study altruism.

The way to study spite is to study endogenous correlation mechanisms. In some kinds of repeated interactions, the shadow of the future may sustain spite. Johnstone and Bshary (2004) recently analyzed the persistence of spite in a repeated-game setting. In repeated contests, a reputation for fighting too hard—to one's own detriment as well as the greater detriment of the opponent—may enable one to win future contests more easily. The application need not be restricted to animal contests.

It might be remarked that this is not really spite in the larger game but only self-interest, just as Hobbes claimed that cooperative behavior in a repeated Prisoner's Dilemma may just be self-interest in the longer view. It is useful to be able to look at the phenomenon from both perspectives.

Successful invasion of a spiteful type into a nonspiteful population can be sustained by local interaction. A strain of E. coli bacteria produces, at some reproductive cost to itself, a poison that kills other strains of E. coli—but to which it is impervious. It cannot invade a large random-mixing population because a few poisoners can't do that much damage to the natives and the natives outreproduce the poisoners. But in a spatial, local interaction setting, a clump of poisoners can take over. These phenomena have been observed in the laboratory, with the random encounters taking place in a well-stirred beaker and local interactions taking place in a petri dish. Theoretical analysis by Durrett and Levin (1994) and Iwasa, Nakamaru, and Levin (1998) is a complement to the local interaction model of evolution of cooperation. If we put this case in the framework of Eshel, Shaked, and Sansone of the last section, we find that a large interaction neighborhood and a small imitation neighborhood robustly favor the evolution of spite (see Skyrms 2004).

Group-selection models are not without their spiteful aspects. Suppose that the farmer never tears down the haystacks—the population islands that they represent remain isolated. Then for a mouse, its haystack becomes its world, and this makes all the difference (see Gardner and West 2004). Its haystack's population becomes a population unto itself, within which evolution takes place. The carrying capacity within a haystack is limited, so we are dealing with a small, finite population. This, in itself, induces negative correlation even if pairs of individuals form at random within the haystack, because an individual does not interact with herself. This effect is negligible in a large population but can be significant in a small population. (For a transparent example, consider a population consisting of four individuals, two C and two D. Population frequencies are 50-50, but each type has probability 2/3 of meeting the other type and 1/3 of meeting its own.)

If a defector is introduced into a haystack full of cooperators—a mutant or a migrant—he can cause problems. If the interaction is Prisoner's Dilemma, defectors will, of course, take over. But in small populations, with some versions of Stag Hunt and even Prisoner's Delight, defectors can still invade as a result of negative correlation.

For any positive value of *e,* the following is a version of Prisoner's Delight—an individual prefers to cooperate no matter what the other does:

PRISONER'S MILD DELIGHT

| | *Cooperate* | *Defect* |
|---|---|---|
| Cooperate | 2 + e | e |
| Defect | 2 | 0 |

For any finite population, there is an *e*—yielding some version of Prisoner's Mild Delight—such that a spiteful defector can invade.[8]

[8] For example, suppose 1 defector is introduced into a population of N cooperators. Individuals pair at random. Since the defector cannot interact with himself, he always pairs with a cooperator for a payoff of 2. Cooperators pair with the defector with probability (1/N) and with other cooperators with probability (N-1)/N, for an average payoff of [(N-1)/N]*2 + e. So if e < (2/N), a spiteful mutant does better than the native cooperators.

These three examples serve to indicate that the evolution of spite is an aspect of the evolution of the social contract that is worthy of more detailed study. There is no reason to believe that they exhaust the mechanisms for negative correlation that may be important in social interactions.

To stop here would be to represent the social contract as a neat and simple package of problems. But the social contract is not neat and simple.

## Bargaining

Prisoner's Delight, Stag Hunt, and Prisoner's Dilemma are not the only games that raise issues central to the social contract. We could separate the issues of cooperating to produce a public good and deciding how that good is to be divided. This is the philosopher's problem of distributive justice, and it brings bargaining games to center stage (Braithwaite 1955; Rawls 1957; Sugden 1986; Binmore 1994, 1998, 2005; Gauthier 1985, 1986).

Consider the simplest Nash Bargaining game. Two players have bottom-line demands for a share of a common good. If the demands exceed the available good, agreement is impossible and players get nothing. Otherwise, they get what they demand. We simplify radically by assuming that there are only three possible demands: one-third, one-half, and two-thirds. Evolutionary dynamics in a large random-encounter environment, with and without persistent random shocks, is well studied.

Allowing differential reproduction to carry a population to equilibrium, there are two possibilities. The population may reach an egalitarian consensus, where all demand one-half. Or it may reach a polymorphic state, where half the population demands two-thirds and half the population demands one-third (Sugden 1986). Greedy players get their two-thirds half the time; modest players get their one-third all the time. This inefficient polymorphism wastes resources, but it is evolutionarily stable and has a significant basin of attraction. Persistent shocks can allow a population to escape from this polymorphic trap and favor the egalitarian norm in the very long run, but the inefficient polymorphism remains a possibility for medium-run behavior.[9]

However, the effect of correlation mechanisms is rarely discussed in connection with Nash Bargaining. If correlation plays an important role in producing a surplus to be divided, might it not also play an important role in deciding how the division takes place? Positive correlation of demand types obviously favors the egalitarian solution. Those who ask for

[9] Most of what we know about this is due to Peyton Young. See Young 1993a, 1993b, 1998; and Binmore, Samuelson, and Young 2003.

equal shares do best when they meet each other. Negative correlation is more complicated. Greedy players who demand two-thirds do very well if paired with modest players who ask for one-third, but not well if paired with those who ask for half. If negative correlation initially allows greedy players to outreproduce all others, it cannot be maintained because they run out of modest players. But a greedy-modest polymorphism is a real possibility if the negative correlation is of the kind that sufficiently disadvantages the egalitarians. Possibilities are multiplied if we allow more demand types. It should be of interest to look at specific correlation mechanisms.

If we allow individuals to bargain with neighbors on a spatial grid, islands of egalitarianism spontaneously form. This generates positive correlation, and if individuals emulate their most prosperous neighbors, egalitarianism takes over the population. This is a very rapid process, as other types who interact along the edges of the egalitarian islands are quickly converted to asking for half (J. Alexander and Skyrms 1999; J. Alexander 2007; Skyrms 2004). Equal sharing is contagious.

If we add costless signaling to a large-population random-encounter evolutionary model of bargaining, complicated correlations arise and then fade away. Cooperators establish a positive correlation with cooperators. Greedy types establish a negative correlation with themselves. Although these correlations are transient, their effect is that the basin of attraction of the egalitarian equilibrium is greatly enlarged (see Skyrms 2004).

Axtell, Epstein, and Young (2006) investigate a related model where individuals have one of two "tags" and can condition their action on the tag of their partner in a bargaining game. But there is a different dynamics. Instead of evolution by replication or imitation, they consider a rational-choice model. Things may fall out in various ways—here is one. When interacting with those having the same tag, individuals share alike. But in interactions between tags, one tag type becomes greedy and always demands two-thirds and the other becomes modest and always demands one-third. In this equilibrium tags are used to set up both positive and negative correlations between behaviors. Both correlations are perfect: demand-half behaviors always meet themselves, and the other two demand behaviors always meet each other. The result is egalitarianism within tag types and inegalitarian distribution between types. Axtell, Epstein, and Young see this as a spontaneous emergence of social classes.

We can also see the spontaneous emergence of social classes in a dynamic social network (Skyrms 2004). The classes are stabilized by rational choice but destabilized by imitation. Depending on the details and timing of the dynamics, the social network may end up egalitarian or with a class structure.

## Division of Labor

So far, negative correlation has played a rather sinister role in this story. It is not always so. In cooperating to produce a common good, organisms sometimes discover the efficiency of division of labor, and find a way to implement it. Modern human societies are marvels in their implementation of division of labor; so are the societies of cells in any multicellular organism. On the most elementary level, we can suppose that there are two kinds of specialists that an individual can become, A and B, and that these specialists are complementary. On the other hand, an individual might not specialize at all but rather, less efficiently, do what both specialists do. This gives us a little division-of-labor game:[10]

division of labor

| | *Specialize A* | *Specialize B* | *Go It Alone* |
|---|---|---|---|
| Specialize A | 0 | 2 | 0 |
| Specialize B | 2 | 0 | 0 |
| Go It Alone | 1 | 1 | 1 |

In a random-encounter setting, specialists do badly. Positive correlation makes it worse. What is required to get division of labor off the ground is the right kind of negative correlation. Not all the correlation mechanisms that we have discussed here do the trick.[11] What works best is dynamic social-network formation, where the network structure evolves quickly. Specialists quickly learn to associate with the complementary specialists, and then specialists outperform those who go it alone. The effect of correlation depends on the nature of the interaction.

## Groups Revisited

Individuals sometimes form groups that have a permanence and uniformity of interaction with other groups that qualified them to be thought of as individuals. This happens at various levels of evolution. We ourselves are such groups of cells. And humans participate in various social corpora—teams, states, ideological groups—that interact with others.

[10] For analysis of different division-of-labor games, motivated by evolution of coviruses, see Wahl 2002.

[11] One population signaling models face is when an individual interacts with another who sends the same signal. See Skyrms 2004.

How such superindividuals are formed and hold together (or don't) is a central issue of both biology (R. Alexander 1979, 1987; Buss 1987; Maynard Smith and Szathmary 1995; Frank 1998, 2003) and social science. There is no one answer, but answers may involve both elements of cooperation and elements of spite. One important factor is the punishment of individuals who act against the interests of the group. A large experimental literature documents the willingness of many individuals to pay to punish "free riders" in public goods–provision games, and shows that such punishment is able to stabilize high levels of cooperation.[12] This is also an important finding of Ostrom's field studies of the self-organized government of commons (1990). Costly punishment is, from an evolutionary point of view, a form of spite—although it is not called by that name in the literature.

The name may strike the reader as overly harsh in the case of the modest, graduated punishments found in Ostrom's field studies of successful cooperative collective management. But in tightly organized superindividuals, punishment can be draconian. Totalitarian regimes or ideologies classify those who violate social norms as traitors or heretics. They may be stoned to death. They have been burned at the stake. The righteous people carrying out such acts no doubt believe that they are engaged in "altruistic punishment." We need also to think about the dark side of punishment.

When groups that can operate more or less like superindividuals have been formed, the interactions of the superindividuals themselves are also liable to the effects of positive and negative correlation described above. They can cooperate to produce a common good, or not. Their interactions may exemplify spite—not only in behavior, like the bacteria, but also in the full psychological sense of the word.

Repeated interactions, alliances, local interaction on a geographical landscape, signals, tags, and network formation all play a role. Division of labor is facilitated by trade networks, and trade may promote both the good and the bad sides of negative correlation.

The negative correlation conducive to spite in small populations may take on a larger significance when we consider interactions between groups. A local population of six interacting nations is perhaps more plausible than a local population of six interacting mice.

[12] For instance, see Ostrom, Walker, and Gardner 1992; and Fehr and Gachter 2000, 2002. Costly punishment is already implicit in the behavior of receivers in ultimatum-game experiments from Güth, Schmittberger, and Schwartze (1982) to Henrich et al. (2004).

## Evolution and the Social Contract

An evolutionary theory of the social contract stands in some contrast with social-contract theory as practiced in the contemporary philosophical treatment of John Harsanyi and John Rawls. They assume that everyone is, in some sense, rational. And they assume that in the relevant choice situation—behind a veil of ignorance—the relevant choosers are all basically the same. They all have the same rational-choice rule,[13] they all have the same basic values, and therefore they all make the same choice.[14] Correlation of types plays no part because it is assumed that there is only one relevant type.

Evolutionary game theory brings different types of individuals into the picture from the beginning. Evolutionary game theory is full of contingency. There are typically many equilibria; there are many possible alternative social contracts. The population might never get to equilibrium but rather cycle or describe a chaotic orbit. Mutation, invention, experimentation, and external environmental shocks add another layer of contingency.

Evolutionary game theory has some affinity with rational-choice theory in the absence of correlation.[15] This vanishes when interactions are correlated. But correlation, positive and negative, is the heart of the social contract. Correlation gets it started. Correlation lets it grow and develop more complex forms. Social institutions and networks evolve to enable and maintain correlation. Correlation explains much of what is admirable and what is despicable in existing social contracts—what we would like to keep and what we would like to change. A better understanding of the dynamics of correlation should be a central concern for Darwinian social philosophy.

[13] But theorists disagree about the nature of rational choice. Rawls minimizes the maximum loss; Harsanyi maximizes the expected payoff.

[14] The theorist tells you what that choice will be.

[15] In large populations, expected fitness can then be calculated using population proportions in place of the subjective probabilities of rational choice theory.

## Notes

I would like to thank my discussants at the Tanner symposium, Eleanor Ostrom, Michael Smith, and Peyton Young, for valuable comments on the lecture. I would also like to thank Jeffrey Barrett, Louis Narens, Don Saari, Rory Smead, Elliott Wagner, Jim Woodward, and Kevin Zollman for comments on earlier drafts, which much improved the lecture.

### references

Alexander, J. M. 2000. "Evolutionary Explanations of Distributive Justice." *Philosophy of Science* 67: 490–516.

———. 2007. *The Structural Evolution of Morality.* Cambridge: Cambridge University Press.

Alexander, J. M., and B. Skyrms. 1999. "Bargaining with Neighbors: Is Justice Contagious?" *Journal of Philosophy* 96: 588–98.

Alexander, R. D. 1979. *Darwinism and Human Affairs.* Seattle: University of Washington Press.

———. 1987. *The Biology of Moral Systems.* New York: de Gruyter.

Axelrod, R. 1981. "The Emergence of Cooperation among Egoists." *American Political Science Review* 75: 306–18.

———. 1984. *The Evolution of Cooperation.* New York: Basic Books.

Axelrod, R., and W. D. Hamilton. 1981. "The Evolution of Cooperation." *Science* 211: 1390–96.

Axtell, R., J. M. Epstein, and H. P. Young. 2006. "The Emergence of Classes in a Multi-agent Bargaining Model." In *Generative Social Science: Studies in Agent-Based Computational Modeling,* 177–95. Princeton: Princeton University Press.

Bergstrom, T. 2002. "Evolution of Social Behavior: Individual and Group Selection Models." *Journal of Economic Perspectives* 16: 231–38.

Bergstrom, T., and O. Stark. 1993. "How Altruism Can Prevail in an Evolutionary Environment." *American Economic Review* 83: 149–55.

Binmore, K. 1994. *Game Theory and the Social Contract I: Playing Fair.* Cambridge: MIT Press.

———. 1998. *Game Theory and the Social Contract II: Just Playing.* Cambridge: MIT Press.

———. 2005. *Natural Justice.* Oxford: Oxford University Press.

Binmore, K., L. Samuelson, and H. P. Young. 2003. "Equilibrium Selection in Bargaining Models." *Games and Economic Behavior* 45: 296–328.

Björnerstedt, J., and J. W. Weibull. 1996. "Nash Equilibrium and Evolution by Imitation." In *The Rational Foundations of Economic Behavior,* edited by K. J. Arrow et al. New York: St. Martin's Press.

Bonacich, P., and T. Liggett. 2003. "Asymptotics of a Matrix-Valued Markov Chain Arising from Sociology." *Stochastic Processes and Their Applications* 104: 155–71.

Braithwaite, R. B. 1955. *The Theory of Games as a Tool for the Moral Philosopher.* Cambridge: Cambridge University Press.

Burlando, R. M., and F. Guala. 2005. "Heterogeneous Agents in Public Goods Experiments." *Experimental Economics* 8: 35–54.

Buss, L. W. 1987. *The Evolution of Individuality.* Princeton: Princeton University Press.

Dewey, J. 1910. *The Influence of Darwin on Philosophy, and Other Essays in Contemporary Thought.* New York: Henry Holt.

Dieckmann, T. 1999. "The Evolution of Conventions with Mobile Players." *Journal of Economic Behavior and Organization* 38: 93–111.

Durrett, R., and S. Levin. 1994. "The Importance of Being Discrete (and Spatial)." *Theoretical Population Biology* 46: 363–94.

Ellison, G. 1993. "Learning, Local Interaction, and Coordination." *Econometrica* 61: 1047–71.

———. 2000. "Basins of Attraction, Long-Run Stochastic Stability, and the Speed of Step-by-Step Evolution." *Review of Economic Studies* 67: 17–45.

Ely, J. 2002. "Local Conventions." *Advances in Theoretical Economics* 2, no. 1. http://www.bepress.com/.

Epstein, J. M. 2006. *Generative Social Science: Studies in Agent-Based Computational Modeling.* Princeton: Princeton University Press.

Eshel, I., and L. L. Cavalli-Sforza. 1982. "Assortment of Encounters and the Evolution of Cooperativeness." *Proceedings of the National Academy of Sciences of the USA* 79: 331–35.

Eshel, I., E. Sansone, and A. Shaked. 1999. "The Emergence of Kinship Behavior in Structured Populations of Unrelated Individuals." *International Journal of Game Theory* 28: 447–63.

Fehr, E., and S. Gachter. 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review* 90: 980–94.

———. 2002. "Altruistic Punishment in Humans." *Nature* 415: 137–40.

Feldman, M., and E. Thomas. 1987. "Behavior-Dependent Contexts for Repeated Plays in the Prisoner's Dilemma II: Dynamical Aspects of the Evolution of Cooperation." *Journal of Theoretical Biology* 128: 297–315.

Fischbacher, U., and S. Gächter. 2006. "Heterogeneous Social Preferences and the Dynamics of Free-Riding in Public Goods." Working paper, University of Zurich.

Frank, S. A. 1995. "George Price's Contributions to Evolutionary Genetics." *Journal of Theoretical Biology* 175: 373–88.

———. 1998. *Foundations of Social Evolution.* Princeton: Princeton University Press.

———. 2003. "Perspective: Repression of Competition and the Evolution of Cooperation." *Evolution* 57: 693–705.

Fudenberg, D., and D. Levine. 1998. *A Theory of Learning in Games.* Cambridge: MIT Press.

Gardner, A., and S. A. West. 2004. "Spite and the Scale of Competition." *Journal of Evolutionary Biology* 17: 1195–1203.

Gauthier, D. 1985. "Bargaining and Justice." *Social Philosophy and Policy* 2: 29–47.

———. 1986. *Morals by Agreement.* Oxford: Oxford University Press.

Gibbard, A. 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgement.* Cambridge: Harvard University Press.

Grafen, A. 1984. "Natural Selection, Kin Selection, and Group Selection." In *Behavioral Ecology: An Evolutionary Approach,* edited by J. R. Krebs and N. B. Davies, 62–84. Sunderland, Mass.: Sinauer.

———. 1985. "A Geometric View of Relatedness." In *Oxford Surveys in Evolutionary Biology,* edited by R. Dawkins and M. Ridley, 2:28–89. Oxford: Oxford University Press.

Greif, A. 1989. "Reputations and Coalitions in Medieval Trade." *Journal of Economic History* 49: 857–82.

———. 2006. *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade.* Cambridge: Cambridge University Press.

Güth, W., R. Schmittberger, and B. Schwartze. 1982. "An Experimental Analysis of Ultimatum Bargaining." *Journal of Economic Behavior and Organization* 3: 367–88.

Hamilton, W. D. 1963. "The Evolution of Altruistic Behavior." *American Naturalist* 97: 354–56.

———. 1964. "The Genetical Evolution of Social Behavior I and II." *Journal of Theoretical Biology* 7: 1–52.

———. 1971. "Selection of Selfish and Altruistic Behavior in Some Extreme Models." In *Man and Beast,* edited by J. F. Eisenberg and W. S. Dillon, 59–91. Washington, D.C.: Smithsonian Institution Press.

———. 1995. *Narrow Roads of Gene Land.* Vol. 1, *Evolution of Social Behavior.* New York: W. H. Freeman.

Hampton, J. 1996. *Hobbes and the Social Contract Tradition.* Cambridge: Cambridge University Press.

Harms, W. 2001. "Cooperative Boundary Populations: The Evolution of Cooperation on Mortality Risk Gradients." *Journal of Theoretical Biology* 213: 299–313.

———. 2004. *Information and Meaning in Evolutionary Processes.* New York: Cambridge University Press.

Harms, W., and B. Skyrms. 2007. "Evolution of Moral Norms." In *Oxford Handbook in the Philosophy of Biology,* edited by Michael Ruse. Oxford: Oxford University Press.

Harsanyi, J. 2007. *Essays on Ethics, Social Behaviour, and Scientific Explanation.* Dordrecht: Reidel.

Hegselmann, R. 1996. "Social Dilemmas in Lineland and Flatland." In *Frontiers of Social Dilemmas Research,* edited by W. B. G. Liebrand and D. Messick, 337–62. Berlin: Springer Verlag.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, and H. Gintis. 2004. *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies.* New York: Oxford University Press.

Hofbauer, J., and K. Sigmund. 1998. *Evolutionary Games and Population Dynamics.* Cambridge: Cambridge University Press.

Iwasa, Y., M. Nakamaru, and S. A. Levin. 1998. "Allelopathy of Bacteria in a Lattice Population: Competition between Colicin-Sensitive and Colicin-Producing Strains." *Evolutionary Ecology* 12: 785–802.

Johnstone, R. A., and R. Bshary. 2004. "Evolution of Spite through Indirect Reciprocity." *Proceedings of the Royal Society of London B* 271: 1917–22.

Kandori, M. 1992. "Social Norms and Community Enforcement." *Review of Economic Studies* 59: 63–80.

Kavka, G. 1986. *Hobbesian Moral and Political Theory.* Princeton: Princeton University Press.

Kitcher, P. 1993. "The Evolution of Human Altruism." *Journal of Philosophy* 10: 497–516.

Liggett, T. M., and S. W. W. Rolles. 2004. "An Infinite Stochastic Model of Social Network Formation." *Stochastic Processes and Their Applications* 113: 65–80.

Luce, R. D., and H. Raiffa. 1957. *Games and Decisions.* New York: Wiley.

Maynard Smith, J. 1964. "Group Selection and Kin Selection." *Nature* 201: 1145–47.

———. 1982. *Evolution and the Theory of Games.* Cambridge: Cambridge University Press.

Maynard Smith, J., and E. Szathmary. 1995. *The Major Transitions in Evolution.* Oxford: Oxford University Press.

Milgrom, P., D. North, and B. Weingast. 1990. "The Role of Institutions in the Revival of Trade: The Law Merchant, Private Judges, and the Champagne Fairs." *Economics and Politics* 2: 1–23.

Nowak, M. A., and R. M. May. 1992. "Evolutionary Games and Spatial Chaos." *Nature* 359: 826–29.

Nowak, M. A., and K. Sigmund. 1998. "Evolution of Indirect Reciprocity by Image Scoring." *Nature* 393: 573–77.

Oechssler, J. 1997. "Decentralization and the Coordination Problem." *Journal of Economic Behavior and Organization* 32: 119–35.

Ostrom, E. 1990. *Governing the Commons.* Cambridge: Cambridge University Press.

Ostrom, E., J. Walker, and R. Gardner. 1992. "Covenants with and without a

Sword: Self-Governance Is Possible." *American Political Science Review* 86: 404–17.

Pacheco, J. M., A. Traulsen, and M. A. Nowak. 2006. "Active Linking in Evolutionary Games." *Journal of Theoretical Biology* 243: 437–43.

Page, T., L. Putterman, and B. Unel. 2005. "Voluntary Association in Public Good Experiments: Reciprocity, Mimicry, and Efficiency." *Economic Journal* 115: 1032–53.

Pemantle, R., and B. Skyrms. 2004a. "Network Formation by Reinforcement Learning: The Long and the Medium Run." *Mathematical Social Sciences* 48: 315–27.

———. 2004b. "Time to Absorption in Discounted Reinforcement Models." *Stochastic Processes and Their Applications* 109: 1–12.

Pollock, G. B. 1989. "Evolutionary Stability in a Viscous Lattice." *Social Networks* 11: 175–212.

Price, G. R. 1970. "Selection and Covariance." *Nature* 227: 520–21.

Ratnieks, F., and K. Visscher. 1989. "Worker Policing in the Honeybee." *Nature* 342: 796–97.

Rawls, J. 1957. "Justice as Fairness." *Journal of Philosophy* 54: 653–62.

———. 1971. *A Theory of Justice.* Cambridge: Harvard University Press.

Robson, A. J. 1990. "Efficiency in Evolutionary Games: Darwin, Nash, and the Secret Handshake." *Journal of Theoretical Biology* 144: 379–96.

Samuelson, L. 1997. *Evolutionary Games and Equilibrium Selection.* Cambridge: MIT Press.

Santos, F. C., J. M. Pacheco, and T. Lenaerts. 2006. "Cooperation Prevails When Individuals Adjust Their Social Ties." *PLoS Computational Biology* 2, no. 10: 1–6.

Scanlon, T. 1998. *What We Owe to Each Other.* Cambridge: Harvard University Press.

Schelling, T. 1960. *The Strategy of Conflict.* Cambridge: Harvard University Press.

Schlag, K. H. 1998. "Why Imitate and If So, How? A Boundedly Rational Approach to Multi-armed Bandits." *Journal of Economic Theory* 78: 130–56.

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

———. 2001. "The Stag Hunt." *Proceedings and Addresses of the American Philosophical Association* 75: 31–41.

———. 2004. *The Stag Hunt and the Evolution of Social Structure.* Cambridge: Cambridge University Press.

———. 2007. "Dynamic Networks and the Stag Hunt: Some Robustness Considerations." *Biological Theory* 2, no. 1: 1–3.

Skyrms, B., and R. Pemantle. 2000. "A Dynamic Model of Social Network Formation." *Proceedings of the National Academy of Sciences of the USA* 97: 9340–46.

———. forthcoming. “Learning to Network.” In *Probability in Science,* edited by E. Eells and J. Fetzer. Chicago: Open Court.

Sober, E., and D. S. Wilson. 1998. *Unto Others: The Evolution and Psychology of Unselfish Behavior.* Cambridge: Harvard University Press.

Sugden, R. 1986. *The Economics of Rights, Co-operation, and Welfare.* Oxford: Basil Blackwell.

Trivers, R. 1971. “The Evolution of Reciprocal Altruism.” *Quarterly Review of Biology* 46: 35–57.

Vanderschraaf, P. 1998. “The Informal Game Theory in Hume’s Account of Convention.” *Economics and Philosophy* 14: 215–47.

———. 2006. “War or Peace: A Dynamical Analysis of Anarchy.” *Economics and Philosophy* 22: 243–79.

Vanderschraaf, P., and J. M. Alexander 2005. “Follow the Leader: Local Interaction with Influence Neighborhoods.” *Philosophy of Science* 72: 86–113.

Wahl, L. M. 2002. “Evolving the Division of Labor: Generalists, Specialists, and Task Allocation.” *Journal of Theoretical Biology* 219: 371–88.

Weibull, J. 1995. *Evolutionary Game Theory.* Cambridge: MIT Press.

Wright, S. 1921. “Systems of Mating III: Assortative Mating Based on Somatic Resemblance.” *Genetics* 6: 144–61.

———. 1945. “Tempo and Mode in Evolution: A Critical Review.” *Ecology* 26: 415–19.

Young, H. P. 1993a. “An Evolutionary Model of Bargaining.” *Journal of Economic Theory* 59: 145–68.

———. 1993b. “The Evolution of Conventions.” *Econometrica* 61: 57–84.

———. 1998. *Individual Strategy and Social Structure.* Princeton: Princeton University Press.

Zollman, K. 2005. “Talking to Neighbors: The Evolution of Regional Meaning.” *Philosophy of Science* 72: 69–85.