Journal of Consciousness Studies
www.imprint-academic.com/jcs

STRATEGIC SUBJECTIVE COMMITMENT

Randolph M. Nesse

Game theory has progressed from analysis of one-move games between two rational agents, to iterated n-person games in which strategies evolve, and actors use prior experience to coordinate their moves. The next step in this direction is to analyse commitment strategies. An individual can influence others by announcing his or her commitment to a future act that would not be in his or her best interests. Spiteful threats can coerce others. Promises to aid someone when nothing can be reciprocated can create deep relationships. Such strategies are inherently paradoxical because the maximum payoff comes from not having to follow through on the commitment, and this is made more likely by expensive signalling of commitments to outlandish threats and promises whose plausibility declines with their magnitude. Nonetheless, the fitness benefits of subjective commitment are substantial and may well have shaped human capacities for revenge and spite, as well as deep attachment and genuine morality.

In just a few pages, Skyrms (2000) manages to clarify several thorny issues at the intersection of game theory and evolution. He straightforwardly shows how a theory based on hyper-rational actors parallels one based on non-rational actors whose strategies are shaped by natural solution. This is bolstered by his further demonstration that if A gives an ESS, then <A,A> is a Nash equilibrium for the corresponding two-person game. He then proceeds to contrast the two perspectives, first noting how individuals who have identities that persist through time can escape from the symmetry requirement, using the game of Chicken as an example, and then showing how an evolutionary approach gives a different solution to the Nash bargaining game, namely, demand half.

Things are complicated further by introducing sequential structure to a game. For instance, the ultimatum game generates subgames in which the rational strategy is to take whatever is offered. Real people, of course, do not do this. The question of 'why not?' is one that would occur only to a game theorist. Skyrms shows how weakly-dominated strategies can persist, and how, with correlated strategies, even strongly-dominated strategies can persist. Exploration of such correlated situations is the strength of the Sober and Wilson book (1998, abstracted in this volume).

The logic and direction of Skyrms' argument is compelling, from hyper-rational actors in a single move game, to correlated interactions that repeat over time and are shaped by evolutionary dynamics. The combination is sufficient to explain much of economic and political behaviour. However, just one step further in this direction is another kind of game that may offer powerful explanations of human relationships, a game that badly needs a better foundation in theory. Much actual human behaviour, especially in personal relationships, is based on commitment strategies (Schelling, 1960). These commitments give rise to wonderful complexities that might well be clarified if theorists like Skyrms were to extend the foundations provided by Hirshleifer (1978) and Frank (1988).

The core idea of commitment is that an actor can, by making a commitment to some future action, influence the behaviour of others. If the commitment is to some action that is Bayes rational, this is not very interesting. Likewise, situations in which parties agree on a contract, with enforcement provisions, have been well studied. If, however, an individual announces a commitment to a future action that would not be in his or her interests, then things become curiouser and curiouser. Such commitments are common and powerful. Some are threats. If a person can convince others that he or she will not swerve in the game of chicken, then they will change their behaviour accordingly in a way that yields an advantage to the committed person. Throwing the steering wheel out of the window is rarely possible, so other strategies must be used to convince others that one is ready to die. Acting wildly irrational is just what is needed. Likewise, employees get an advantage if they can convince a supervisor that they will quit unless a raise is granted. A commitment to kill someone unless a ransom is paid is structurally the same. In all these cases, having to follow through on the commitment would yield a net loss, but if a person can convince others that he or she will nonetheless follow through, then there is no need to, and the maximum payoff is obtained. The difficulty is obvious, and creates the central paradox of commitment. How can a person convince others that he or she will, in the future, do something that would be irrational? Various commitment devices have been well described, such as contracts and depriving oneself of negotiating options. My

interest, however, is in subjective commitment. We humans have strong feelings that get us to do things that are otherwise not rational. Trivers (1981) showed the role of evolved emotions in mediating reciprocity, and Robert Frank (1988) has been a clear and eloquent advocate for the selective benefits of emotions in establishing commitments. The desire for spiteful revenge is an excellent example. Falling in love, or, rather, staying in love, is another.

On this more positive side, commitments are promises to help others. Huge efforts have gone into analysing mutual aid relationships in terms of reciprocity and the Prisoner's Dilemma (Axelrod, 1997). While these theories capture some of what goes on, I think they are fundamentally incomplete when applied to personal relationships based on commitment. To see why, look at how careful people are to avoid treating their friends as reciprocity exchange partners. If you offer a friend a ride to the airport, and on the way you say, 'Well, I just want to be clear that you will give me a ride next time I need one. Will you agree to that?', you may get the ride, but lose a friend. And, after your previous friend spreads gossip about what kind of person you really are, you may find yourself trying to understand strange smiles and frosty reserve at future social gatherings. Of course, there is also the other problem of expecting people to fulfill their commitments when there is a basic imbalance in the reciprocity foundations of a relationship. Commitment is not an alternative to reciprocity, it is a way of promising to extend credit beyond the available collateral, that benefits both parties in the long run, on the average.

Consider also the fate of a recent fad in marital therapy. A few years ago some therapists tried to analyse marriages in terms of the exchanges they involved. Spouses rebelled, uniting in their opposition to viewing their behaviours in terms of reciprocity. It seemed inhuman to them. And it was. In our personal attachments, we are extremely careful to see and present our actions as the results of 'love' or other feelings. This is because we want to have partners who will be there for us when we really need them, namely, when we have nothing to offer in exchange. An economic/game theory/naive evolutionary view of human relationships as reciprocal exchanges has spread widely in our culture. This may undermine people's capacity for deep personal relationships, because people who believe that others are capable only of exchange relationships are unable to make emotional commitments. In return, no one loves them either. Their beliefs create a social world that is genuinely cold and empty. This is a realm where the products of social construction strongly influence fitness.

Psychologists have long emphasized the importance of 'basic trust' (Balint, 1979; Erikson, 1980), and the personality pathology that results if it is absent. What individuals believe about human nature has dramatic effects on how they conduct their relationships. On the macro scale, this, in turn, shapes the nature of a society. Psychologists also emphasize 'attachment', usually with reference to the benefits mother and baby get from staying close together (Cassidy and Shaver, 1999). Adult attachments to non-kin may or may not be based on the same brain mechanisms, but I suspect their function is quite different — to make commitment strategies possible. It seems plausible that the emotions that mediate such relationships, including loyalty and grief, may have evolved specifically to facilitate commitments. Relationships based on commitment are a further step in the direction of Skyrms' argument. Incorporating foresight and information transmission to correlated and repeated interactions leads directly to the study of commitments.

I have often emphasized how people are similar to other organisms, but commitment strategies may make us distinctive, if not unique. As soon as cognition gets to the point where people can communicate their intentions, benefits from subjective commitments become available. These benefits may be so substantial that they become a powerful selective force for further increases in intelligence and foresight. The advantages of conscious thought, with its ability to anticipate the outcomes of possible future actions, are amplified in a social setting where people use, and must cope with, commitment strategies. The difficulty, once again, is to convince others that one will follow through on a commitment, or to determine if another will follow through. The best predictors are past behaviour and expensive public expressions of intent that cannot be violated without resulting in major reputational costs. This gives rise to the central paradox of commitment strategies. The optimal strategy is to convince others that one will follow through, so that one does not have to do so. But this usually requires making extreme threats and following through on certain occasions. Thus, the Mafia must burn a certain number of businesses to maintain its protection rackets, and a person who threatens suicide must, at some point, do something very dangerous in order to be taken seriously. Still more ominously, the nuclear strategy of mutual assured destruction rests on this logic.

On the positive side, the courting suitor must give expensive gifts and spend inordinate amounts of time with his intended, otherwise she will conclude that he is in it just for what he can get, and will be unlikely to stay with her decades later when she may be sick or less desirable. Zahavi (1976) has suggested that members of such pairs engage in 'testing of the bond'. By withdrawing or acting disabled, they can test the prospective partner for willingness to help when there is little likelihood of being quickly repaid. If the partners pass such tests, the bond becomes a committed attachment, by which we mean that it is based, *not* on reciprocity or rational calculation, but non-rational emotions. How could such emotions evolve? They certainly can lead to exploitation, so risks abound, but individuals who are capable of making wise and deep commitments have a huge advantage over those who just play reciprocity games. By signalling that they will behave according to commitments instead of rationality, they get major advantages. Especially if they don't have to fulfill their commitments too often. When a very expensive commitment is called in, say when a spouse gets Alzheimer's disease, the psychological wrenching is terrible. That is a subject for another essay. Likewise, there is wonderful complexity yet to be explored in how people form groups, often based on religion, to enforce their mutual commitments.

Finally, the evolutionary benefits of commitment may provide the needed foundation for understanding the origins of the moral passions. Moral behaviour depends, in large measure, on fulfilling one's commitments. Helping one's kin and exchanging favours is individually generous (irrespective of whether it is evolutionarily selfish). But explicitly moral behaviour is, necessarily, outside of the reciprocity framework. Morality requires living up to prior commitments, specifically those that require action that will be to the individual's disadvantage. The fitness benefits of mechanisms to make and keep commitments may have shaped the capacity for morality. All of this will be far better understood when we have a solid game theory foundation for strategies based on commitment.

References

Axelrod, Robert M. (1997), *The Complexity of Cooperation: Agent-based Models of Competition and Collaboration* (Princeton: Princeton University Press).
Balint, Michael (1979), *The Basic Fault* (New York: Brunner/Mazel).
Cassidy, Jude and Shaver, Philip R. (1999), *Handbook of Attachment: Theory, Research, and Clinical Applications* (New York: Guilford Press).
Erikson, Eric H. (1980), *Identity and the Life Cycle* (New York: Norton).
Frank, Robert H. (1988), *Passions Within Reason: The Strategic Role of The Emotions* (New York: W.W. Norton).
Hirshleifer, Jack (1978), 'Competition, cooperation, and conflict in economics and biology', *Journal of the American Economic Association*, **68**, pp. 238–43.
Schelling, Thomas C. (1960), *The Strategy of Conflict* (Cambridge: Harvard University Press).
Skyrms, B. (2000), 'Game theory, rationality and evolution of the social contract', *Journal of Consciousness Studies*, **7** (1–2), pp. 269–84.
Sober, Elliot and Wilson, David S. (1998), *Unto Others: The Evolution and Psychology of Unselfish Behaviour* (Cambridge, MA: Harvard University Press).
Trivers, Robert L. (1981), 'Sociobiology and politics', in *Sociobiology and Human Politics*, ed. E. White (Toronto: Lexington).
Zahavi, Amotz (1976), 'The testing of a bond', *Animal Behaviour*, **25**, pp. 246–7.

DISTRIBUTIVE JUSTICE AND THE NASH BARGAINING SOLUTION

Christopher D. Proulx

Skyrms has pointed out differences between the results of rational choice theory and evolutionary game theory. This commentary argues that there is a great deal of agreement on the Nash Bargaining Solution, which maximizes the product of player payoffs, in both rational-choice-based and evolution-based theories of equilibrium selection. While evolutionary game theory has the potential to explain how we arrive at the behavioural rules that govern what we do, realistic models will require calibration through laboratory experiments. Indeed, experimental evidence strongly supports the Nash Bargaining Solution.

Professor Skyrms' paper (2000) compares rational choice theory and evolutionary game theory. His research has important implications for the notion of distributive justice. Since his paper contains many references to his book *Evolution of the Social Contract* (Skyrms, 1996), inspiration for some of these comments arises there as well as his paper.

I: Experiments and the Evolution of a Theory

Evolutionary game theory became popular in economics partly because of its flexibility in accounting for experimental observations which are apparently inconsistent with rational choice theory. While there is only one way to be rational, there are many ways to be boundedly rational, and this flexibility is a strength and a weakness of evolutionary game theory. The sensitivity of results of evolutionary models to their assumptions means that modeling issues are important. Which issues are appropriate in which contexts is an empirical matter, and evolutionary theory needs empirical observation to guide its development.

That said, we shouldn't throw out the insights gained from standard game theory in favour of some evolutionary theory describing how people decide to do what they do, just as we shouldn't throw out algebra in favour of a theory of developmental psychology which tells us how people come to think about addition. However, we should collect data to determine the bounds of applicability of our theories.