Journal of Economic Perspectives— Volume 3, Number 4— Fall 1989— Pages 85–97

Spontaneous Order

Robert Sugden

In a fishing village on the Yorkshire coast there used to be an unwritten rule about the gathering of driftwood after a storm. Whoever was first onto a stretch of the shore after high tide was allowed to take whatever he wished, without interference from later arrivals, and to gather it into piles above the high-tide line. Provided he placed two stones on the top of each pile, the wood was regarded as his property, for him to carry away when he chose. If, however, a pile had not been removed after two more high tides, this ownership right lapsed (Walmsley, 1932, pp. 70–71). The writer who describes this "first-on" rule does not tell us how it came into existence. Probably its origins had been long forgotten. Nor does he tell us why people obeyed it: only that they did. But we can be sure that the inhabitants of a fishing village would not have appealed to law courts or police to enforce a custom about driftwood. Somehow this rule was self-enforcing. The first-on rule is an example of what Friedrich Hayek (1960, 1979) calls "spontaneous order."

An economist would notice the efficiency properties of a rule that quickly establishes ownership rights over unowned but valuable objects. The principle that the first person on the shore is given a free hand to collect what he wants allows the gathering of driftwood to be carried out in a way that is economical of labor as compared with a system in which everyone scrambles to get to the best wood first. And the principle of the marked piles does away with the need to guard the wood before it is taken home. But if a rule has never consciously been chosen, can its efficiency be an explanation of its existence? To think so would be to ignore the lesson of the Prisoner's Dilemma: even if everyone is better off when everyone follows the rule than when no one follows it, following that rule is not necessarily rational for any individual.

■ *Robert Sugden is Professor of Economics, School of Economic and Social Studies, University of East Anglia, Norwich, England.*

So how are we to explain the existence of this rule? And why this particular rule for assigning property rights, rather than any other? The first-on rule is only one of many rules, or as I shall say, conventions, that could fix property rights in driftwood. Why not allocate each person a day for collecting wood? Or why not use a lottery? Either of these alternatives might have been more efficient than the first-on rule. By using a kind of race (the race to be first on the shore) to assign property rights, the first-on rule encourages everyone to expend effort in an attempt to win; this competitive expenditure of effort in the race is a deadweight loss. Are some sorts of convention —even if inefficient—more likely to emerge or more able to survive than others?

My concern is to try to explain how rules regulating human action can evolve without conscious human design, and can maintain themselves without there being any formal machinery for enforcing them. I want to be able to say something about the kinds of rules that are likely to evolve and survive. And I want to find how these rules link with rationality and with morality.

Why, it might be asked, should economists concern themselves with such questions? One answer is that the market itself is in important respects a spontaneous order. Many of the institutions of a market economy are conventions that no one has designed, but that have simply evolved. Think of the arrangements by which buyers and sellers make contact. How does it come to be common knowledge that advertisements for a particular kind of job will be placed in one newspaper rather than another, and that it is the task of the buyer of labor, not the seller, to advertise? What makes one part of a city the business district? What makes some stock markets more important than others? Although markets may work more smoothly when property rights are defined by formal laws and enforced by the state, they can come into existence and persist without any such external support. Think of how markets in foreign currency, gambling, prostitution, alcohol and narcotics can continue despite the attempts of governments to suppress them. Such markets can continue only because the participants recognise *de facto* property rights that the state does not. This raises the possibility that the institution of property itself may ultimately be a form of spontaneous order.

An analysis of spontaneous order may also throw light on some theoretical problems about rationality. Game theorists typically assume that their games are played by ideally rational individuals who have full knowledge of each other's preferences and attitudes to risk. Many game theorists have regarded it as self-evident that, in such a situation, there must be a uniquely rational strategy for each player, which can be identified by deductive reasoning (for example, Harsanyi and Selten, 1988). If this is true, then individuals who follow conventions cannot be fully rational. (The essential feature of a convention is that it is one of several possible solutions to a game.) Conversely, if it can be rational to follow a convention, the claim that every game has a uniquely rational solution must be false.

Finally, the idea of spontaneous order may have implications for how we approach normative questions. Economists tend to think of moral judgments as judgments about the overall welfare of society, made from some neutral standpoint. They are attracted to moral theories like classical utilitarianism, or John Rawls's (1972) theory of justice, which allow social welfare functions to be constructed

according to simple sets of rational principles. Hayek (1960, 1979) offers an entirely different perspective. For Hayek, the idea that we can, as it were, stand outside our society and rationally appraise its institutions is a dangerous illusion. The institutions and the moral beliefs of a free society, he argues, are the unplanned consequences of a process of evolution. The conventions which create order in a free society are supported by moral beliefs: people believe that they ought to keep to these conventions. But there is no independent principle of justice that provides a rational basis for these beliefs. The belief that one ought to follow a convention is the product of the same process of evolution as the convention itself. Thus the study of spontaneous order may help to explain why we have some of the moral beliefs that we do have, without in any way being able to show that we *ought* to have them. It is to this enterprise that this paper, and the book (Sugden, 1986) on which it is based, belongs.

Conventions

The game of Chicken provides a simple model of how rules of property might come into existence. Imagine two individuals disputing about which of them should take something they both want (perhaps driftwood). Each has a choice between two (pure) strategies, an aggressive one and a conciliatory one. Mixing avian metaphors, I shall use the language of theoretical biology and call these strategies "Hawk" and "Dove." The intuitive idea is that to play Hawk is to hold out for all the good at the risk of deadlock or conflict. To play Dove is to seek compromise but to be ready to back down at the slightest sign of determination by one's opponent. If one player chooses Hawk while the other chooses Dove, the former scores 1 (which may be thought of as the utility of the good) and the latter 0. If both choose Dove, each scores 0.5: we may imagine that they agree to divide the good equally. If both choose Hawk, each scores -1. (Nothing depends on my having chosen this particular number; all that matters is that each player's score in a Hawk-Hawk encounter is negative.) The crucial assumption here is that if you knew your opponent was sure to be aggressive, it would pay you to be conciliatory.

Each player has a range of options in this game. A player might adopt the "pure" strategy of playing Hawk with certainty, or of playing Dove with certainty. Or he might adopt a "mixed" strategy, playing Hawk with some probability p and Dove with probability $1 - p$. The game is said to be in (Nash) equilibrium if the strategy chosen by each player is a best reply to his opponent's strategy.

I wish to argue that the only *stable* equilibria (a term to be defined more completely later) are ones in which the two players behave differently. But this can make sense only if the players are distinguished in *some* way; if they were literally identical, they would be unable to reason to different conclusions. I shall therefore introduce an explicit mechanism for labelling the players. Suppose that each player, before making a move, receives a signal, which may take the value A or B. For each player the prior probability of each value is 0.5. The signals are perfectly and negatively correlated, such that if one player receives the signal A, that player can be sure the opponent has received the signal B, and vice versa. Thus these signals label

the players as "*A*" and "*B*" and this labelling is common knowledge. (For example, if we treat it as a matter of chance who reaches the shore first after a storm, *A* might be the first on the shore and *B* the second.) Notice, however, that the signals provide no information about the structure of the game: they merely provide a point of reference for the players.[1]

This version of Chicken has three Nash equilibria. One is a mixed-strategy equilibrium in which each player plays Hawk with probability 1/3 irrespective of the signal received. The other two are pure-strategy equilibria. In one, the player who is *A* plays Hawk while his opponent plays Dove. In the other, *B* plays Hawk and *A* plays Dove. For the moment, I shall concentrate on these pure-strategy equilibria, which I shall call conventions. (I shall define "convention" formally later, and explain why the mixed-strategy equilibrium is unstable.)

The rule that whichever player has received (say) the *A*-signal should take the disputed good seems arbitrary: although it is self-sustaining—each player will choose to follow this rule provided he expects his opponent to follow it—just the same is true of the opposite rule. Indeed it is arbitrary that the players should use this particular signal to coordinate their behavior: any signal that gave one label to one player and a different label to the other would serve equally well. Any such convention may be understood as a *de facto* rule of property. The rule assigns the good to whichever of the two players has been identified by a particular signal. Each player receives the signal and each knows the convention (and knows that the other knows it, and so on). Given the behavior of the other, each player benefits by following the convention. In this sense the rule is self-enforcing.

Rationality and Experience

Is it rational to follow conventions? There is a tradition in game theory, traceable to John von Neumann and Oskar Morgenstern (1947, especially pp. 146–148), of analyzing games as unrepeated interactions between players who are fully rational, who know each other's utility functions, and whose rationality is common knowledge. On this view, the ultimate objective of game theory is to show that rational analysis uniquely prescribes a particular strategy for each player in a game. It is as if each player sits in a room by himself, knowing nothing about the other player except his utility function and that he is rational, and knowing nothing about how the game may have been played by other people. Each player must decide what to do, applying unlimited powers of rationality to this severely restricted information and to nothing else. The guiding idea—which is assumed rather than proved—is that a fully rational individual, given this information, must be able to reach a determinate conclusion; and that any two rational individuals, reasoning from the same data, must reach the

[1]The idea that the players of a game may be able to make their choice of strategies contingent on some jointly-observed random event corresponds with Robert Aumann's (1987) concept of coordinated equilibrium.

same conclusion (see, for example, Aumann, 1987). If this program could be carried out, we should have a theory of behavior that was based on axioms of rationality and on nothing more. For rational players, conventions would be redundant.

One of the major achievements of Thomas Schelling (who was featured in the Spring 1989 issue of this journal) has been to show that games like Chicken cannot be "solved" in this way. The ideally rational but completely inexperienced players of classical game theory would find they had insufficient data to determine what they should do. In contrast, ordinary people with limited rationality but some degree of experience and imagination might have no difficulty in coordinating their behavior. On this view, the program of classical game theory is a blind alley: it requires us to throw away the information that players need if they are to work out what it is rational for them to do.

The problem is that rational analysis, as understood in classical game theory, has a circular character. Certainly we can be sure that *if there were* a uniquely rational strategy for each player in any game, these strategies would be in Nash equilibrium with one another. (If one player has a uniquely rational strategy, the other, being rational, will be able to work out what it is; thus each must choose a strategy that is a best reply to the one he knows his opponent will play.) But this does *not* imply that if a game has a unique Nash equilibrium, that equilibrium is uniquely rational. To prove this latter proposition, we need the additional assumption that every game has a uniquely rational solution, and this is precisely what is in question. And in any case, games like Chicken have more than one Nash equilibrium.

To see the circular nature of reasoning about rationality, suppose we begin with the hypothesis that "*A* plays Hawk, *B* plays Dove" is uniquely prescribed by rationality. Then whichever player is *A* can deduce that the opponent, being rational, will play Dove. And so, since Hawk is the unique best reply to Dove, "*A* plays Hawk" *is* uniquely prescribed by rational analysis. Similarly, whichever player is *B* can deduce that his opponent will play Hawk, to which the unique best reply is Dove. So the original hypothesis is self-fulfilling: if both players believed it to be true, each would behave in a way that would make it true for the other. But, of course, exactly the same can be said about the hypothesis that rational analysis uniquely prescribes "*A* plays Dove, *B* plays Hawk."[2] There seems to be no way in which rational analysis could discriminate between two such self-fulfilling hypotheses. The implication seems to be that a convention such as "*A* plays Hawk, *B* plays Dove" is *consistent with*, but not *prescribed by*, rationality. When people follow such a convention, they are guided by something more than the axioms of rational choice, as economists normally understand them.

What, then, is this missing ingredient? Schelling (1960, pp. 54–58) finds it in the concept of *prominence* (sometimes called "salience" or the idea of a "focal point"). A famous "coordination game" of Schelling's illustrates the idea. You are paired off with

[2]The same argument cannot be applied to the mixed-strategy equilibrium. The supposition that it is *uniquely* rational for each player to play Hawk with probability 1/3 is self-contradictory: if you know your opponent will play this particular mixed strategy, any reply is as good as any other.

a partner, whom you are trying to meet. He is trying to meet you too. You cannot communicate with one another. You are each told to choose one place in New York City to go in the hope of meeting the other. Where should you go? The problem is to think of the place that your partner would be most likely to choose, given that he is trying to think of the place you would be most likely to choose, and given that he knows you are trying to think.... As Schelling shows, people are often remarkably good at solving this sort of problem. That is, they can converge on the same answers: the only test of a good answer is that it agrees with other people's. (For example: more than half of Schelling's respondents, who were all from New Haven, Connecticut, presumably in the late 1950s, knew to go to Grand Central Station.) Some ways of coordinating behavior seem to strike people as more obvious than others: this is the property of prominence.

If people can coordinate their behavior without communicating with one another, they must be drawing—consciously or unconsciously—on some fund of ideas that they have in common. The most important source of such ideas, I suggest, is common experience. Suppose I am driving down a narrow lane somewhere in England, and meet another car coming the other way. There is just room for two cars to pass. To which side of the road should I steer? I would steer left, expecting the other driver to do the same. But what are the grounds for this expectation? The game, we may assume, is entirely symmetrical; there is no advantage to our both steering left rather than right. If this were a genuinely unrepeated game, and if all I had to go on was the knowledge that my opponent was rational, I would be at a loss to know what to do. Because I have played the game before, I can draw on my experience of English driving, which tells me that drivers almost always steer left in situations like this.

A rational-choice theorist might still ask why this experience is relevant. If my opponent and I are both rational, why isn't our behavior determined by the payoffs of the particular game we are playing? Why should it matter to us what other people have done in the past? The answer, surely, is that there is no uniquely rational solution to our problem. If we are to coordinate our behavior, as we both wish to do, we must rely on some shared notion of prominence. Our common experience of English driving provides the clue we need. Steering left is prominent because it is common knowledge that this is what people generally do: we have each observed this, we can each assume the other has observed it, and so on. What classical game theory is requiring is that rational individuals should ignore as irrelevant the information that comes to them because they are human beings with common experiences—the very information they need to use in order to coordinate their behavior. This is surely far too narrow a view of rationality.

Evolutionary Stability

Conventions, I have argued, cannot be understood if we use the starting point of classical game theory—perfectly rational individuals in unrepeated interactions. It may be more useful to put less stress on rationality and to think of conventions as the

product of evolutionary processes. A fruitful way to begin is to look at how game theory has been used to explain evolutionary processes in biology. Here the central concept is that of *evolutionary stability*, developed by John Maynard Smith and his collaborators (Maynard Smith and Price, 1973; Maynard Smith and Parker, 1976; Maynard Smith, 1982) to explain how animals of a given species behave when they come into conflict with one another. In these theories, interactions between animals are modelled as games, in which the payoffs are measured in terms of biological fitness. If we substitute utility for fitness and learning for natural selection, this approach can be adapted to explain human behavior.

Imagine a large population from which pairs of individuals are repeatedly drawn at random to play a particular two-person game. (I use a two-person game for convenience: the generalization is obvious.) If, as in my version of Chicken, there is some signal which assigns labels to the players, then a strategy must specify what is to be done given every possible value of the signal. For example, one strategy in Chicken is "If A, play Hawk; if B, play Dove." Players do not know, or do not remember, the identities of their opponents; a strategy is successful, then, to the extent that it performs well against opponents in general. There is no assumption that players are rational enough to work out optimal strategies by deductive reasoning; instead it is assumed that through a process of trial and error and imitation, they gravitate towards successful strategies. (This is the human analogue of mutation and natural selection.) An *evolutionarily stable strategy* (or ESS) is a pattern of behavior such that, if it is generally followed in the population, any small number of people who deviate from it will do less well than the others. This, then, is a state of rest in the evolutionary process. I shall define a convention as any ESS in a game that has two or more ESS's. The idea here is that a convention is one of two or more rules of behavior, any one of which, once established, would be self-enforcing.

To test whether a strategy is evolutionarily stable, we must consider not only whether any individual can gain by deviating unilaterally, but also whether any small group of individuals would gain if they happened to deviate in the same way at the same time. The significance of this test can be seen by looking again at Chicken. Recall that this game has three Nash equilibria. These can be described by the following three strategies: (i) If A, play Hawk; if B, play Dove, (ii) If A, play Dove; if B, play Hawk, and (iii) Whether you are A or B, play Hawk with probability $1/3$. Clearly, any strategy that does not correspond with a Nash equilibrium cannot be evolutionarily stable. (If a given strategy is not a Nash equilibrium, then in a situation in which the strategy is generally followed, some individuals can gain by deviating unilaterally.) But not all Nash equilibria are evolutionarily stable.

Each of the strategies (i) and (ii) is the *unique* best reply to itself. (That is, these strategies correspond with "strong" or "strict" Nash equilibria.) This is a sufficient condition for evolutionary stability. To say that a given strategy is the unique best reply to itself is to say that, against opponents who play that strategy, individuals who deviate from the strategy do less well than those who do not deviate. Thus in a situation in which almost everyone follows the strategy (that is, in which the proportion of deviants is vanishingly small), deviants must do less well than non-deviants. But while strategies (i) and (ii) are evolutionarily stable, strategy (iii) is not. To see

why not, suppose almost everyone follows this strategy, but a few individuals simultaneously hit on the idea of playing strategy (i). Against an opponent who plays Hawk with probability 1/3, *every* strategy, pure or mixed, gives the same expected utility (that is, 1/3). So to the extent that his opponents play strategy (iii), an individual loses nothing by playing (i) rather than (iii). But on the rare occasions when (i)-playing deviants meet, they will be able to coordinate with one another. Thus the deviants will do slightly better overall than the rest of the population, and so there will be a tendency for the deviant strategy to be repeated and imitated. But the more frequently the deviant strategy is played, the greater is the incentive to play it. Thus strategy (iii) is not evolutionarily stable: as Maynard Smith and Parker (1976) show, the only ESS's in Chicken are those that exploit asymmetries.

How far does the idea of an ESS differ from equilibrium conditions used in conventional game theory? As I have shown, evolutionary stability is a stronger requirement than (or, in the language of game theory, a "refinement" of) Nash equilibrium. There are some similarities between an ESS and Reinhard Selten's (1975) concept of a *trembling-hand equilibrium*, which is also a refinement of Nash equilibrium. Roughly, Selten supposes that there is some probability, vanishingly small but not zero, that each strategy that is available to any player will be played by mistake: this is the tremble. Then each player's equilibrium strategy must be a best reply to the opponent's, after making allowance for the possibility that the opponent will tremble. Deviant play in the evolutionary approach might be regarded as a kind of tremble, since *in the first instance* it is not governed by any rational calculus, but is essentially random.

However, there is an important difference between the two ways of thinking about nonrational behavior. In Selten's theory, the frequency of each type of tremble is taken as given—as the product of some unexplained psychological mechanism. To test whether a particular state of affairs is an equilibrium, we ask whether rational players are optimizing, given their knowledge of the frequency of trembles. The results of this kind of analysis can depend critically on the assumptions that are made about trembles; but it is not at all clear what constitutes a satisfactory hypothesis here. Since trembles are irrational, they resist explanation in terms of the concepts of classical game theory. This issue is pursued in Binmore (1987). In the evolutionary approach, in contrast, the crucial mechanism is not the responses of rational players to the possibility of nonrational play, but the tendency for deviant play, if successful, to be repeated and imitated. Deviant play, then, is more like experiment than error. Although the initial appearance of any deviant strategy is unexplained (analogously with the role of mutation in biological theories), the extent to which it is then played depends on its degree of success. In this sense, deviant play is explained within the theory.

Which Convention?

In games like Chicken, evolutionary processes will produce conventions that exploit asymmetries between the players. But *which* asymmetries? The evolutionary

approach can also be used to throw light on the question of which conventions are most likely to evolve and survive.

One implication of evolutionary theory is that conventions can be evolutionarily stable even if they are not Pareto-efficient. This can be seen if we introduce some asymmetry into the payoffs of Chicken. Suppose the utility of the good is 1.1 for the *A*-player and 0.9 for the *B*-player; if both players play Dove, *A* gets 0.55 and *B* gets 0.45. Otherwise the game is exactly as before. As in the original version, there are two ESSs: "If *A*, play Hawk; if *B*, play Dove" and "If *A*, play Dove; if *B*, play Hawk." Recall that in any game, each player is equally likely to be labelled as *A* or *B*. Viewed from the start of the game—before the labelling signal is received—the convention under which *A* plays Hawk gives each player an expected utility of 0.55 (i.e. a fifty-fifty chance of 1.1 or zero), while the convention under which *B* plays Hawk gives each an expected utility of 0.45. Thus the second convention is Pareto-inferior to the first. Nevertheless, once established, the inefficient convention is self-perpetuating: no individual or small group can gain by deviating from it.

If conventions were the result of deliberate collective choice, we might expect that inefficient conventions would not be chosen. But conventions are not chosen; they evolve. If we are to explain why one convention is found rather than another, it is not very useful to start from a comparison between a world in which everyone follows one convention and a world in which everyone follows the other: either of these worlds, once achieved, would be self-perpetuating. Instead we must consider the process by which conventions evolve. More particularly, we must look at how they *start* to evolve. Once a convention has started to evolve—once significantly more people are following it than are following any other convention—a self-reinforcing process is in motion. The conventions that establish themselves will be the ones that can take root (biological metaphors are almost unavoidable) most quickly in a convention-free world.

A convention can start to evolve as soon as some people believe that other people are following it. But what gives rise to this initial belief? One possibility is that the same forces are at work as enable people to coordinate their actions without communication in unrepeated games. Some forms of coordination are more prominent than others, and people have a prior expectation of finding the most prominent ones. But, I have argued, prominence is largely a matter of common experience. The implication is that conventions may spread by analogy from one context to another. If it is a matter of common knowledge that a particular convention is followed in one situation, then that convention acquires prominence for other, analogous situations. For example: on my journey to work there is a narrow bridge, not wide enough for two vehicles to pass. If two drivers approach from opposite directions, which of them should give way? Coming on this problem for the first time, my prior expectation was when the drivers came into view of one another, whoever was closer to the bridge would be given the right of way. This expectation—which proved correct—was based on an analogy with the "first come, first served" principle.

If conventions can spread by analogy, then the conventions that are best able to spread are those that are most susceptible to analogy. Thus we should expect to find family relationships among conventions, and not just a chaos of arbitrary and

unrelated rules. One such family relationship, I suggest, is the idea of favoring first possessors and first arrivals. This lies behind both the "first-on" rule in the driftwood example and the "closer driver" rule in the example of the bridge. It also lies behind the "first come, first served" principle of queuing and the "last in, first out" rule for determining which workers should be laid off first in a recession. The same idea is of enormous importance in international affairs. Think of how the positions reached by the American and Soviet armies at the end of the Second World War determined the political map of Europe, each power tacitly respecting the other's claim to the areas its armies had reached first.

Of course, it would be surprising if everyone had exactly the same expectations about which convention would evolve in any given context. Even if everyone is looking for analogies with similar problems, the concepts of similarity and analogy are subjective: different people may draw different analogies. However, in looking for analogies, people are playing another coordination game: each is trying to pick, not the analogy that most appeals to him, but the same analogy as everyone else. Thus we should expect some common principles for drawing analogies to evolve.

Even so, several different conventions might start to evolve simultaneously, each corresponding with a different set of expectations. But over time there will be a tendency for people to gravitate towards whichever convention is most successful; the other conventions (to use another biological metaphor) will die out. So we need to consider what makes for success at this early stage. The most obvious factor is simply popularity: other things equal, the more people follow a convention, the more it pays people to follow it. This takes us back to prominence again.

Another factor is versatility. A convention is versatile if someone who follows it can expect to do reasonably well against opponents following any of the other conventions that might be beginning to evolve at the same time. In the vital early stages of evolution, this may be more important than doing well against opponents like oneself.[3]

These explanations of how conventions start to evolve share a common feature: they do not necessarily favor rules that are Pareto-efficient. An inefficient convention may be more prominent than an efficient one. (For example, rules favoring first arrivals seem to have prominence, even though they can lead to wasteful races.) Because of the tendency for conventions to spread by analogy, we should not necessarily expect them to be well-adapted (in an economic efficiency sense) to the particular problems of coordination that they resolve. Similarly, versatility is a matter of the payoffs from following a convention *before* it becomes firmly established; once the convention is established, its versatility is redundant. Evolution will tend to favor versatile but inefficient conventions relative to ones that are less versatile but more efficient.

[3]Robert Axelrod (1981, 1984) argues that "tit-for-tat" (that is, cooperate in the first round and then repeat your opponent's last move) is a particularly versatile strategy in a repeated Prisoner's Dilemma game, and sees this as one of the main reasons for its success in his famous tournaments.

Conventions and Norms

Rules of behavior, I have been arguing, can evolve spontaneously. Up to now I have meant by a "rule" nothing more than an established pattern of behavior. But now I shall argue that such patterns can become rules in a stronger sense. People can come to believe that they *ought* to act in ways that maintain these patterns: conventions can become norms. This will not be a moral argument. I have nothing to say about what moral beliefs people ought to hold. My concern is to explain the beliefs they *do* hold.

How conventions can become norms was first explained by David Hume (1740, Bk. 3, part 2, sec. 1-3). My analysis of the evolution of conventions is in many ways similar to—and inspired by—Hume's account of the origin of principles of justice. Hume argues that rules of property are conventions that evolve spontaneously; if we are to explain why these rules take the particular forms they do, we must look to "the imagination" rather than to "reason and public interest." Hume's idea of the importance of imagination is remarkably similar to Schelling's concept of prominence. But having argued that principles of justice (and in particular, rules of property) evolve out of the repeated interactions of individuals pursuing their separate interests, Hume goes on to argue that we "annex the idea of virtue to justice."

The mechanism that can transform conventions into norms is the human desire for the approval of others. Although this desire is rarely considered by modern economists, introspection surely tells us that it is at least as fundamental as the desire for most consumption goods. That we desire approval should not be surprising: we are, after all, social animals, biologically fitted to live in groups.

For most of us, being the focus of another person's ill-will, resentment or anger is a source of unease—something we prefer to avoid. This is a psychological externality: one person's *state of mind*, as interpreted by another person, can affect that other person's happiness or utility. This is not to be confused with punishment, which is an *act* by which one person harms another. Most people have the ability to inflict some harm on most others—for example, by physical violence or by theft. But this works both ways. Because you have the ability to harm me, you can choose to punish me when I breach a rule; but you must take account of my ability to retaliate. So if we are roughly equal in the ability to inflict harm on each other, your punishing me is likely to be costly for you as well as for me. If we are to explain acts of punishment, then, we have to explain why those who punish incur these costs. In contrast, your feeling ill-will towards me is not an act of choice on your part; it is merely a psychological state. For your ill-will to cause me unease, it is not even necessary that you should choose to express it: it is sufficient that I can infer your state of mind.[4]

To see the significance of all this, consider any convention that assigns *de facto* property rights in a valuable resource. Suppose the convention is well-established.

[4]Skeptical readers may try the following experiment. Have a meal at an expensive restaurant and leave without giving a tip. Do you feel uneasy as you do so, even if the waiter is perfectly polite?

Then each person has a well-grounded expectation—an expectation grounded in induction from experience—that other people will follow the convention. Given this expectation, each person finds it in his interest to follow the convention. And given that a person is following the convention himself, he not only *expects* the people with whom he interacts to demand no more than the convention allows them, he also *wants* them to behave in this way. Think of the first-on rule for collecting wood. Suppose you are on the shore before me, but I start collecting wood. You had expected me to let you have the pick of the wood. My action is harming you by frustrating an expectation that you had good reason to hold. Further, you have reason to suspect that I know that this is what I am doing: I presumably know the convention as well as you do. My action will surely provoke anger and resentment from you. To explain these reactions, we do not need to appeal to any prior moral beliefs: that I am frustrating your expectations is sufficient explanation. And even if you do not express your feelings, I will be able to deduce what they are from my knowledge that I am frustrating your expectations.

In addition, anyone who is favored by the convention on at least some occasions is likely to regard any breach of the convention as an indirect threat to himself. Thus someone who demands more than the convention allows him will tend to arouse the resentment, not only of those who are directly harmed by this demand, but also of third parties. Suppose someone sees me taking the wood that, according to the first-on rule, is yours. He relies on this rule to protect his claims when he is first on the shore. So the existence of people like me is a threat to him. He will thus be inclined to sympathize with, and in some degree to share, your resentment against me.

Why, it might be asked, does the third party not see my breach of the convention as setting a precedent that he could profitably follow when he is the second person on the shore? Why does he sympathize with you rather than with me? Because the convention is well-established. Occasional breaches of the convention by mavericks like me will not cause it to collapse. Given that almost everyone follows the convention, my action can only harm other people.

This argument does not depend on any assumption that everyone benefits from the existence of the convention, or that the convention increases the overall welfare of society.[5] All that matters for the argument is that breaches of the convention are harmful *to all those people who follow it*. Admittedly, this would not be true of a convention which *always* favored the members of one group of people over another (men over women, for example), since then the disfavored group would have no reason to disapprove of breaches of the convention. But it could be true of a convention whose overall effect was to benefit one group at the expense of another. The point is this: the standpoint from which behavior is judged is a state of affairs in

[5]Much of the argument in this section has followed David Lewis (1969, especially p. 99). Lewis, however, restricts the application of this argument to "coordination problems"—games in which the players' main concern is to coordinate their behavior in some way, and in which they are indifferent (or almost indifferent) between alternative conventions. Thus, unlike me, he stops short of arguing that conventions that favor some people at the expense of others can become norms.

which the convention is generally followed. The convention, we might say, is being judged from inside, not outside.

Conclusion

Order in human affairs, I have argued, can arise spontaneously, in the form of conventions. These are patterns of behavior that are self-perpetuating—that can replicate themselves. In particular, rules of property—the essential preconditions for markets to work—can evolve in this way. These rules are not the result of any process of collective choice. Nor do they result from the kind of abstract rational analysis employed in classical game theory, in which individuals are modelled as having unlimited powers of deductive reasoning but no imagination and no common human experience. In this sense, at least, conventions are not the product of our reason.

Nor are these patterns of behavior necessarily efficient. They have evolved because they are more successful at replicating themselves than other patterns: if they can be said to have any purpose or function, it is simply replication. They do not serve any overarching social purpose; thus they cannot, in general, be justified in terms of any system of morality that sees society as having an overall objective or welfare function. The conventions that we follow may, however, have moral force for us. But if they do, that is because our moral beliefs are the products of the same process of evolution.

References

Aumann, R. J., "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica*, 1987, *55*, 1–18.

Axelrod, R., "The Emergence of Cooperation Among Egoists," *American Political Science Review*, 1981, *75*, 941–973.

Axelrod, R., *The Evolution of Cooperation*. New York: Basic Books, 1984.

Binmore, K., "Modeling Rational Players," *Economics and Philosophy*, 1987, *3*, 179–214.

Harsanyi, J. C., and R. Selten, *A General Theory of Equilibrium Selection in Games*. Cambridge: MIT Press, 1988.

Hayek, F., *The Constitution of Liberty*. London: Routledge and Kegan Paul, 1960.

Hayek, F., *Law, Legislation and Liberty*. London: Routledge and Kegan Paul, 1979. (In three volumes: Vol. 1 published 1973, Vol. 2 published 1976, Vol. 3 published 1979.)

Hume, D. (1740), *A Treatise of Human Nature*. 2nd Edition, Selby-Bigge, L.A., ed. Oxford: Clarendon Press, 1978.

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

Maynard Smith, J., *Evolution and the Theory of Games*. Cambridge: Cambridge University Press, 1982.

Maynard Smith, J., and G. Parker, "The Logic of Asymmetric Contests," *Animal Behavior*, 1976, *24*, 159–175.

Maynard Smith, J., and G. Price, "The Logic of Animal Conflict," *Nature*, 1973, *246*, 15–18.

Neumann, J. von, and O. Morgenstern, *Theory of Games and Economic Behavior*, 2nd Edition. Princeton: Princeton University Press, 1947.

Rawls, J., *A Theory of Justice*. Oxford: Oxford University Press, 1972.

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

Selten, R., "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, 1975, *4*, 25–55.

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

Walmsley, L., *Three Fevers*. London: Collins, 1932.