Chapter Five

A Behavioral Approach to the Rational Choice Theory of Collective Action[1]

Elinor Ostrom[2]

Let me start with a provocative statement. You would not be reading this article if it were not for some of our ancestors learning how to undertake collective action to solve social dilemmas. Successive generations have added to the stock of everyday knowledge about how to instil productive norms of behavior in their children and to craft rules to support collective action that produces public goods and avoids 'tragedies of the commons'.[3] What our ancestors and contemporaries have learned about engaging in collective action for mutual defense, child rearing, and survival is not, however, understood or explained by the extant theory of collective action.

Yet, the theory of collective action is *the* central subject of political science. It is the core of the justification for the state. Collective-action problems pervade international relations, face legislators when devising public budgets, permeate public bureaucracies, and are at the core of explanations of voting, interest group formation, and citizen control of governments in a democracy. If political scientists do not have an empirically grounded theory of collective action, then we are hand-waving at our central questions. I am afraid that we do a lot of hand-waving.

The lessons of effective collective action are not simple – as is obvious from human history and the immense tragedies that humans have endured, as well as the successes we have realized. As global relationships become even more intricately intertwined and complex, however, our survival becomes more dependent on empirically grounded scientific understanding. We have not yet developed a *behavioural theory of collective action* based on models of the individual consistent

---

1. Presidential Address, American Political Science Association, 1997 published initially in *American Political Science Review*, vol. 92, 1998, pp. 1–22.

2. The author gratefully acknowledges the support of the National Science Foundation (Grant #SBR-9319835 and SBR-9521918), the Ford Foundation, the Bradley Foundation, and the MacArthur Foundation. My heartiest thanks go to James Alt, Jose Apesteguia, Patrick Brandt, Kathryn Firmin-Sellers, Roy Gardner, Derek Kauneckis, Fabrice Lehoucq, Margaret Levi, Thomas Lyon, Tony Matejczyk, Mike McGinnis, Trudi Miller, John Orbell, Vincent Ostrom, Eric Rasmusen, David Schmidt, Sujai Shivakumar, Vernon Smith, Catherine Tucker, George Varughese, Jimmy Walker, John Williams, Rick Wilson, Toshio Yamagishi, and Xin Zhang for their comments on earlier drafts and to Patty Dalecki for all her excellent editorial and moral support.

3. The term 'tragedy of the commons' refers to the problem that common-pool resources, such as oceans, lakes, forests, irrigation systems, and grazing lands, can easily be overused or destroyed if property rights to these resources are not well defined (*see* Hardin 1968).

with empirical evidence about how individuals make decisions in social-dilemma situations. A behavioural commitment to theory grounded in empirical inquiry is essential if we are to understand such basic questions as why face-to-face communication so consistently enhances cooperation in social dilemmas or how structural variables facilitate or impede effective collective action.

Social dilemmas occur whenever individuals in interdependent situations face choices in which the maximization of short-term self-interest yields outcomes leaving all participants worse off than feasible alternatives. In a public-good dilemma, for example, all those who would benefit from the provision of a public good – such as pollution control, radio broadcasts, or weather forecasting – find it costly to contribute and would prefer others to pay for the good instead. If everyone follows the equilibrium strategy, then the good is not provided or is underprovided.

Yet, everyone would be better off if everyone were to contribute.

Social dilemmas are found in all aspects of life, leading to momentous decisions affecting war and peace as well as the mundane relationships of keeping promises in everyday life. Social dilemmas are called by many names, including the public-good or collective-good problem (Olson 1965; P. Samuelson 1954), shirking (Alchian and Demsetz 1972), the free-rider problem (Edney 1979; Grossman and Hart 1980), moral hazard (Holmstrom 1982), the credible commitment dilemma (Williams, Collins and Lichbach 1997), generalized social exchange (Ekeh 1974; Emerson 1972a, 1972b; Yamagishi and Cook 1993), the tragedy of the commons (G. Hardin 1968) and exchanges of threats and violent confrontations (Boulding 1963). The prisoners' dilemma has become the best-known social dilemma in contemporary scholarship. Among the types of individuals who are posited to face these kinds of situations are politicians (Geddes 1994), international negotiators (Sandler 1992; Snidal 1985), legislators (Shepsle and Weingast 1984), managers (Miller 1992), workers (Leibenstein 1976), long-distance traders (Greif, Milgrom and Weingast 1994), ministers (Bullock and Baden 1977), oligopolists (Comes, Mason and Sandler 1986), labour union organisers (Messick 1973), revolutionaries (Lichbach 1995), homeowners (Boudreaux and Holcombe 1989), even cheerleaders (Hardy and Latané 1988) and, of course, all of us – whenever we consider trusting others to cooperate with us on long-term joint endeavours.

In prehistoric times, simple survival was dependent both on the aggressive pursuit of self-interest and on collective action to achieve cooperation in defence, food acquisition, and child rearing. Reciprocity among close kin was used to solve social dilemmas, leading to a higher survival rate for those individuals who lived in families and used reciprocity within the family (Hamilton 1964). As human beings began to settle in communities and engage in agriculture and long-distance trade, forms of reciprocity with individuals other than close kin were essential to achieve mutual protection, to gain the benefits of long-distance trading, and to build common facilities and preserve common-pool resources.[4] Evolutionary

4. The term 'reciprocal altruism' is used by biologists and evolutionary theorists to refer to strategies of conditional cooperation with non kin that produce higher benefits for the individuals who follow these strategies if they interact primarily with others who are reciprocators (Trivers 1971). Since these strategies benefit the individual using them in the long term, I prefer the term 'reciprocity'.

psychologists have produced substantial evidence that human beings have evolved the capacity – similar to that of learning a language – to learn reciprocity norms and general social rules that enhance returns from collective action (Cosmides and Tooby 1992). At the same time, cognitive scientists have also shown that our genetic inheritance does not give us the capabilities to do unbiased, complex, and full analyses without substantial acquired knowledge and practice as well as reliable feedback from the relevant environment. Trial-and-error methods are used to learn individual skills as well as rules and procedures that increase the joint returns individuals may obtain through specialization, coordination, and exchange. All long-enduring political philosophies have recognized human nature to be complex mixtures of the pursuit of self-interest combined with the capability of acquiring internal norms of behavior and following enforced rules when understood and perceived to be legitimate. Our evolutionary heritage has hardwired us to be boundedly self-seeking at the same time that we are capable of learning heuristics and norms, such as reciprocity, that help achieve successful collective action.

One of the most powerful theories used in contemporary social sciences – rational choice theory – helps us understand humans as self-interested, short-term maximisers. Models of complete rationality have been highly successful in predicting marginal behavior in competitive situations in which selective pressures screen out those who do not maximize external values, such as profits in a competitive market or the probability of electoral success in party competition. Thin models of rational choice have been unsuccessful in explaining or predicting behavior in one-shot or finitely repeated social dilemmas in which the theoretical prediction is that no one will cooperate. In indefinitely (or infinitely) repeated social dilemmas, standard rational choice models predict a multitude of equilibria ranging from the very best to the very worst of available outcomes without any hypothesized process for how individuals might achieve more productive outcomes and avert disasters.[5] Substantial evidence from experiments demonstrates that cooperation levels for most one-shot or finitely repeated social dilemmas far exceed the predicted levels and are systematically affected by variables that play no theoretical role in affecting outcomes. Field research also shows that individuals systematically engage in collective action to provide local public goods or manage common-pool resources without an external authority to offer inducements or impose sanctions. Simply assuming that individuals use long-range thinking 'to achieve the goal of establishing and/or maintaining continued mutual cooperation' (Pruitt and Kimmel 1977: 375) is not a sufficient theory either. It does not explain why some groups fail to obtain joint outcomes easily available to them or why initial cooperation can break down.

5. *See* Farrell and Maskin (1989) for a different approach to this problem.

We now have enough scholarship from multiple disciplines to expand the range of rational choice models we use. For at least five reasons, we need to formulate a behavioural theory of boundedly rational and moral behavior.

First, behavior in social dilemmas is affected by many structural variables, including size of group, heterogeneity of participants, their dependence on the benefits received, their discount rates, the type and predictability of transformation processes involved, the nesting of organisational levels, monitoring techniques, and the information available to participants.[6] In theories that predict either zero or 100 percent cooperation in one-shot or finitely repeated dilemmas, structural variables do not affect levels of cooperation at all. A coherent explanation of the relationship among structural variables and the likelihood of individuals solving social dilemmas depends on developing a behavioural theory of rational choice. This will allow scholars who stress structural explanations of human behavior and those who stress individual choice to find common ground, rather than continue the futile debate over whether structural variables or individual attributes are the most important.

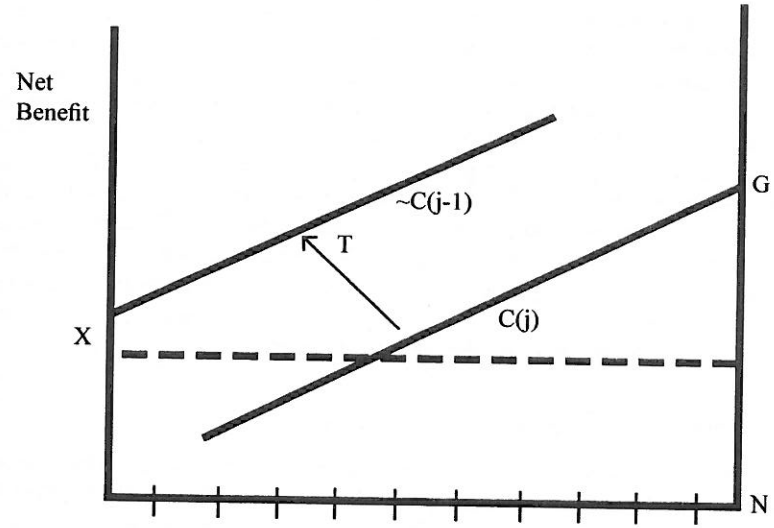
Second, scholars in all the social and some biological sciences have active research programs focusing on how groups of individuals achieve collective action. An empirically supported theoretical framework for the analysis of social dilemmas would integrate and link their efforts. Essential to the development of such a framework is a conception of human behavior that views complete rationality as one member of a family of rationality models rather than the only way to model human behavior. Competitive institutions operate as a scaffolding structure so that individuals who fail to learn how to maximize some external value are no longer in the competitive game (Alchian 1950; Clark 1995; Satz and Ferejohn 1994). If all institutions involved strong competition, then the thin model of rationality used to explain behavior in competitive markets would be more useful. Models of human behavior based on theories consistent with our evolutionary and adaptive heritage need to join the ranks of theoretical tools used in the social and biological sciences.

Third, sufficient work by cognitive scientists, evolutionary theorists, game theorists, and social scientists in all disciplines (Axelrod 1984; Boyd and Richerson 1988, 1992; Cook and Levi 1990; Guth and Kliemt 1995; Sethi and Somanathan 1996; Simon 1985, 1997) on the use of heuristics and norms of behavior, such as reciprocity, has already been undertaken. It is now possible to continue this development toward a firmer behavioural foundation for the study of collective action to overcome social dilemmas.

Fourth, much of our current public policy analysis – particularly since Garrett Hardin's (1968) evocative paper, 'The Tragedy of the Commons' – is based on an

6. This is only a short list of the more important variables found to affect behavior within social dilemmas (for summary overviews of this literature, *see* Goetze and Orbell 1988; Ledyard 1995; Lichbach 1996; E. Ostrom 1990, 1998; E. Ostrom, Gardner, and Walker 1994; Sally 1995; Schroeder 1995).

Figure 5.1: N-person Social Dilemma



Note: *N* players choose between cooperating (C) or not cooperating (~C). When j individuals cooperate, their payoffs are always lower than the j-1 individuals who do not cooperate. The predicted outcome is that no one will cooperate and all players will receive X benefits. The temptation (*T*) not to cooperate is the increase in benefit any co-operator would receive for switching to not cooperating. If all cooperate, they all receive G-X more benefits than if all do not cooperate.

assumption that rational individuals are helplessly trapped in social dilemmas from which they cannot extract themselves without inducement or sanctions applied from the outside. Many policies based on this assumption have been subject to major failure and have exacerbated the very problems they were intended to ameliorate (Arnold and Campbell 1986; Baland and Platteau 1996; Morrow and Hull 1996). Policies based on the assumptions that individuals can learn how to devise well-tailored rules and cooperate conditionally when they participate in the design of institutions affecting them are more successful in the field (Berkes 1989; Bromley *et al.* 1992; Ellickson 1991; Feeny *et al.* 1990; McCay and Acheson 1987; McKean and Ostrom 1995; Pinkerton 1989; Yoder 1994).

Fifth, the image of citizens we provide in our textbooks affects the long-term viability of democratic regimes. Introductory textbooks that presume rational citizens will be passive consumers of political life – the masses – and focus primarily on the role of politicians and officials at a national level – the elite – do not inform future citizens of a democratic polity of the actions they need to know and can undertake. While many political scientists claim to eschew teaching the

normative foundations of a democratic polity, they actually introduce a norm of cynicism and distrust without providing a vision of how citizens could do anything to challenge corruption, rent seeking,[7] or poorly designed policies.

The remainder of this article is divided into six sections. In the first I briefly review the theoretical predictions of currently accepted rational choice theory related to social dilemmas. The next will summarize the challenge to the sole reliance on a complete model of rationality presented by extensive experimental research. Then I examine two major empirical findings that begin to show how individuals achieve results that are 'better than rational' (Cosmides and Tooby 1994) by building conditions in which reciprocity, reputation, and trust can help to overcome the strong temptations of short-run self-interest. The following section raises the possibility of developing second-generation models of rationality, and the next develops an initial theoretical scenario. I conclude by examining the implications of placing reciprocity, reputation, and trust at the core of an empirically tested, behavioural theory of collective action.

Theoretical predictions for social dilemmas

The term 'social dilemma' refers to a large number of situations in which individuals make independent choices in an interdependent situation (Dawes 1975: 1980; R. Hardin 1971). In all N-person social dilemmas, a set of participants has a choice of contributing (C) or not contributing ($-C$) to a joint benefit. While I represent this as an either-or choice in Figure 5.1, it frequently is a choice of how much to contribute rather than whether to contribute or not.

If everyone contributes, they get a net positive benefit (G). Everyone faces a temptation (T) to shift from the set of contributors to the set of those who do not contribute. The theoretical prediction is that everyone will shift and that no one will contribute. If this happens, then the outcome will be at the intercept. The difference between the predicted outcome and everyone contributing is $G - X$. Since the less-valued payoff is at a Nash equilibrium, no one is independently motivated to change his or her choice, given the choices of other participants. These situations are dilemmas because at least one outcome exists that yields greater advantage for *all* participants. Thus, a Pareto-superior alternative exists, but rational participants making isolated choices are not predicted to realize this outcome. A conflict is posed between individual and group rationality. The problem of collective action raised by social dilemmas is to find a way to avoid Pareto-inferior equilibria and to move closer to the optimum. Those who find ways to coordinate strategies in some fashion receive a 'cooperators' dividend' equal to the difference between the predicted outcome and the outcome achieved.

Many models of social dilemmas exist in the literature (*see* Schelling 1978 and Lichbach 1996 for reviews of alternative formalizations). In all models, a

7. The term 'rent seeking' refers to nonproductive activities directed toward creating opportunities for profits higher than would be obtained in an open, competitive market.

set of individuals is involved in a game in which a strategy leading to a Nash equilibrium for a single iteration of the game yields less than an optimal outcome for all involved. The equilibrium is thus Pareto inferior. The optimal outcome could be achieved if those involved 'cooperated' by selecting strategies other than those prescribed by an equilibrium solution to a noncooperative game (Harsanyi and Selten 1988). Besides these assumptions regarding the structure of payoffs in the one-shot version of the game, other assumptions are made in almost all formal models of social dilemmas. (1) All participants have common knowledge of the exogenously fixed structure of the situation and of the payoffs to be received by all individuals under all combinations of strategies. (2) Decisions about strategies are made independently, often simultaneously. (3) In a symmetric game, all participants have available the same strategies. (4) No external actor (or central authority) is present to enforce agreements among participants about their choices.

When such a game is finitely repeated, participants are assumed to solve the game through backward induction. Assumptions about the particular payoff functions differ. Rather than describe the Nash equilibrium and Pareto-efficient outcome for all models considered in this article, aggregate behavior consistent with the Nash equilibrium will be described as zero cooperation, while behavior consistent with the efficient outcome will be described as 100 percent cooperation.

The grim predictions evoked considerable empirical challenges as well as important theoretical breakthroughs. The predictions ran counter to so many everyday experiences that some scholars turned to survey and field studies to examine the level of voluntary contributions to public goods (*see* Lichbach 1995 for a summary). Others turned to the experimental lab and confirmed much higher than predicted levels of cooperation in one-shot experiments. Game theorists were challenged to rethink their own firm conclusions and to pose new models of when cooperation might emerge (*see* Benoit and Krishna 1985).

The introduction of two kinds of uncertainty into repeated games – about the number of repetitions and about the types of players participating in a social dilemma – has led to more optimistic predictions. When individuals, modelled as fully rational actors with low discount rates, interact in a repeated social dilemma whose end point is determined stochastically, it is now theoretically well established that it is *possible* for them to achieve optimal or near optimal outcomes and avoid the dominant strategies of one-shot and finitely repeated games that yield suboptimal outcomes (Fudenberg and Maskin 1986). This is possible when players achieve self-enforcing equilibria by committing themselves to punish noncooperators sufficiently to deter noncooperation. Kreps *et al.* (1982) introduced a second kind of uncertainty related to whether all the players use complete rationality as their guide to action. The probability of the presence of an 'irrational' player, who reciprocates cooperation with cooperation, is used as the grounds for a completely rational player to adopt the strategy of cooperating early in a sequence of games and switching to noncooperation at the end. Once either of these two forms of uncertainty is introduced, the number of possible equilibria explode in number (Abreau 1988). Everything is predicted: optimal outcomes, the Pareto-inferior Nash equilibria, and everything in between.

To generate predictions other than noncooperation, theorists using standard rational choice theory have found it necessary to assume real uncertainty about the duration of a situation or to assume that some players are 'irrational' in their willingness to reciprocate cooperation with cooperation. To assume that if some players *irrationally* choose reciprocity, then others can *rationally* choose reciprocity is a convoluted explanation – to say the least – of the growing evidence that reciprocity is a core norm used by many individuals in social dilemma situations.

The lack of a general fit

In all the social sciences, experiments have been conducted on various types of social dilemmas for several decades. While some scholars question the value of laboratory experiments for testing the predictions of major theories in the social sciences, this method has many advantages. First, one can design experiments that test multiple predictions from the same theory under controlled conditions. Second, replication is feasible. Third, researchers can challenge whether a particular design adequately captures the theoretically posited variables and conduct further experiments to ascertain how changes in a design affect outcomes. The evidence discussed below is based on multiple studies by diverse research teams. Fourth, experimental methods are particularly relevant for studying human choice under diverse institutional arrangements. Subjects in experimental studies draw on the modes of analysis and values they have learned throughout their lives to respond to diverse incentive structures. Experiments thus allow one to test precisely whether individuals behave within a variety of institutional settings as predicted by theory (Plott 1979; Smith 1982).

In this section, I will summarize four consistently replicated findings that directly challenge the general fit between behavior observed in social-dilemma experiments and the predictions of noncooperative game theory using complete rationality and complete information for one-shot and finitely repeated social dilemmas. I focus first on the fit between theory and behavior, because the theoretical predictions are unambiguous and have influenced so much thinking across the social sciences. Experiments on market behavior do fit the predictions closely (*see* Davis and Holt 1993 for an overview). If one-shot and finitely repeated social-dilemma experiments were to support strongly the predictions of noncooperative game theory, then we would have a grounded theory with close affinities to a vast body of economic theory for which there is strong empirical support. We would need to turn immediately to the problem of indefinitely repeated situations for which noncooperative game theory faces an embarrassment of too many equilibria. As it turns out, we have a different story to tell. The four general findings are as follows:

1. High levels of initial cooperation are found in most types of social dilemmas, but the levels are consistently less than optimal.
2. Behavior is not consistent with backward induction in finitely repeated social dilemmas.
3. Nash equilibrium strategies are not good predictors at the individual level.
4. Individuals do not learn Nash equilibrium strategies in repeated social dilemmas.

High but suboptimal levels of initial cooperation

Most experimental studies of social dilemmas with the structure of a public-goods provision problem have found levels of cooperative actions in one-shot games, or in the first rounds of a repeated game, that are significantly above the predicted level of zero.[8] 'In a wide variety of treatment conditions, participants rather persistently contributed 40 to 60 percent of their token endowments to the [public good], far in excess of the 0 percent contribution rate consistent with a Nash equilibrium' (Davis and Holt 1993: 325). Yet, once an experiment is repeated, cooperation levels in public-good experiments tend to decline. The individual variation across experiment sessions can be very great.[9] While many have focused on the unexpectedly high rates of cooperation, it is important to note that in sparse institutional settings with no feedback about individual contributions, cooperation levels never reach the optimum. Thus, the prediction of zero levels of cooperation can be rejected, but cooperation at a suboptimal level is consistently observed in sparse institutional settings.

Behavior in social dilemmas inconsistent with backward induction

In all finitely repeated experiments, players are predicted to look ahead to the last period and determine what they would do in that period. In the last period, there is no future interaction; the prediction is that they will not cooperate in that round. Since that choice would be determined at the beginning of an experiment, the players are presumed to look at the second to-last period and ask themselves what they would do there. Given that they definitely would not cooperate in the

8. *See* Isaac, McCue, and Plott 1985; Kim and Walker 1984; Marwell and Ames 1979, 1980, 1981; Orbell and Dawes 1991, 1993; Schneider and Pommerehne 1981. An important exception to this general finding is that when subjects are presented with an experimental protocol with an opportunity to invest tokens in a common-pool resource (the equivalent of harvesting from a common pool), they tend to overinvest substantially in the initial rounds (*see* E. Ostrom, Gardner and Walker 1994 and comparison of public goods and common-pool resource experiments in Goetze 1994 and E. Ostrom and Walker 1997). Ledyard (1995) considers common-pool resource dilemmas to have the same underlying structure as public good dilemmas, but behavior in common-pool resource experiments without communication is consistently different from public good experiments without communication. With repetition, outcomes in common-pool resource experiments approach the Nash equilibrium from below rather than from above, as is typical in public good experiments.

9. In a series of eight experiments with different treatments conducted by Isaac, Walker and Thomas (1984), in which the uniform theoretical prediction was zero contributions, contribution rates varied from nearly 0 percent to around 75 percent of the resources available to participants.

last period, it is assumed that they also would not cooperate in the second-to-last period. This logic would then extend backward to the first round (Luce and Raiffa 1957: 98–9).

While backward induction is still the dominant method used in solving finitely repeated games, it has been challenged on theoretical grounds (Binmore 1997: R. Hardin 1997). Furthermore, as discussed above, uncertainty about whether others use norms like tit-for-tat rather than follow the recommendations of a Nash equilibrium may make it rational for a player to signal a willingness to cooperate in the early rounds of an iterated game and then defect at the end (Kreps *et al.* 1982). What is clearly the case from experimental evidence is that players do not use backward induction in their decision-making plans in an experimental laboratory Amnon Rapoport (1997: 122) concludes from a review of several experiments focusing on resource dilemmas that 'subjects are not involved in or capable of backward induction'.[10]

Nash equilibrium strategies do not predict individual behavior in social dilemmas

From the above discussion, it is obvious that individuals in social dilemmas tend not to use the predicted Nash equilibrium strategy, even though this is a good predictor at both an individual and group level in other types of situations. While outcomes frequently approach Nash equilibria at an aggregate level, the variance of individual actions around the mean is extremely large. When groups of eight subjects made appropriation decisions in repeated common-pool resource experiments of 20 to 30 rounds, the unique symmetric Nash equilibrium strategy was never played (Walker, Gardner and Ostrom 1990). Nor did individuals use Nash equilibrium strategies in repeated public good experiments (Dudley 1993; Isaac and Walker 1991, 1993). In a recent set of thirteen experiments involving seven players making ten rounds of decisions without communication or any other institutional structure, Walker *et al.* (1997) did not observe a single individual choice of a symmetric Nash equilibrium strategy in the 910 opportunities available to subjects. Chan *et al.* (1996: 58) also found little evidence to support the use of Nash equilibria when they examined the effect of heterogeneity of income on outcomes: 'It is clear that the outcomes of the laboratory sessions reported here cannot be characterized as Nash equilibria outcomes'.

Individuals do not learn Nash equilibrium strategies in social dilemmas

In repeated experiments without communication or other facilitating institutional conditions, levels of cooperation fall (rise) toward the Nash equilibrium in public-good (common-pool resource) experiments. Some scholars have speculated that

10. Subjects in Centipede games also do not use backward induction (*see* McKelvey and Palfrey 1992).

it just takes some time and experience for individuals to learn Nash equilibrium strategies (Ledyard 1995). But this does not appear to be the case. In all repeated experiments, there is considerable pulsing as subjects obtain outcomes that vary substantially with short spurts of increasing and decreasing levels of cooperation while the general trend is toward an aggregate that is consistent with a Nash equilibrium (Isaac, McCue and Plott 1985; E. Ostrom, Gardner and Walker 1994).[11] Furthermore, there is substantial variation in the strategies followed by diverse participants within the same game (Dudley 1993; Isaac and Walker 1988b; E. Ostrom, Gardner and Walker 1994).

It appears that subjects learn something other than Nash strategies in finitely repeated experiments. Isaac, Walker and Williams (1994) compare the rate of decay when experienced subjects are explicitly told that an experiment will last 10, 40, or 60 rounds. The rate of decay of cooperative actions is inversely related to the number of decision rounds. Instead of learning the noncooperative strategy, subjects appear to be learning how to cooperate at a moderate level for even longer periods. Cooperation rates approach zero only in the last few periods, whenever these occur.

Two internal ways out of the social dilemmas

The combined effect of these four frequently replicated, general findings represents a strong rejection of the predictions derived from a complete model of rationality. Two more general findings are also contrary to the predictions of currently accepted models. At the same time, they also begin to show how individuals are able to obtain results that are substantially 'better than rational' (Cosmides and Tooby 1994), at least as *rational* has been defined in currently accepted models. The first is that simple, cheap talk allows individuals an opportunity to make conditional promises to one another and potentially to build trust that others will reciprocate. The second is the capacity to solve second-order social dilemmas that change the structure of the first-order dilemma.

Communication and collective action

In noncooperative game theory, players are assumed to be unable to make enforceable agreements.[12] Thus, communication is viewed as cheap talk (Farrell 1987). In a social dilemma, self-interested players are expected to use communication to try to convince others to cooperate and promise cooperative action, but then to choose the Nash equilibrium strategy when they make their

11. The pulsing cannot be explained using a complete model of rationality, but it can be explained as the result of a heuristic used by subjects to raise or lower their investments depending upon the average return achieved on the most recent round (*see* E. Ostrom, Gardner and Walker 1994).

12. In cooperative game theory, in contrast, it is assumed that players can communicate and make enforceable agreements (Harsanyi and Selten 1988: 3).

private decision (Barry and Hardin 1982: 381; Farrell and Rabin 1996: 113).[13] Or, as Gary Miller (1992: 25) expresses it: 'It is obvious that simple communication is not sufficient to escape the dilemma'.[14]

From this theoretical perspective, face-to-face communication should make no difference in the outcomes achieved in social dilemmas. Yet, consistent, strong, and replicable findings are that substantial increases in the levels of cooperation are achieved when individuals are allowed to communicate face to face.[15] This holds true across all types of social dilemmas studied in laboratory settings and in both one-shot and finitely repeated experiments. In a meta-analysis of more than 100 experiments involving more than 5,000 subjects conducted by economists, political scientists, sociologists, and social psychologists, Sally (1995) finds that opportunities for face-to-face communication in one-shot experiments significantly raise the cooperation rate, on average, by more than 45 percentage points. When subjects are allowed to talk before each decision round in repeated experiments, they achieve 40 percentage points more on average than in repeated games without communication. No other variable has as strong and consistent an effect on results as face-to-face communication. Communication even has a robust and positive effect on cooperation levels when individuals are not provided with feedback on group decisions after every round (Cason and Khan 1996).

The efficacy of communication is related to the capability to talk face to face. Sell and Wilson (1991, 1992), for example, developed a public-good experiment in which subjects could signal promises to cooperate via their computer terminal. There was much less cooperation than in the face-to-face experiments using the same design (Isaac and Walker 1988a, 1991). Rocco and Warglien (1995) replicated all aspects of prior common-pool resource experiments, including the efficacy of face-to-face communication.[16] They found, however, that subjects who had to rely on computerized communication did not achieve the same increase in efficiency as did those who were able to communicate face to face.[17] Palfrey and Rosenthal (1988) report that no significant difference occurred in a provision point public-good experiment in which subjects could send a computerized message stating whether they intended to contribute.

13. In social-dilemma experiments, subjects make anonymous decisions and are paid privately. The role of cheap talk in coordination experiments is different since there is no dominant strategy. In this case, preplay communication may help players coordinate on one of the possible equilibria (*see* Cooper, DeJong and Forsythe 1992).

14. As Aumann (1974) cogently points out, the players are faced with the problem that whatever they agree upon has to be self-enforcing. That has led Aumann and most game theorists to focus entirely on Nash equilibria which, once reached, are self-enforcing. In coordination games, cheap talk can be highly efficacious.

15. *See* E. Ostrom, Gardner and Walker 1994 for extensive citations to studies showing a positive effect of the capacity to communicate. Dawes, McTavish and Shaklee 1977; Frey and Bohnet 1996; Hackett, Schlager and Walker 1994; Isaac and Walker 1988a, 1991; Orbell, Dawes and van de Kragt 1990; Orbell, van de Kragt and Dawes 1988, 1991; E. Ostrom, Gardner and Walker 1994; Sally 1995.

16. Moir (1995) also replicated these findings with face-to-face communication.

17. Social psychologists have found that groups who perform tasks using electronic media do much better if they have had an opportunity to work face to face prior to the use of electronic communication only (Hollingshead, McGrath and O'Connor 1993).

The reasons offered by those doing experimental research for why communication facilitates cooperation include (1) transferring information from those who can figure out an optimal strategy to those who do not fully understand what strategy would be optimal, (2) exchanging mutual commitment, (3) increasing trust and thus affecting expectations of others' behavior, (4) adding additional values to the subjective payoff structure, (5) reinforcement of prior normative values, and (6) developing a group identity (Davis and Holt 1993; Orbell, Dawes and van de Kragt 1990; Orbell, van de Kragt and Dawes 1988; E. Ostrom and Walker 1997). Carefully crafted experiments demonstrate that the effect of communication is not primarily due to the first reason. When information about the individual strategy that produces an optimal joint outcome is clearly presented to subjects who are not able to communicate, the information makes little difference in outcomes achieved (Isaac, McCue and Plott 1985; Moir 1995).

Consequently, exchanging mutual commitment, increasing trust, creating and reinforcing norms, and developing a group identity appear to be the most important processes that make communication efficacious. Subjects in experiments do try to extract mutual commitment from one another to follow the strategy they have identified as leading to their best joint outcomes. They frequently go around the group and ask each person to promise the others that they will follow the joint strategy. Discussion sessions frequently end with such comments as: 'Now remember everyone that we all do much better if we all follow *X* strategy' (*see* transcripts in E. Ostrom, Gardner and Walker 1994). In repeated experiments, subjects use communication opportunities to lash out verbally at unknown individuals who did not follow mutually agreed strategies, using such evocative terms as scumbuckets and finks. Orbell, van de Kragt and Dawes (1988) summarize the findings from ten years of research on one-shot public-good experiments by stressing how many mutually reinforcing processes are evoked when communication is allowed.[18] Without increasing mutual trust in the promises that are exchanged, however, expectations of the behavior of others will not change. Given the very substantial difference in outcomes, communication is most likely to affect individual trust that others will keep to their commitments. As discussed below, the relationships among trust, conditional commitments, and a reputation for being trustworthy are key links in a second-generation theory of boundedly rational and moral behavior.

As stakes increase and it is difficult to monitor individual contributions, communication becomes less efficacious, however. E. Ostrom, Gardner and Walker (1994) found that subjects achieved close to fully optimal results when each subject had relatively low endowments and was allowed opportunities for face-to-face communication. When endowments were substantially increased – increasing the temptation to cheat on prior agreements – subjects achieved far more in communication experiments as contrasted to non-communication experiments but less than in small-stake situations. Failures to achieve collective action in field settings in which communication has been feasible point out that communication alone is not a sufficient mechanism to assure successful collective action under all conditions.

18. *See* also Banks and Calvert (1992a, 1992b) for a discussion of communication in incomplete information games.

Innovation and collective action

Changing the rules of a game or using scarce resources to punish those who do not cooperate or keep agreements are usually not considered viable options for participants in social dilemmas, since these actions create public goods. Participants face a second-order social dilemma (of equal or greater difficulty) in any effort to use costly sanctions or change the structure of a game (Oliver 1980). The predicted outcome of any effort to solve a second-order dilemma is failure.

Yet, participants in many field settings and experiments do exactly this. Extensive research on how individuals have governed and managed common-pool resources has documented the incredible diversity of rules designed and enforced by participants themselves to change the structure of underlying social-dilemma situations (Blomquist 1992; Bromley *et al.* 1992; Lam 1998; McKean 1992; E. Ostrom 1990; Schlager 1990; Schlager and Ostrom 1993; Tang 1992). The particular rules adopted by participants vary radically to reflect local circumstances and the cultural repertoire of acceptable and known rules used generally in a region. Nevertheless, general design principles characterize successfully self-organised, sustainable, local, regional, and international regimes (E. Ostrom 1990). Most robust and long-lasting common-pool regimes involve clear mechanisms for monitoring rule conformance and graduated sanctions for enforcing compliance. Thus, few self-organised regimes rely entirely on communication alone to sustain cooperation in situations that generate strong temptations to break mutual commitments. Monitors – who may be participants themselves – do not use strong sanctions for individuals who rarely break rules. Modest sanctions indicate to rule breakers that their lack of conformance has been observed by others. By paying a modest fine, they rejoin the community in good standing and learn that rule infractions are observed and sanctioned. Repeated rule breakers are severely sanctioned and eventually excluded from the group. Rules meeting these design principles reinforce contingent commitments and enhance the trust participants have that others are also keeping their commitments.

In field settings, innovation in rules usually occurs in a continuous trial-and-error process until a rule system is evolved that participants consider yields substantial net benefits. Given the complexity of the physical world that individuals frequently confront, they are rarely ever able to 'get the rules right' on the first or second try (E. Ostrom 1990). In highly unpredictable environments, a long period of trial and error is needed before individuals can find rules that generate substantial positive net returns over a sufficiently long time horizon. Nonviolent conflict may be a regular feature of successful institutions when arenas exist to process conflict cases regularly and, at times, to innovate new rules to cope with conflict more effectively (V. Ostrom 1987; V. Ostrom, Feeny, and Picht 1993).

In addition to the extensive field research on changes that participants make in the structure of situations they face, subjects in a large number of experiments have also solved second-order social dilemmas and consequently moved the outcomes in their first-order dilemmas closer to optimal levels (Dawes, Orbell, and van de Kragt 1986; Messick and Brewer 1983; Rutte and Wilke 1984; Sato 1987; van de Kragt, Orbell and Dawes 1983; Yamagishi 1992). Toshio Yamagishi (1986), for example, conducted experiments with subjects who had earlier completed a questionnaire including items from a scale measuring trust. Subjects who ranked higher on the trust scale consistently contributed about 20 percent more to collective goods than those who ranked lower. When given an opportunity to contribute to a substantial 'punishment fund' to be used to fine the individual who contributed the least to their joint outcomes, however, low-trusting individuals contributed significantly more to the punishment fund and also achieved the highest level of cooperation. In the last rounds of this experiment, they were contributing 90 percent of their resources to the joint fund. These results, which have now been replicated with North American subjects (Yamagishi 1988a, 1988b), show that individuals who are initially the least trusting are willing to contribute to sanctioning systems and then respond more to a change in the structure of the game than those who are initially more trusting.

E. Ostrom, Walker and Gardner (1992) also examined the willingness of subjects to pay a 'fee' in order to 'fine' another subject. Instead of the predicted zero use of sanctions, individuals paid fees to fine others at a level significantly above zero.[19] When sanctioning was combined with a single opportunity to communicate or a chance to discuss and vote on the creation of their own sanctioning system, outcomes improved dramatically. With only a single opportunity to communicate, subjects were able to obtain an average of 85 percent of the optimal level of investments (67 percent with the costs of sanctioning subtracted). Those subjects who met face to face and agreed by majority vote on their own sanctioning system achieved 93 percent of optimal yield. The level of defections was only 4 percent, so that the costs of the sanctioning system were low, and net benefits were at a 90 percent level (E. Ostrom, Walker and Gardner 1992).

Messick and his colleagues have undertaken a series of experiments designed to examine the willingness of subjects to act collectively to change institutional structures when facing common-pool resource dilemmas (see Messick *et al.* 1983; Samuelson *et al.* 1984; C. Samuelson and Messick 1986). In particular, they have repeatedly given subjects the opportunity to relinquish their individual decisions concerning withdrawals from the common resource to a leader who is given the authority to decide for the group. They have found that 'people want to change the rules and bring about structural change when they observe that the common resource is being depleted' (C. Samuelson and Messick 1995; 147). Yet, simply having an unequal distribution of outcomes is not a sufficient inducement to affect the decision whether to change institutional structure.

What do these experiments tell us? They complement the evidence from field settings and show that individuals temporarily caught in a social-dilemma

19. Furthermore, they invested more when the fine was lower or when it was more efficacious, and they tended to direct their fines to those who had invested the most on prior rounds. Given the cost of the sanctioning mechanism, subjects tended to overuse it and to end up with a less efficient outcome after sanctioning costs were subtracted from their earnings. This finding is consistent with the Boyd and Richerson (1992) result that moralistic strategies may result in negative net outcomes.

structure are likely to invest resources to innovate and change the structure itself in order to improve joint outcomes. They also strengthen the earlier evidence that the currently accepted, noncooperative gametheoretical explanation relying on a particular model of the individual does not adequately predict behavior in one-shot and finitely repeated social dilemmas. Cooperative game theory does not provide a better explanation. Since both cooperative and noncooperative game theory predict extreme values, neither provides explanations for the conditions that tend to enhance or detract from cooperation levels.

The really big puzzle in the social sciences is the development of a consistent theory to explain why cooperation levels vary so much and why specific configurations of situational conditions increase or decrease cooperation in first- or second-level dilemmas. This question is important not only for our scientific understanding but also for the design of institutions to facilitate individuals' achieving higher levels of productive outcomes in social dilemmas. Many structural variables affect the particular innovations chosen and the sustainability and distributional consequences of these institutional changes (Knight 1992). A coherent theory of institutional change is not within reach, however, with a theory of individual choice that predicts no innovation will occur. We need a second-generation theory of boundedly rational, innovative, and normative behavior.

Toward second-generation models of rationality

First-generation models of rational choice are powerful engines of prediction when strong competition eliminates players who do not aggressively maximize immediate external values. While incorrectly confused with a general theory of human behavior, complete rationality models will continue to be used productively by social scientists, including the author. But the thin model of rationality needs to be viewed, as Selten (1975) points out, as the limiting case of bounded or incomplete rationality. Consistent with all models of rational choice is a general theory of human behavior that views all humans as complex, fallible learners who seek to do as well as they can given the constraints that they face and who are able to learn heuristics, norms, rules, and how to craft rules to improve achieved outcomes.

Learning heuristics, norms, and rules

Because individuals are boundedly rational, they do not calculate a complete set of strategies for every situation they face. Few situations in life generate information about all potential actions that one can take, all outcomes that can be obtained, and all strategies that others can take. In a model of complete rationality, one simply assumes this level of information. In field situations, individuals tend to use heuristics – rules of thumb – that they have learned over time regarding responses that tend to give them good outcomes in particular kinds of situations. They bring these heuristics with them when they participate in laboratory experiments. In frequently encountered, repetitive situations, individuals learn better and better heuristics that are tailored to particular situations.

With repetition, sufficiently large stakes, and strong competition, individuals may learn heuristics that approach best-response strategies.

In addition to learning instrumental heuristics, individuals also learn to adopt and use norms and rules. By *norms* I mean that the individual attaches an internal valuation – positive or negative – to taking particular types of action. Crawford and Ostrom (1995) refer to this internal valuation as a delta parameter that is added to or subtracted from the objective costs of an action.[20] Andreoni (1989) models individuals who gain a 'warm glow' when they contribute resources that help others more than they help themselves in the short term. Knack (1992) refers to negative internal valuations as 'duty'.[21] Many norms are learned from interactions with others in diverse communities about the behavior that is expected in particular types of situations (Coleman 1987). The change in preferences represents the internalization of particular moral lessons from life (or from the training provided by one's elders and peers).[22] The strength of the commitment (Sen 1977) made by an individual to take particular types of future actions (telling the truth, keeping promises) is reflected in the size of the delta parameter. After experiencing repeated benefits from other people's cooperative actions, an individual may resolve that s/he should always initiate cooperative actions in the future.[23] Or, after many experiences of being the 'sucker' in such experiences, an individual may resolve never to be the first to cooperate.

Since norms are learned in a social milieu, they vary substantially across cultures, across individuals within any one culture, within individuals across different types of situations they face, and across time within any particular situation. The behavioural implications of assuming that individuals acquire norms do not vary substantially from the assumption that individuals learn to use heuristics. One may think of norms as heuristics that individuals adopt from a moral perspective, in that these are the kinds of actions they wish to follow in living their life. Once some members of a population acquire norms of behavior, they affect the expectations of others.

By *rules* I mean that a group of individuals has developed shared understandings that certain actions in particular situations must, must not, or may be undertaken and that sanctions will be taken against those who do not conform. The distinction

20. When constructing formal models, one can include overt delta parameters in the model (*see* Crawford and Ostrom 1995; Palfrey and Rosenthal 1988). Alternatively, one can assume that these internal delta parameters lead individuals to enter new situations with differing probabilities that they will follow norms such as reciprocity. These probabilities not only vary across individuals but also increase or decrease as a function of the specific structural parameters of the situation and, in repeated experiments, the patterns of behavior and outcomes achieved in that situation over time.

21. The change in valuations that an individual may attach to an action-outcome linkage may be generated strictly internally or may be triggered by external observation and, thus, a concern with how others will evaluate the normative appropriateness of actions.

22. Gouldner (1960; 171) considers norms of reciprocity to be universal and as important in most cultures as incest taboos, even though the 'concrete formulations may vary with time and place'.

23. *See* Selten (1986) for a discussion of his own and John Harsanyi's (1977) conception of 'rule utilitarianism' as contrasted to 'act utilitarianism'.

between internalized but widely shared norms for what are appropriate actions in broad types of situations and rules that are self-consciously adopted for use in particular situations is at times difficult to draw when doing fieldwork. Analytically, individuals can be thought of as learning norms of behavior that are general and fit a wide diversity of particular situations. Rules are artifacts related to particular actions in specific situations (V. Ostrom 1980, 1997). Rules are created in private associations as well as in more formalized public institutions, where they carry the additional legal weight of being enforced legal enactments.[24] Rules can enhance reciprocity by making mutual commitments clear and overt. Alternatively, rules can assign authority to act so that benefits and costs are distributed inequitably and thereby destroy reliance on positive norms.

Reciprocity: An especially important class of norms

That humans rapidly learn and effectively use heuristics, norms, and rules is consistent with the lessons learned from evolutionary psychology (*see* Barkow, Cosmides and Tooby 1992), evolutionary game theory (*see* Guth and Kliemt 1996; Hirshleifer and Rasmusen 1989),[25] biology (Trivers 1971) and bounded rationality (Selten 1990, 1991; Selten, Mitzkewitz and Uhlich 1997; Simon 1985). Humans appear to have evolved specialized cognitive modules for diverse tasks, including making sense out of what is seen (Marr 1982), inferring rules of grammar by being exposed to adult speakers of a particular language (Pinker 1994) and increasing their long-term returns from interactions in social dilemmas (Cosmides and Tooby 1992). Humans dealt with social dilemmas related to rearing and protecting offspring, acquiring food, and trusting one another to perform future promised action millennia before such oral commitments could be enforced by external authorities (de Waal 1996). Substantial evidence has been accumulated (and reviewed in Cosmides and Tooby 1992) that humans inherit a strong capacity to learn reciprocity norms and social rules that enhance the opportunities to gain benefits from coping with a multitude of social dilemmas.

Reciprocity refers to a family of strategies that can be used in social dilemmas involving (1) an effort to identify who else is involved, (2) an assessment of the likelihood that others are conditional cooperators, (3) a decision to cooperate initially with others if others are trusted to be conditional cooperators, (4) a refusal to cooperate with those who do not reciprocate, and punishment of those who betray trust. All reciprocity norms share the common ingredients that individuals tend to react to the positive actions of others with positive responses and the negative actions of others with negative responses. Reciprocity is a basic norm taught in all societies (*see* Becker 1990; Blau 1964; Gouldner 1960; Homans 1961; Oakerson 1993; V. Ostrom 1997; Thibaut and Kelley 1959).

24. Crawford and Ostrom (1995) discuss these issues in greater depth. *See* also Piaget ([1932] 1969).

25. The evolutionary approach has been strongly influenced by the work of Robert Axelrod (*see*, in particular, Axelrod 1984, 1986; Axelrod and Hamilton 1981; and Axelrod and Keohane 1985).

By far the most famous reciprocal strategy – tit-for-tat – has been the subject of considerable study from an evolutionary perspective. In simulations, pairs of individuals are sampled from a population, and they then interact with one another repeatedly in a prisoners' dilemma game. Individuals are each modeled as if they had inherited a strategy that included the fixed maxims of always cooperate, always defect, or the reciprocating strategy of tit-for-tat (cooperate first, and then do whatever the others did in the last round). Axelrod and Hamilton (1981) and Axelrod (1984) have shown that when individuals are grouped so that they are more likely to interact with one another than with the general population, and when the expected number of repetitions is sufficiently large, reciprocating strategies such as tit-for-tat can successfully invade populations composed of individuals following an all-defect strategy. The size of the population in which interactions are occurring may need to be relatively small for reciprocating strategies to survive potential errors of players (Bendor and Mookherjee 1987; but *see* Boyd and Richerson 1988, 1992; Hirshleifer and Rasmusen 1989; Yamagishi and Takahashi 1994).

The reciprocity norms posited to help individuals gain larger cooperators' dividends depend upon the willingness of participants to use retribution to some degree. In tit-for-tat, for example, an individual must be willing to 'punish' a player who defected in the last round by defecting in the current round. In grim trigger, an individual must be willing to cooperate initially but then 'punish' everyone for the rest of the game if any defection is noticed in the current round.[26]

Human beings do not inherit particular reciprocity norms via a biological process. The argument is more subtle. Individuals inherit an acute sensitivity for learning norms that increase their own long-term benefits when confronting social dilemmas with others who have learned and value similar norms. The process of growing up in any culture provides thousands of incidents (learning trials) whereby parents, siblings, friends, and teachers provide the specific content of the type of mutual expectations prevalent in that culture. As Mueller (1986) points out, the first dilemmas that humans encounter are as children. Parents reward and punish them until cooperation is a learned response. In the contemporary setting, corporate managers strive for a trustworthy corporate reputation by continuously reiterating and rewarding the use of key principles or norms by corporate employees (Kreps 1990).

Since particular reciprocity norms are *learned,* not everyone learns to use the same norms in all situations. Some individuals learn norms of behavior that are not so 'nice'. Clever and unscrupulous individuals may learn how to lure others into dilemma situations and then defect on them. It is possible to gain substantial

26. The grim trigger has been used repeatedly as a support for cooperative outcomes in infinitely (or indefinitely) repeated games (Fudenberg and Maskin 1986). In games in which substantial joint benefits are to be gained over the long term from mutual cooperation, the threat of the grim trigger is thought to be sufficient to encourage everyone to cooperate. A small error on the part of one player or exogenous noise in the payoff function, however, makes this strategy a dangerous one to use in larger groups, where the cooperators' dividend may also be substantial.

resources by such means, but one has to hide intentions and actions, to keep moving, or to gain access to power over others. In any group composed only of individuals who follow reciprocity norms, skills in detecting and punishing cheaters could be lost. If this happens, it will be subject to invasion and substantial initial losses by clever outsiders or local deviants who can take advantage of the situation. Being too trusting can be dangerous. The presence of some untrustworthy participants hones the skills of those who follow reciprocity norms.

Thus, individuals vary substantially in the probability that they will use particular norms, in how structural variables affect their level of trust and willingness to reciprocate cooperation in a particular situation, and in how they develop their own reputation. Some individuals use reciprocity only in situations in which there is close monitoring and strong retribution is likely. Others will only cooperate in dilemmas when they have publicly committed themselves to an agreement and have assurances from others that their trust will be returned. Others find it easier to build an external reputation by building their own personal identity as someone who always trusts others until proven wrong. If this trust proves to be misplaced, then they stop cooperating and either exit the situation or enter a punishment phase. As Hoffman, McCabe and Smith (1996a; 23–4) express it:

> A one-shot game in the laboratory is part of a life-long sequence, not an isolated experience that calls for behavior that deviates sharply from one's reputational norm. Thus we should expect subjects to rely upon reciprocity norms in experimental settings unless they discover in the process of participating in a particular experiment that reciprocity is punished and other behaviours are rewarded. In such cases they abandon their instincts and attempt other strategies that better serve their interests.

In any population of individuals, one is likely to find some who use one of three reciprocity norms when they confront a repeated social dilemma:[27]

1. Always cooperate first; stop cooperating if others do not reciprocate; punish noncooperators if feasible.
2. Cooperate immediately only if one judges others to be trustworthy; stop cooperating if others do not reciprocate; punish noncooperators if feasible.
3. Once cooperation is established by others, cooperate oneself; stop cooperating if others do not reciprocate; punish noncooperators if feasible.

In addition, one may find at least three other norms:

27. This is not the complete list of all types of reciprocity norms, but it captures the vast majority.

1. Never cooperate.
2. Mimic (1) or (2), but stop cooperating if one can successfully free ride on others.
3. Always cooperate (an extremely rare norm in all cultures).

The proportion of individuals who follow each type of norm will vary from one subpopulation to another and from one situation to another.[28] Whether reciprocity is advantageous to individuals depends sensitively on the proportion of other individuals who use reciprocity and on an individual's capacity to judge the likely frequency of reciprocators in any particular situation and over time. When there are many others who use a form of reciprocity that always cooperates first, then even in one-shot situations cooperation may lead to higher returns when diverse situations are evaluated together. Boundedly rational individuals would expect other boundedly rational individuals to follow a *diversity* of heuristics, norms, and strategies rather than expect to find others who adopt a single strategy – except in those repeated situations in which institutional selection processes sort out those who do not search out optimal strategies. Investment in detection of other individuals' intentions and actions improves one's own outcomes. One does not have to assume that others are 'irrational' in order for it to be rational to use reciprocity (Kreps *et al.* 1982).

Evidence of the use of reciprocity in experimental settings

Laboratory experiments provide evidence that a substantial proportion of individuals use reciprocity norms even in the very short-term environments of an experiment (McCabe, Rassenti and Smith 1996). Some evidence comes from experiments on ultimatum games. In such games, two players are asked to divide a fixed sum of money. The first player suggests a division to the second, who then decides to accept or reject the offer. If the offer is accepted, then the funds are divided as proposed. If it is rejected, then both players receive zero. The predicted equilibrium is that the first player will offer a minimal unit to the second player, who will then accept anything more than zero. This prediction has repeatedly been falsified, starting with the work of Guth, Schmittberger and Schwarze (1982; *see* Frey and Bohnet 1996; Guth and Tietz 1990; Roth 1995; Samuelson, Gale, and Binmore 1995).[29] Subjects assigned to the first position tend to offer substantially

28. The proportion of individuals who follow the sixth norm – cooperate always – will be minuscule or nonexistent. Individuals following the first norm will be those, along with those following the sixth norm, who cooperate in the first few rounds of a finitely repeated experimental social dilemma without prior communication. Individuals following the second norm will cooperate (immediately) in experiments if they have an opportunity to judge the intentions and trustworthiness of the other participants and expect most of the others to be trustworthy. Those following the third norm will cooperate (after one or a few rounds) in experiments in which others cooperate.

29. The results obtained by Hoffman, McCabe and Smith (1996b) related to dictator games under varying conditions of social distance are also quite consistent with the behavioural approach of this chapter.

more than the minimum unit. They frequently offer the 'fair' division of splitting the sum. Second movers tend to reject offers that are quite small. The acceptance level for offers tends to cluster around different values in diverse cultures (Roth *et al.* 1991). Given that the refusal to accept the funds offered contradicts a basic tenet in the complete model of rationality, these findings have represented a major challenge to the model's empirical validity in this setting.

Several hypotheses have been offered to explain these findings, including a 'punishment hypothesis' and a 'learning hypothesis'.

The punishment hypothesis is in essence a reciprocity argument. In contrast to adaptive learning, punishment attributes a motive to the second mover's rejection of an unequal division asserting that it is done to punish the first mover for unfair treatment. This propensity toward negative reciprocity is the linchpin of the argument. Given this propensity, first movers should tend to shy away from the perfect equilibrium offer out of fear of winding up with nothing (Abbink *et al.* 1996; 6).

Abbink and his colleagues designed an experiment in which the prediction of the learning and punishment hypotheses is clearly different and found strong support for the punishment hypothesis. 'We found that second movers were three times more likely to reject the unequal split when doing so punished the first mover [...] than when doing so rewarded the first mover' (Abbink *et al.* 1996; 15–6). Consequently, second movers do appear to punish first movers who propose unfair divisions.

Two additional findings from one-shot social dilemmas provide further evidence of the behavioural propensities of subjects. First, those who intend to cooperate in a particular one-shot social dilemma also expect cooperation to be returned by others at a much higher rate than those who intend to defect (Dawes, McTavish and Shaklee 1977; Dawes, Orbell and van de Kragt 1986). As Orbell and Dawes (1991; 519) summarize their own work: 'One of our most consistent findings throughout these studies – a finding replicated by others' work – is that cooperators expect significantly more cooperation than do defectors'. Second, when there is a choice whether to participate in a social dilemma, those who intend to cooperate exhibit a greater willingness to enter such transactions (Orbell and Dawes 1993). Given these two tendencies, reciprocators are likely to be more optimistic about finding others following the same norm and disproportionately enter more voluntary social dilemmas than nonreciprocators. Given both propensities, the feedback from such voluntary activities will generate confirmatory evidence that they have adopted a norm which serves them well over the long run.

Thus, while individuals vary in their propensity to use reciprocity, the evidence from experiments shows that a substantial proportion of the population drawn on by social science experiments has sufficient trust that others are reciprocators to cooperate with them even in one-shot, no-communication experiments. Furthermore, a substantial proportion of the population is also willing to punish noncooperators (or individuals who do not make fair offers) at a cost to themselves. Norms are learned from prior experience (socialization) and are affected by situational variables yielding systematic differences among experimental designs.

The level of trust and resulting levels of cooperation can be increased by (1) providing subjects with an opportunity to see one another (Frey and Bohnet 1996; Orbell and Dawes 1991), (2) allowing subjects to choose whether to enter or exit a social-dilemma game (Orbell and Dawes 1991, 1993; Orbell, Schwartz-Shea and Simmons 1984; Schuessler 1989; Yamagishi 1988c; Yamagishi and Hayashi 1996), (3) sharing the costs equally if a minimal set voluntarily contributes to a public good (Dawes, Orbell, and van de Kragt 1986), (4) providing opportunities for distinct punishments of those who are not reciprocators (Abbink *et al.* 1996; McCabe, Rassenti and Smith 1996) and, as discussed above, (5) providing opportunities for face-to-face communication.

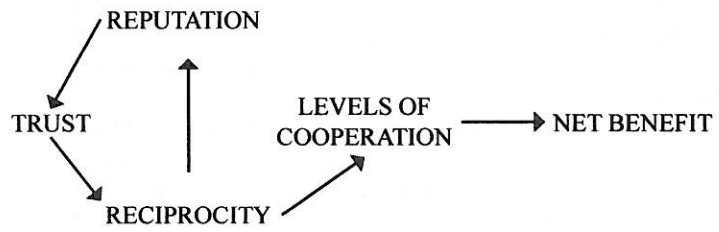
The core relationships: reciprocity, reputation, and trust

When many individuals use reciprocity, there is an incentive to acquire a *reputation* for keeping promises and performing actions with short-term costs but longterm net benefits (Keohane 1984; Kreps 1990; Milgrom, North, and Weingast 1990; Miller 1992). Thus, trustworthy individuals who trust others with a reputation for being trustworthy (and try to avoid those who have a reputation for being untrustworthy) can engage in mutually productive social exchanges, even though they are dilemmas, so long as they can limit their interactions primarily to those with a reputation for keeping promises. A reputation for being trustworthy, or for using retribution against those who do not keep their agreements or keep up their fair share, becomes a valuable asset. In an evolutionary context, it increases fitness in an environment in which others use reciprocity norms. Similarly, developing *trust* in an environment in which others are trustworthy is also an asset (Braithwaite and Levi 1998; Fukuyama 1995; Gambetta 1988; Putnam 1993). Trust is the expectation of one person about the actions of others that affects the first person's choice, when an action must be taken before the actions of others are known (Dasgupta 1997: 5). In the context of a social dilemma, trust affects whether an individual is willing to initiate cooperation in the expectation that it will be reciprocated. Boundedly rational individuals enter situations with an initial probability of using reciprocity based on their own prior training and experience.

Thus, at the core of a behavioural explanation are the links between the trust that individuals have in others, the investment others make in trustworthy reputations, and the probability that participants will use reciprocity norms (*see* Figure 5.2). This mutually reinforcing core is affected by structural variables as well as the past experiences of participants. In the initial round of a repeated dilemma, individuals do or do not initiate cooperative behavior based on their own norms, how much trust they have that others are reciprocators (based on any information they glean about one another) and how structural variables affect their own and their expectation of others' behavior.

If initial levels of cooperation are moderately high, then individuals may learn to trust one another, and more may adopt reciprocity norms. When more individuals use reciprocity norms, gaining a reputation for being trustworthy is a better investment. Thus, levels of trust, reciprocity, and reputations for being

Figure 5.2: The Core Relationship



trustworthy are positively reinforcing. This also means that a decrease in any one of these can lead to a downward spiral. Instead of explaining levels of cooperation directly, this approach leads one to link structural variables to an inner triangle of trust, reciprocity, and reputation as these, in turn, affect levels of cooperation and net benefits.

Communication and the core relationships

With these core relationships, one can begin to explain why repeated face-to-face communication substantially changes the structure of a situation (*see* discussion in E. Ostrom, Gardner and Walker 1994: 199). With a repeated chance to see and talk with others, a participant can assess whether s/he trusts others sufficiently to try to reach a simple contingent agreement regarding the level of joint effort and its allocation. In a contingent agreement, individuals agree to contribute X resources to a common effort so long as at least Y others also contribute. Contingent agreements do not need to include all those who benefit. The benefit to be obtained from the contribution of Y proportion of those affected may be so substantial that some individuals are willing to contribute so long as Y proportion of others also agree and perform.

Communication allows individuals to increase (or decrease) their trust in the reliability of others.[30] When successful, individuals change their expectations from the initial probability that others use reciprocity norms to a higher probability that others will reciprocate trust and cooperation. When individuals are symmetric in assets and payoffs, the simplest agreement is to share a contribution level equally that closely approximates the optimum joint outcome. When individuals are not symmetric, finding an agreement is more difficult, but various fairness norms can be used to reduce the time and effort needed to achieve an agreement (*see* Hackett, Dudley and Walker 1995; Hackett, Schlager and Walker 1994).

30. Frank, Gilovich and Regan (1993) found, for example, that the capacity of subjects to predict whether others would play cooperatively was significantly better than chance after a face-to-face group discussion. Kikuchi, Watanabe and Yamagishi (1996) found that high trusters predicted other players' trustworthiness significantly better than did low trusters.

Contingent agreements may deal with punishment of those who do not cooperate (Levi 1988). How to punish noncooperative players, keep one's own reputation, and sustain any initial cooperation that has occurred in N-person settings is more difficult than in two-person settings.[31] In an N-person, uncertain situation, it is difficult to interpret from results that are less than expected whether one person cheated a lot, several people cheated a little, someone made a mistake, or everyone cooperated and an exogenous random variable reduced the expected outcome. If there is no communication, then the problem is even worse. Without communication and an agreement on a sharing formula, individuals can try to signal a willingness to cooperate through their actions, but no one has agreed to any particular contribution. Thus, no one's reputation (external or internal) is at stake.

Once a verbal agreement in an N-person setting is reached, that becomes the focal point for further action within the context of a particular ongoing group. If everyone keeps to the agreement, then no further reaction is needed by someone who is a reciprocator. If the agreement is not kept, however, then an individual following a reciprocity norm – without any prior agreement regarding selective sanctions for nonconformance – needs to punish those who did not keep their commitment. A frequently posited punishment is the grim trigger, whereby a participant plays the Nash equilibrium strategy forever upon detecting any cheating. Subjects in repeated experiments frequently discuss the use of a grim trigger to punish mild defections but reject the idea because it would punish everyone – not just the cheater(s) (E. Ostrom, Gardner and Walker 1994). A much less drastic punishment strategy is the measured reaction. 'In a measured reaction, a player reacts mildly (if at all) to a small deviation from an agreement. Defections trigger mild reactions instead of harsh punishments. If defections continue over time, the measured response slowly moves from the point of agreement toward the Nash equilibrium' (pp. 199–200).

For several reasons, this makes sense as the initial 'punishment' phase in an N-person setting with a minimal institutional structure and no feedback concerning individual contributions. If only a small deviation occurs, then the cooperation of most participants is already generating positive returns. By keeping one's own reaction close to the agreement, one keeps up one's own reputation for cooperation, keeps cooperation levels higher, and makes it easier to restore full conformance. Using something like a grim trigger immediately leads to the unravelling of the agreement and the loss of substantial benefits over time. To supplement the measured reaction, effort is expended on determining who is

31. In a two-person situation of complete certainty, individuals can easily follow the famous tit-for-tat (or tit-for-tat or exit) strategy even without communication. When a substantial proportion of individuals in a population follows this norm, and they can identify with whom they have interacted in the past (to either refuse future interactions or to punish prior uncooperative actions) and when discount rates are sufficiently low, tit-for-tat has been shown to be a highly successful strategy, yielding higher payoffs than are available to those using other strategies (Axelrod 1984, 1986). With communication, it is even easier.

breaking the agreement, on using verbal rebukes to try to get that individual back in line, and on avoiding future interactions with that individual.[32]

Thus, understanding how trust, reciprocity, and reputation feed one another (or their lack, which generates a cascade of negative effects) helps to explain why repeated, face-to-face communication has such a major effect. Coming to an initial agreement and making personal promises to one another places at risk an individual's own identity as one who keeps one's word, increases trust, and makes reciprocity an even more beneficial strategy. Tongue-lashing can be partially substituted in a small group for monetary losses and, when backed by measured responses, can keep many groups at high levels of cooperation. Meeting only once can greatly increase trust, but if some individuals do not cooperate immediately, the group never has a further opportunity to hash out these problems. Any evidence of lower levels of cooperation undermines the trust established in the first meeting, and there is no further opportunity to build trust or use verbal sanctioning. It is also clearer now why sending anonymous, computerized messages is not as effective as face-to-face communication. Individuals judge one another's trustworthiness by watching facial expressions and hearing the way something is said. It is hard to establish trust in a group of strangers who will make decisions independently and privately without seeing and talking with one another.

Illustrative theoretical scenarios

I have tried to show the need for the development of second-generation models of rationality in order to begin a coherent synthesis of what we know from empirical research on social dilemmas. Rather than try to develop a new formal model, I have stayed at the theoretical level to identify the attributes of human behavior that should be included in future formal models. The individual attributes that are particularly important in explaining behavior in social dilemmas include the expectations individuals have about others' behavior (trust), the norms individuals learn from socialization and life's experiences (reciprocity) and the identities individuals create that project their intentions and norms (reputation). Trust, reciprocity, and reputation can be included in formal models of individual behavior (*see* the works cited by Boyd and Richerson 1988; Guth and Yaari 1992; Nowak and Sigmund 1993).

In this section, I construct theoretical scenarios of how exogenous variables combine to affect endogenous structural variables that link to the core set of relationships shown in Figure 5.2. It is not possible to relate all structural variables

32. In a series of 18 common-pool resource experiments, each involving eight subjects in finitely repeated communication experiments, E. Ostrom, Gardner and Walker (1994; 215) found that subjects kept to their agreements or used measured responses in two-thirds of the experiments. In these experiments, joint yields averaged 89 percent of optimum. In the six experiments in which some players deviated substantially from agreements and measured responses did not bring them back to the agreement, cooperation levels were substantially less, and yields averaged 43 percent of optimum (which is still far above zero levels of cooperation).

in one large causal model, given the number of important variables and the fact that many depend for their effect on the values of other variables. It is possible, however, to produce coherent, cumulative, theoretical scenarios that start with relatively simple baseline models. One can then begin the systematic exploration of what happens as one variable is changed. Let me illustrate what I mean by theoretical scenarios.

Let us start with a scenario that should be conducive to cooperation – a small group of ten farmers who own farms of approximately the same size. These farmers share the use of a creek for irrigation that runs by their relatively flat properties. They face the problem each year of organising one collective workday to clear out the fallen trees and brush from the prior winter. All ten expect to continue farming into the indefinite future. Let us assume that the creek delivers a better water supply directly in response to how many days of work are completed. All farmers have productive opportunities for their labor that return more at the margin than the return they would receive from their own input into this effort. Thus, free riding and hoping that the others contribute labor is objectively attractive. The value to each farmer, however, of participation in a successful collective effort to clear the creek is greater than the costs of participating.

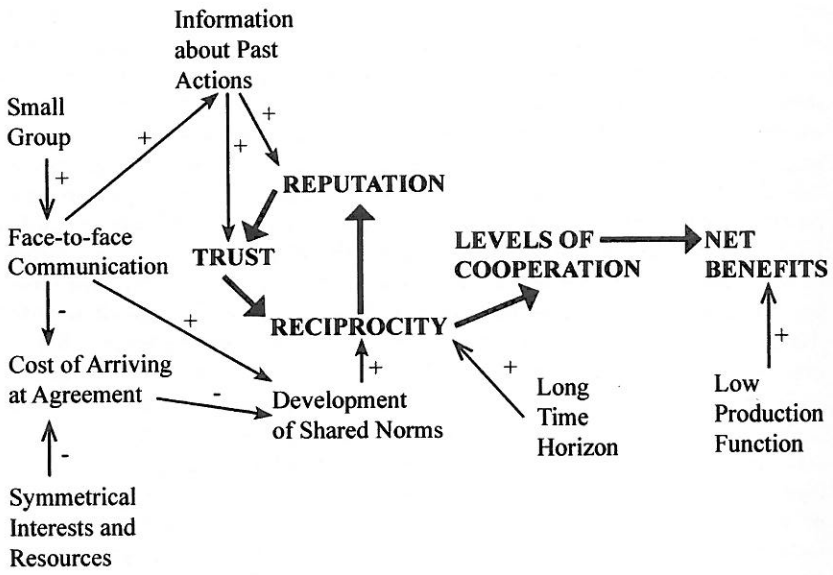
Now let us examine how some structural variables affect the likelihood of collective action (*see* Figure 5.3). As a small group, it would be easy for them to engage in face-to-face communication. Since their interests and resources are relatively symmetric, arriving at a fair, contingent agreement regarding how to share the work should not be too difficult. One simple agreement that is easy to monitor is that they all work on the same day, but each is responsible for clearing the part of the creek going through his or her property. Conformance to such an agreement would be easy to verify. While engaged in discussions, they can reinforce the importance of everyone participating in the workday. In face-to-face meetings, they can also gossip about anyone who failed to participate in the past, urge them to change their ways, and threaten to stop all labor contributions if they do not 'shape up'. Given the small size of the group, its symmetry, and the relatively low cost of providing the public good, combined with the relatively long time horizon, we can predict with some confidence that a large proportion of individuals facing such a situation will find a way to cooperate and overcome the dilemma. Not only does the evidence from experimental research support that prediction, but also substantial evidence from the field is consistent with this explanation (*see* E. Ostrom 1998).

This is a rough but coherent causal theory that uses structural variables (small size, symmetry of assets and resources, long time horizon, and a low-cost production function) to predict with high probability that participants can themselves solve this social dilemma. Changes in any of the structural variables of this relatively easy scenario affect that prediction. Even a small change may suffice to reverse the predicted outcome. For example, assume that another local farmer buys five parcels of land with the plan to farm them for a long time. Now there are only six farmers, but one of them holds half the relevant assets. If that farmer shares the norm that it is fair to share work allocated to a collective benefit

in the same ratio as the benefits are allocated, then the increased heterogeneity will not be a difficult problem to overcome. They would agree – as farmers around the world have frequently agreed (*see* Lam 1998; Tang 1992) – to share the work in proportion to the amount of land they own. If the new farmer uses a different concept of fairness, then the smaller group may face a more challenging problem than the larger group due to its increased heterogeneity.

Now, assume that the five parcels of land are bought by a local developer to hold for future use as a suburban housing development. The time horizon of one of the six actors – the developer – is extremely short with regard to investments in irrigation. From the developer's perspective, he is not a 'free rider', as he sees no benefit to clearing out the creek. Thus, such a change actually produces several: A decrease in the N of the group, an introduction of an asymmetry of interests and resources, and the presence of one participant with half the resources but a short time horizon and no interest in the joint benefit. This illustrates how changes in one structural variable can lead to a cascade of changes in the others, and thus how difficult it is to make simple bivariate hypotheses about the effect of one variable on the level of cooperation. In particular, this smaller group is much less likely to cooperate than the larger group of ten symmetric farmers, exactly the reverse of the standard view of the effect of group size.

Figure 5.3: A Simple Scenario



Implications

The implications of developing second-generation models of empirically grounded, boundedly rational, and moral decision making are substantial. Puzzling research questions can now be addressed more systematically. New research questions will open up. We need to expand the type of research methods regularly used in political science. We need to increase the level of understanding among those engaged in formal theory, experimental research, and field research across the social and biological sciences. The foundations of policy analysis need rethinking. And civic education can be based on empirically validated theories of collective action empowering citizens to use the 'science and art of association' (Tocqueville [1835 and 1840] 1945) to help sustain democratic polities in the twenty-first century.

Implications for research

What the research on social dilemmas demonstrates is a world of *possibility* rather than of *necessity*. We are neither trapped in inexorable tragedies nor free of moral responsibility for creating and sustaining incentives that facilitate our own achievement of mutually productive outcomes. We cannot adopt the smug presumption of those earlier group theorists who thought groups would always form whenever a joint benefit would be obtained. We can expect many groups to fail to achieve mutually productive benefits due to their lack of trust in one another or to the lack of arenas for low-cost communication, institutional innovation, and the creation of monitoring and sanctioning rules (V. Ostrom 1997). Nor can we simply rest assured that only one type of institution exists for all social dilemmas, such as a competitive market, in which individuals pursuing their own preferences are led to produce mutually productive outcomes. While new institutions often facilitate collective action, the key problems are to design new rules, motivate participants to conform to rules once they are devised, and find and appropriately punish those who cheat. Without individuals viewing rules as appropriate mechanisms to enhance reciprocal relationships, no police force and court system on earth can monitor and enforce all the needed rules on its own. Nor would most of us want to live in a society in which police were really the thin blue line enforcing all rules.

While I am proposing a further development of second-generation theories of rational choice, theories based on complete but thin rationality will continue to play an important role in our understanding of human behavior. The clear and unambiguous predictions stemming from complete rational choice theories will continue to serve as a critical benchmark in conducting empirical studies and for measuring the success or failure of any other explanation offered for observed behavior. A key research question will continue to be: What is the difference between the predicted equilibrium of a complete rationality theory and observed behavior? Furthermore, game theorists are already exploring ways of including reputation, reciprocity, and various norms of behavior in game-theoretic models (*see* Abbink *et al.* 1996; Girth 1995; Kreps 1990; Palfrey and Rosenthal 1988; Rabin 1994; Selten 1990, 1991). Thus, bounded and complete rationality models may become more complementary in the next decade than appears to be the case today.

For political scientists interested in diverse institutional arrangements, complete rational choice theories provide well-developed methods for analysing the vulnerability of institutions to the strategies devised by talented, analytically sophisticated, short-term hedonists (Brennan and Buchanan 1985). Any serious institutional analysis should include an effort to understand how institutions – including ways of organising legislative procedures, formulas used to calculate electoral weights and minimal winning coalitions, and international agreements on global environmental problems – are vulnerable to manipulation by calculating, amoral participants.[33] In addition to the individuals who have learned norms of reciprocity in any population, others exist who may try to subvert the process so as to obtain very substantial returns for themselves while ignoring the interests of others. One should always know the consequences of letting such individuals operate in any particular institutional setting.

The most immediate research questions that need to be addressed using second-generation models of human behavior relate to the effects of structural variables on the likelihood of organising for successful modes of collective action. It will not be possible to relate all structural variables in one large causal theory, given that they are so numerous and that many depend for their effect on the values of other variables. What is possible, however, is the development of coherent, cumulative, theoretical scenarios that start with relatively simple baseline models and then proceed to change one variable at a time, as briefly illustrated above. From such scenarios, one can proceed to formal models and empirical testing in field and laboratory settings. The kind of *theory* that emerges from such an enterprise does not lead to the global bivariate (or even multivariate) predictions that have been the ideal to which many scholars have aspired. Marwell and Oliver (1993) have constructed such a series of theoretical scenarios for social dilemmas involving large numbers of heterogeneous participants in collective action. They have come to a similar conclusion about the nature of the theoretical and empirical enterprise: 'This is not to say that general theoretical predictions are impossible using our perspective, only that they cannot be simple and global. Instead, the predictions that we can validly generate must be complex, interactive, and conditional' (p. 25).

As political scientists, we need to recognize that political systems are complexly organised and that we will rarely be able to state that one variable is always positively or negatively related to a dependent variable. One can do comparative statics, but one must know the value of the other variables and not simply assume that they vary around the average.

The effort to develop second-generation models of boundedly rational and moral behavior will open up a variety of *new* questions to be pursued that are of major importance to all social scientists and many biologists interested in

33. Consequently, research on the effect of institutional arrangements on strategies and outcomes continues to be crucial to future developments. *See* Agrawal 1998; Alt and Shepsle 1990; Bates 1989; Dasgupta 1993; Eggertsson 1990; Gibson 1999; Levi 1997; V. Ostrom 1997; V. Ostrom, Feeny, and Picht 1993; Scharpf 1997.

human behavior. Among these questions are: How do individuals gain trust in other individuals? How is trust affected by diverse institutional arrangements? What verbal and visual clues are used in evaluating others' behavior? How do individuals gain common understanding so as to craft and follow self-organised arrangements (V. Ostrom 1990)? John Orbell (personal communication) posits a series of intriguing questions: 'Why do people join together in these games in the first place? How do we select partners in these games? How do our strategies for selecting individual partners differ from our strategies for adding or removing individuals from groups?'

An important set of questions is related to how institutions enhance or restrict the building of mutual trust, reciprocity, and reputations. A recent set of studies on tax compliance raises important questions about the trust heuristics used by citizens and their reactions to governmental efforts to monitor compliance (*see* Scholz 1998). Too much monitoring may have the counterintuitive result that individuals feel they are *not* trusted and thus become less trustworthy (Frey 1993). Bruno Frey (1997) questions whether some formal institutional arrangements, such as social insurance and paying people to contribute effort, reduce the likelihood that individuals continue to place a positive intrinsic value on actions taken mainly because of internal norms. Rather, they may assume that formal organisations are charged with the responsibility of taking care of joint needs and that reciprocity is no longer needed (*see also* Taylor 1987).

Since all rules legitimate the use of sanctions against those who do not comply, rules can be used to assign benefits primarily to a dominant coalition. Those who are, thus, excluded have no motivation to cooperate except in order to avoid sanctions. Using first-generation models, that is what one expects in any case. Using second-generation models, one is concerned with how constitutional and collective-choice rules affect the distribution of benefits and the likelihood of reciprocal cooperation. While much research has been conducted on long-term successful self-organised institutions, less has been documented about institutions that never quite got going or failed after years of success. More effort needs to be made to find reliable archival information concerning these failed attempts and why they failed.

It may be surprising that I have relied so extensively on experimental research. I do so for several reasons. As theory becomes an ever more important core of our discipline, experimental studies will join the ranks of basic empirical research methods for political scientists. As an avid field researcher for the past 35 years, I know the importance and difficulty of testing theory in field settings – particularly when variables function interactively. Large-scale field studies will continue to be an important source of empirical data, but frequently they are a very expensive and inefficient method for addressing how institutional incentives combine to affect individual behavior and outcomes. We can advance much faster and more coherently when we examine hypotheses about contested elements among diverse models or theories of a coherent framework. Careful experimental research designs frequently help sort out competing hypotheses more effectively than does trying to find the precise combination of variables in the field. By adding

experimental methods to the battery of field methods already used extensively, the political science of the twenty-first century will advance more rapidly in acquiring well-grounded theories of human behavior and of the effect of diverse institutional arrangements on behavior. Laboratory research will still need to be complemented by sound field studies to meet the criteria of external validity.

Implications for policy

Using a broader theory of rationality leads to potentially different views of the state. If one sees individuals as helpless, then the state is the essential external authority that must solve social dilemmas for everyone. If, however, one assumes individuals can draw on heuristics and norms to solve some problems and create new structural arrangements to solve others, then the image of what a national government might do is somewhat different. There is a very considerable role for large-scale governments, including national defence, monetary policy, foreign policy, global trade policy, moderate redistribution, keeping internal peace when some groups organise to prey on others, provision of accurate information and of arenas for resolving conflicts with national implications, and other large-scale activities. But national governments are too small to govern the global commons and too big to handle smaller scale problems.

To achieve a complex, multitiered governance system is quite difficult. Many types of questions are raised. How do different kinds of institutions support or undermine norms of reciprocity both within hierarchies (Miller 1992) and among members of groups facing collective action problems (Frohlich and Oppenheimer 1970; Galjart 1992)? Field studies find that monitoring and graduated sanctions are close to universal in all robust common-pool resource institutions (E. Ostrom 1990). This tells us that without some external support of such institutions, it is unlikely that reciprocity *alone* completely solves the more challenging common-pool resource problems. Note that sanctions are graduated rather than initially severe. Our current theory of crime – based on a strict expected value theory – does not explain this. If people can learn reciprocity as the fundamental norm for organising their lives, and if they agree to a set of rules contingent upon others following these rules, then graduated sanctions do something more than deter rule infractions.

Reciprocity norms can have a dark side. If punishment consists of escalating retribution, then groups who overcome social dilemmas may be limited to very tight circles of kin and friends, who cooperate only with one another, embedded in a matrix of hostile relationships with outsiders (R. Hardin 1995). This pattern can escalate into feuds, raids, and overt warfare (Boyd and Richerson 1992; Chagnon 1988; Elster 1985; Kollock 1993). Or tight circles of individuals who trust one another may discriminate against anyone of a different colour, religion, or ethnicity. A focus on the return of favours for favours can also be the foundation for corruption. It is in everyone else's interest that some social dilemmas are *not* resolved, such as those involved in monopolies and cartel formation, those that countervene basic moral standards and legal relationships, and those that restrict

the opportunities of an open society and an expanding economy. Policies that provide alternative opportunities for those caught in dysfunctional networks are as important as those that stimulate and encourage positive networks (Dasgupta 1997).

Implications for civic education

Human history teaches us that autocratic governments often wage war on their own citizens as well as on those of other jurisdictions. Democracies are characterized by the processing of conflict among individuals and groups without resort to massive killings. Democracies are, however, themselves fragile institutions that are vulnerable to manipulation if citizens and officials are not vigilant (V. Ostrom 1997). For those who wish the twenty-first century to be one of peace, we need to translate our research findings on collective action into materials written for high school and undergraduate students. All too many of our textbooks focus exclusively on leaders and, worse, only national-level leaders. Students completing an introductory course on American government, or political science more generally, will not learn that they play an essential role in sustaining democracy. Citizen participation is presented as contacting leaders, organising interest groups and parties, and voting. That citizens need additional skills and knowledge to resolve the social dilemmas they face is left unaddressed. Their moral decisions are not discussed. We are producing generations of cynical citizens with little trust in one another, much less in their governments. Given the central role of trust in solving social dilemmas, we may be creating the very conditions that undermine our own democratic ways of life. It is ordinary persons and citizens who craft and sustain the workability of the institutions of everyday life. We owe an obligation to the next generation to carry forward the best of our knowledge about how individuals solve the multiplicity of social dilemmas – large and small – that they face.

References

Abbink, K., Bolton, G. E., Sadrieh, A. and Tang, F. F. (1996) 'Adaptive Learning versus Punishment in Ultimatum Bargaining', *Discussion paper no. B-381,* Rheinische FriedrichWilhelms-Universitat Bonn.

Abreau, D. (1988) 'On the Theory of Infinitely Repeated Games with Discounting', *Econometrica* 80 (4): 383–96.

Agrawal, A. (1998) *Greener Pastures: Exchange, Politics and Community among a Mobile Pastoral People*, Durham, NC: Duke University Press.

Alchian, A. A. (1950) 'Uncertainty, Evolution, and Economic Theory', *Journal of Political Economy* 58 (3): 211–21.

Alchian, A. A. and Demsetz, H. (1972) 'Production, Information Costs, and Economic Organisation', *American Economic Review* 62(December): 777–95.

Alt, James E. and Shepsle, K. A. (eds) (1990) *Perspectives on Positive Political Economy*, New York: Cambridge University Press.

Andreoni, J. (1989) 'Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence', *Journal of Political Economy* 97 (December): 1, 447–51, 458.

Arnold, J. E. M. and Campbell, J. G. (1986) 'Collective Management of Hill Forests in Nepal: The Community Forestry Development Project', in *Proceedings of the Conference on Common Property Resource Management,* National Research Council. Washington, DC: National Academy Press, pp. 425–54.

Aumann, R. J. (1974) 'Subjectivity and Correlation in Randomized Strategies', *Journal of Mathematical Economics* 1(March): 67–96.

Axelrod, Robert (1984) *The Evolution of Cooperation*, New York: Basic Books.

— (1986) 'An Evolutionary Approach to Norms', *American Political Science Review* 80 (December): 1095–111.

Axelrod, R. and Hamilton, W. D. (1981) 'The Evolution of Cooperation', *Science* 211 (March): 1390–6.

Axelrod, R. and Keohane, R. O. (1985) 'Achieving Cooperation under Anarchy: Strategies and Institutions', *World Politics* 38 (October): 226–54.

Baland, J.-M. and Platteau, J.-P. (1996) *Halting Degradation of Natural Resources: Is There a Role for Rural Communities,* Oxford: Clarendon Press.

Banks, Jeffrey S. and Calvert, Randall L. (1992a) 'A Battle-of-the-Sexes Game with Incomplete Information', *Games and Economic Behavior* 4 (July): 347–72.

— (1992b) 'Communication and Efficiency in Coordination Games', Working paper, Department of Economics and Department of Political Science, University of Rochester, New York.

Barkow, J. H., Cosmides, L. and Tooby, J. (eds) (1992) *The Adapted Mind. Evolutionary Psychology and the Generation of Culture*, Oxford: Oxford University Press.

Barry, B. and Hardin, R. (1982) *Rational Man and Irrational Society? An Introduction and Source Book*, Beverly Hills, CA: Sage.

Bates, R. H. (1989) *Beyond the Miracle of the Market: The Political Economy of Agrarian Development in Kenya*, New York: Cambridge University Press.

Becker, L. C. (1990) *Reciprocity*, Chicago: University of Chicago Press.

Bendor, J. and Dilip, M. (1987) 'Institutional Structure and the Logic of Ongoing Collective Action', *American Political Science Review* 81 (March): 129–54.

Benoit, J.-P. and Krishna, V. (1985) 'Finitely Repeated Games', *Econometrica* 53 (July): 905–22.

Berkes, F. (ed.) (1989) *Common Property Resources: Ecology and Community-Based Sustainable Development*, London: Belhaven.

Binmore, K. (1997) 'Rationality and Backward Induction', *Journal of Economic Methodology* 4:23–41.

Blau, P. M. (1964) *Exchange of Power in Social Life*, New York: Wiley.

Blomquist, W. (1992) *Dividing the Waters: Governing Groundwater in Southern California*, San Francisco, CA: Institute for Contemporary Studies Press.

Boudreaux, D. J. and Holcombe, R. G. (1989) 'Government by Contract', *Public Finance Quarterly* 17 (July): 264–80.

Boulding, K. E. (1963) 'Towards a Pure Theory of Threat Systems', *American Economic Review* 53 (May): 424–34.

Boyd, R. and Richerson, P. J. (1988) 'The Evolution of Reciprocity in Sizable Groups', *Journal of Theoretical Biology* 132 (June): 337–56.

— (1992) 'Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizable Groups', *Ethology and Sociobiology* 13 (May): 171–95.

Braithwaite, V. and Levi, M. (eds) (1998) *Trust and Governance*, New York: Russell Sage Foundation.

Brennan, G. and Buchanan, J. (1985) *The Reason of Rules*, Cambridge: Cambridge University Press.

Bromley, D. W., Feeny, D., McKean, M., Peters, P., Gilles, J., Oakerson, R. C., Runge, F. and Thomson, J. (eds) (1992) *Making the Commons Work: Theory, Practice, and Policy,* San Francisco, CA: Institute for Contemporary Studies Press.

Bullock, K. and Baden, J. (1977) 'Communes and the Logic of the Commons', in G. Hardin and J. Baden (eds), *Managing the Commons,* San Francisco, CA: Freeman, pp. 182–99.

Cason, T. N. and Khan, F. U. (1996) 'A Laboratory Study of Voluntary Public Goods Provision with Imperfect Monitoring and Communication', Working paper, Department of Economics, University of Southern California, Los Angeles.

Chagnon, N. A. (1988) 'Life Histories, Blood Revenge, and Warfare in a Tribal Population', *Science* 239 (February): 985–92.

Chan, K., Mestelman, S., Moir, R. and Muller, A. (1996) 'The Voluntary Provision of Public Goods under Varying Endowments', *Canadian Journal of Economics* 29 (1): 54–69.

Clark, A. (1995) 'Economic Reason: The Interplay of Individual Learning and External Structure', Working paper. Department of Philosophy, Washington University in St. Louis.

Coleman, J. S. (1987) 'Norms as Social Capital', in G. Radnitzky and P. Bernholz (eds) *Economic Imperialism: The Economic Approach Applied Outside the Field of Economics,* New York: Paragon House pp. 133–55.

Cook, K. S. and Levi, M. (1990) *The Limits of Rationality,* Chicago: University of Chicago Press.

Cooper, R., DeJong, D. V. and Forsythe, R. (1992) 'Communication in Coordination Games', *Quarterly Journal of Economics* 107 (2): 739–71.

Comes, R., Mason, C. F. and Sandler, T. (1986) 'The Commons and the Optimal Number of Firms', *Quarterly Journal of Economics* 101(August): 641–6.

Cosmides, L. and Tooby, J. (1992) 'Cognitive Adaptations for Social Exchange', in Jerome H. Barkow, L. Cosmides, and J. Tooby (eds) *The Adapted Mind, Evolutionary Psychology and the Generation of Culture,* New York: Oxford University Press, pp. 163–228.

— (1994) 'Better than Rational: Evolutionary Psychology and the Invisible Hand', *American Economic Review* 84 (May): 327–32.

Crawford, S. E. S. and Ostrom, E. (1995) 'A Grammar of Institutions', *American Political Science Review* 89 (September): 582–600.

Dasgupta, P. S. (1993) *An Inquiry into Well-Being and Destitution,* Oxford: Clarendon Press.

— (1997) 'Economic Development and the Idea of Social Capital', Working paper, Faculty of Economics, University of Cambridge.

Davis, D. D. and Holt, C. A. (1993) *Experimental Economics*, Princeton, NJ: Princeton University Press.

Dawes, R. M. (1975) 'Formal Models of Dilemmas in Social Decision Making', in M. F. Kaplan and S. Schwartz (eds) *Human Judgment and Decision Processes: Formal and Mathematical Approaches,* New York: Academic Press, pp. 87–108.

— (1980) 'Social Dilemmas', *Annual Review of Psychology* 31: 169–93.

Dawes, R. M., McTavish, J. and Shaklee, H. (1977) 'Behavior, Communication, and Assumptions about Other People's Behavior in a Commons Dilemma Situation', *Journal of Personality and Social Psychology* 35(1): 1–11.

Dawes, R. M., Orbell, J. M. and van de Kragt, A. (1986) 'Organising Groups for Collective Action', *American Political Science Review* 80 (December): 1171–85.

de Waal, F. (1996) *Good Natured: The Origins of Right and Wrong in Humans and Other Animals,* Cambridge, MA: Harvard University Press.

Dudley, D. (1993) 'Essays on Individual Behavior in Social Dilemma Environments: An Experimental Analysis', PhD diss., Indiana University.

Edney, J. (1979) 'Freeriders en Route to Disaster', *Psychology Today* 13 (December): 80–102.

Eggertsson, T. (1990) *Economic Behavior and Institutions,* New York: Cambridge University Press.

Ekeh, P. P. (1974) *Social Exchange Theory: The Two Traditions,* Cambridge, MA: Harvard University Press.

Ellickson, R. C. (1991) *Order without Law: How Neighbours Settle Disputes*, Cambridge, MA: Harvard University Press.

Elster, J. (1985) *Sour Grapes: Studies in the Subversion of Rationality,* Cambridge: Cambridge University Press.

Emerson, R. (1972a) 'Exchange Theory, Part I: A Psychological Basis for Social Exchange', in J. Berger, M. Zelditch and B. Anderson (eds) *Sociological Theories in Progress,* Vol. 2. Boston: Houghton Mifflin, pp. 38–57.

— (1972b) 'Exchange Theory, Part II: Exchange Relations and Networks', in J. Berger, M. Zelditch and B. Anderson (eds) *Sociological Theories in Progress,* Vol. 2. Boston: Houghton Mifflin, pp. 58–87.

Farrell, J. (1987) 'Cheap Talk, Coordination, and Entry', *Rand Journal of Economics*, 18 (Spring): 34–9.

Farrell, J. and Maskin, E. (1989) 'Renegotiation in Repeated Games', *Games and Economic Behavior* 1 (December): 327–60.

Farrell, J. and Rabin, M. (1996) 'Cheap Talk', *Journal of Economic Perspectives* 10 (Summer): 103–18.

Feeny, D., Berkes, F., McCay, B. J. and Acheson, J. M. (1990) 'The Tragedy of the Commons: Twenty-Two Years Later', *Human Ecology* 18 (1) :1–19.

Frank, R. H., Gilovich, T. and Regan, D. T. (1993) 'The Evolution of One-Shot Cooperation: An Experiment', *Ethology and Sociobiology* 14 (July): 247–56.

Frey, B. S. (1993) 'Does Monitoring Increase Work Effort? The Rivalry with Trust and Loyalty', *Economic Inquiry* 31 (October): 663–70.

— (1997) *Not Just for the Money: An Economic Theory of Personal Motivation,* Cheltenham, UK: Edward Elgar.

Frey, B. S. and Bohnet, I. (1996) 'Cooperation, Communication and Communitarianism: An Experimental Approach', *Journal of Political Philosophy* 4 (4): 322–36.

Frohlich, N. and Oppenheimer, J. (1970) 'I Get By with a Little Help from My Friends', *World Politics* 23(October): 104–20.

Fudenberg, D. and Maskin, E. (1986) 'The Folk Theorem in Repeated Games with Discounting or with Incomplete Information', *Econometrica* 54 (3): 533–54.

Fukuyama, F. (1995) *Trust: The Social Virtues and the Creation of Prosperity,* New York: Free Press.

Galjart, B. (1992) 'Cooperation as Pooling: A Rational Choice Perspective', *Sociologia Ruralis* 32 (4): 389–407.

Gambetta, D. (ed.) (1988) *Trust: Making and Breaking Cooperative Relations,* Oxford: Basil Blackwell.

Geddes, B. (1994) *Politician's Dilemma: Building State Capacity in Latin America,* Berkeley: University of California Press.

Gibson, C. (1999) *Politicians, Peasants and Poachers: The Political Economy of Wildlife in Africa,* Cambridge: Cambridge University Press.

Goetze, D. (1994) 'Comparing Prisoner's Dilemma, Commons Dilemma, and Public Goods Provision Designs in Laboratory Experiments', *Journal of Conflict Resolution* 38 (March): 56–86.

Goetze, D. and Orbell, J. (1988) 'Understanding and Cooperation in Social Dilemmas', *Public Choice* 57 (June): 275–9.

Gouldner, A. W. (1960) 'The Norm of Reciprocity: A Preliminary Statement', *American Sociological Review* 25 (April): 161–78.

Greif, A., Milgrom, P. and Weingast, B. R. (1994) 'Coordination, Commitment, and Enforcement: The Case of the Merchant Guild', *Journal of Political Economy* 102 (August): 745–76.

Grossman, S. J. and Hart, O. D. (1980) 'Takeover Bids, the Free-Rider Problem, and the Theory of the Corporation', *Bell Journal of Economics* 11 (Spring): 42–64.

Guth, W. (1995) 'An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives', *International Journal of Game Theory* 24 (4): 323–44.

Guth, W. and Kliemt, H. (1995) 'Competition or Cooperation. On the Evolutionary Economics of Trust, Exploitation and Moral Attitudes', Working paper, Humboldt University, Berlin.

— (1996) 'Towards a Completely Indirect Evolutionary Approach-a Note', *Discussion Paper* 82, Economics Faculty, Humboldt University, Berlin.

Guth, W., Schmittberger, R. and Schwarze, B. (1982) 'An Experimental Analysis of Ultimatum Bargaining', *Journal of Economic Behavior and Organisation* 3 (December): 367–88.

Guth, W. and Tietz, R. (1990) 'Ultimatum Bargaining Behavior. A Survey and Comparison of Experimental Results', *Journal of Economic Psychology* 11 (September): 417–49.

Guth, W. and Yaari, M. (1992) 'An Evolutionary Approach to Explaining Reciprocal Behavior in a Simple Strategic Game', in U. Witt (ed.) *Explaining Process and Change: Approaches to Evolutionary Economics,* Ann Arbor: University of Michigan Press, pp. 23–34.

Hackett, S., Dudley, D. and Walker, J. (1995) 'Heterogeneities, Information and Conflict Resolution: Experimental Evidence on Sharing Contracts', in R. O. Keohane and E. Ostrom (eds) *Local Commons and Global Interdependence: Heterogeneity and Cooperation in Two Domains,* London: Sage, pp. 93–124.

Hackett, S., Schlager, E. and Walker, J. (1994) 'The Role of Communication in Resolving Commons Dilemmas: Experimental Evidence with Heterogeneous Appropriators', *Journal of Environmental Economics and Management* 27 (September): 99–126.

Hamilton, W. D. (1964) 'The Genetical Evolution of Social Behavior', *Journal of Theoretical Biology* 7(July): 1–52.

Hardin, G. (1968) 'The Tragedy of the Commons', *Science* 162 (December): 1243–8.

Hardin, R. (1971) 'Collective Action as an Agreeable n-Prisoners' Dilemma', *Science* 16 (September-October): 472–81.

— (1995) *One for All: The Logic of Group Conflict,* Princeton, NJ: Princeton University Press.

— (1997) 'Economic Theories of the State', in D. C. Mueller (ed.) *Perspectives on Public Choice: A Handbook,* Cambridge: Cambridge University Press, pp. 21–34.

Hardy, C. J. and Latane, B. (1988) 'Social Loafing in Cheerleaders: Effects of Team Membership and Competition', *Journal of Sport and Exercise Psychology* 10 (March): 109–14.

Harsanyi, J. (1977) 'Rule Utilitarianism and Decision Theory', *Erkenntnis* 11 (May): 25–53.

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

Hirshleifer, D. and Rasmusen, E. (1989) 'Cooperation in a Repeated Prisoner's Dilemma with Ostracism', *Journal of Economic Behavior and Organisation* 12 (August): 87–106.

Hoffman, E., McCabe, K. and Smith, V. (1996a) 'Behavioural Foundations of Reciprocity: Experimental Economics and Evolutionary Psychology', Working paper, Department of Economics, University of Arizona, Tucson.

— (1996b) 'Social Distance and Other-Regarding Behavior in Dictator Games', *American Economic Review* 86 (June): 653–60.

Hollingshead, A. B., McGrath, J. E. and O'Connor, K. M. (1993) 'Group Task Performance and Communication Technology: A Longitudinal Study of Computer-Mediated versus Face-to-Face Work Groups', *Small Group Research* 24 (August): 307–33.

Holmstrom, B. (1982) 'Moral Hazard in Teams', *Bell Journal of Economics* 13 (Autumn): 324–40.

Homans, G. C. (1961) *Social Behavior: Its Elementary Forms,* New York: Harcourt, Brace, & World.

Isaac, R. M., McCue, K. and Plott, C. S. R. (1985) 'Public Goods Provision in an Experimental Environment', *Journal of Public Economics* 26 (February): 51–74.

Isaac, R. M. and Walker, J. (1988a) 'Communication and Free-Riding Behavior: The Voluntary Contribution Mechanism', *Economic Inquiry* 26 (October): 585–608.

— (1988b) 'Group Size Effects in Public Goods Provision: The Voluntary Contributions Mechanism', *Quarterly Journal of Economics* 103 (February) :179–99.

— (1991) 'Costly Communication: An Experiment in a Nested Public Goods Problem', in Thomas R. Palfrey (ed.) *Laboratory Research in Political Economy,* Ann Arbor: University of Michigan Press, pp. 269–86.

— (1993) 'Nash as an Organising Principle in the Voluntary Provision of Public Goods: Experimental Evidence', Working paper, Indiana University, Bloomington.

Isaac, R. M., Walker, J. and Thomas, S. (1984) 'Divergent Evidence on Free Riding: An Experimental Examination of Some Possible Explanations', *Public Choice* 43 (2): 113–49.

Isaac, R. M., Walker, J. and Williams, A. W. (1994) 'Group Size and the Voluntary Provision of Public Goods: Experimental Evidence Utilizing Large Groups', *Journal of Public Economics* 54 (May): 1–36.

Keohane, R. O. (1984) *After Hegemony*, Princeton, NJ: Princeton University Press.
Kikuchi, M., Watanabe, Y. and Yamagishi, T. (1996) 'Accuracy in the Prediction of Others' Trustworthiness and General Trust: An Experimental Study', *Japanese Journal of Experimental Social Psychology* 37 (1): 23–36.
Kim, O. and Walker, M. (1984) 'The Free Rider Problem: Experimental Evidence', *Public Choice* 43 (1): 3–24.
Knack, S. (1992) 'Civic Norms, Social Sanctions, and Voter Turnout', *Rationality and Society* 4 (April) :133–56.
Knight, J. (1992) *Institutions and Social Conflict,* Cambridge: Cambridge University Press.
Kollock, P.(1993) 'An Eye for an Eye Leaves Everyone Blind: Cooperation and Accounting Systems', *American Sociological Review* 58 (6): 768–86.
Kreps, D. M. (1990) 'Corporate Culture and Economic Theory', in J. E. Alt and K. A. Shepsle (eds) *Perspectives on Positive Political Economy,* New York: Cambridge University Press, pp. 90–143.
Kreps, D. M., Milgrom, P., Roberts, J. and Wilson, R. (19820 'Rational Cooperation in the Finitely Repeated Prisoner's Dilemma', *Journal of Economic Theory* 27 (August): 245–52.
Lam, W. F. (1998) *Institutions, Infrastructure, and Performance in the Governance and Management of Irrigation Systems: The Case of Nepal,* San Francisco, CA: Institute for Contemporary Studies Press. Forthcoming.
Ledyard, J. (1995) 'Public Goods: A Survey of Experimental Research', in J. Kagel and A. Roth (eds) *The Handbook of Experimental Economics,* Princeton, NJ: Princeton University Press, pp. 111–94.
Leibenstein, H. (1976) *Beyond Economic Man,* Cambridge, MA: Harvard University Press.
Levi, M. (1988) *Of Rule and Revenue*, Berkeley: University of California Press.
— (1997) *Consent, Dissent, and Patriotism*, New York: Cambridge University Press.
Lichbach, M. I. (1995) *The Rebel's Dilemma*, Ann Arbor: University of Michigan Press.
— (1996) *The Cooperator's Dilemma,* Ann Arbor: University of Michigan Press.
Luce, R. D. and Raiffa, H. (1957) *Games and Decisions: Introduction and Critical Survey,* New York: Wiley.
Marr, D. (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information,* San Francisco, CA: W. H. Freeman.
Marwell, G. and Ames, R. E. (1979) 'Experiments on the Provision of Public Goods I: Resources, Interest, Group Size, and the Free Rider Problem', *American Journal of Sociology* 84 (May): 1335–60.
— (1980) 'Experiments on the Provision of Public Goods II: Provision Points, Stakes, Experience and the Free Rider Problem', *American Journal of Sociology* 85 (January): 926–37.

— (1981) 'Economists Free Ride: Does Anyone Else?', *Journal of Public Economics* 15 (November): 295–310.
Marwell, G. and Oliver, P. (1993) *The Critical Mass in Collective Action: A Micro-Social Theory*, New York: Cambridge University Press.
McCabe, K., Rassenti, S. and Smith, V. (1996) 'Game Theory and Reciprocity in Some Extensive Form Bargaining Games', Working paper, Economic Science Laboratory, University of Arizona, Tucson.
McCay, B. J. and Acheson, J. M. (1987) *The Question of the Commons: The Culture and Ecology of Communal Resources,* Tucson: University of Arizona Press.
McKean, M. (1992) 'Success on the Commons: A Comparative Examination of Institutions for Common Property Resource Management', *Journal of Theoretical Politics* 4 (July): 247–82.
McKean, M. and Ostrom, E. (1995) 'Common Property Regimes in the Forest: Just a Relic from the Past?', *Unasylva* 46 (January): 3–15.
McKelvey, R. D. and Thomas P. (1992) 'An Experimental Study of the Centipede Game', *Econometrica* 60 (July): 803–36.
Messick, D. M. (1973) 'To Join or Not to Join: An Approach to the Unionization Decision', *Organisational Behavior and Human Performance* 10 (August): 146–56.
Messick, D. M. and Brewer, M. B. (1983) 'Solving Social Dilemmas: A Review', in L. Wheeler and P. Shaver (eds) *Annual Review of Personality and Social Psychology,* Beverly Hills, CA: Sage, pp. 11–44.
Messick, D. M., Wilke, H. A. M., Brewer, M. B., Kramer, R. M., Zemke, P. E. and Lui, L. (1983) 'Individual Adaptations and Structural Change as Solutions to Social Dilemmas', *Journal of Personality and Social Psychology* 44 (February): 294–309.
Milgrom, P. R., North, D. C. and Weingast, B. R. (1990) 'The Role of Institutions in the Revival of Trade: The Law Merchant, Private Judges, and the Champagne Fairs', *Economics and Politics* 2 (March): 1–23.
Miller, G. (1992) *Managerial Dilemmas: The Political Economy of Hierarchy,* New York: Cambridge University Press.
Moir, R. (1995) 'The Effects of Costly Monitoring and Sanctioning upon Common Property Resource Appropriation', Working paper, Department of Economics, University of New Brunswick, Saint John.
Morrow, C. E. and Hull, R. W. (1996) 'Donor-Initiated Common Pool Resource Institutions: The Case of the Yanesha Forestry Cooperative', *World Development* 24 (10): 164157.
Mueller, D. (1986) 'Rational Egoism versus Adaptive Egoism as Fundamental Postulate for a Descriptive Theory of Human Behavior', *Public Choice* 51 (1): 3–23.
Nowak, M. A. and Sigmund, K. (1993) 'A Strategy of Win-Stay, Lose-Shift that Outperforms Tit-for-Tat in the Prisoner's Dilemma Game', *Nature* 364 (July): 56–8.

Oakerson, R. J. (1993) 'Reciprocity: A Bottom-Up View of Political Development', in V. Ostrom, D. Feeny and H. Picht (eds) *Rethinking Institutional Analysis and Development: Issues, Alternatives, and Choices,* San Francisco, CA: Institute for Contemporary Studies Press, pp. 141–58.

Oliver, P. (1980) 'Rewards and Punishments as Selective Incentives for Collective Action: Theoretical Investigations', *American Journal of Sociology* 85 (May): 1356–75.

Olson, M. (1965) *The Logic of Collective Action: Public Goods and the Theory of Groups,* Cambridge, MA: Harvard University Press.

Orbell, J. M. and Dawes, R. M. (1991) 'A 'Cognitive Miser' Theory of Cooperators' Advantage', *American Political Science Review* 85 (June): 515–28.

— (1993) 'Social Welfare, Cooperators' Advantage, and the Option of Not Playing the Game', *American Sociological Review* 58 (December): 787–800.

Orbell, J. M., Dawes, R. M. and van de Kragt, A. (1990) 'The Limits of Multilateral Promising', *Ethics* 100 (April): 616–27.

Orbell, J. M., Schwartz-Shea, P. and Simmons, R. (1984) 'Do Cooperators Exit More Readily than Defectors?', *American Political Science Review* 78 (March): 147–62.

Orbell, J. M., van de Kragt, A. and Dawes, R. M. (1988) 'Explaining Discussion-Induced Cooperation', *Journal of Personality and Social Psychology* 54 (5): 811–9.

Ostrom, E. (1990) *Governing the Commons: The Evolution of Institutions for Collective Action,* New York: Cambridge University Press.

— (1998) 'Self-Governance of Common-Pool Resources', in P. Newman (ed.) *The New Palgrave Dictionary of Economics and the Law,* London: Macmillan.

Ostrom, E., Gardner, R. and Walker, J. (1992) 'Covenants with and without a Sword: Self-Governance Is Possible', *American Political Science Review* 86 (June): 404–17.

— (1994) *Rules, Games, and Common-Pool Resources,* A. Arbor: University of Michigan Press.

— (1997) 'Neither Markets Nor States: Linking Transformation Processes in Collective Action Arenas', in D. C. Mueller (ed.) *Perspectives on Public Choice: A Handbook,* Cambridge: Cambridge University Press, pp. 35–72.

Ostrom, V. (1980) 'Artisanship and Artifact', *Public Administration Review* 40 (July-August): 309–17.

— (1987) *The Political Theory of a Compound Republic: Designing the American Experiment,* 2nd rev. edn. San Francisco, CA: Institute for Contemporary Studies Press.

— (1990) 'Problems of Cognition as a Challenge to Policy Analysts and Democratic Societies', *Journal of Theoretical Politics* 2 (3): 243–62.

— (1997) *The Meaning of Democracy and the Vulnerability of Democracies: A Response to Tocqueville's Challenge,* Ann Arbor: University of Michigan Press.

Ostrom, V., Feeny, D. and Picht, H. (eds) (1993) *Rethinking Institutional Analysis and Development: Issues, Alternatives, and Choices*, San Francisco, CA: Institute for Contemporary Studies Press.

Palfrey, T. R. and Rosenthal, H. (1988) 'Private Incentives in Social Dilemmas', *Journal of Public Economics* 35 (April): 309–32.

Piaget, J. [1932] (1969) *The Moral Judgment of the Child,* New York: Free Press.

Pinker, S. (1994) *The Language Instinct*, New York: W. Morrow.

Pinkerton, E. (ed.) (1989) *Co-operative Management of Local Fisheries: New Directions for Improved Management and Community Development*, Vancouver: University of British Columbia Press.

Plott, C. R. (1979) 'The Application of Laboratory Experimental Methods to Public Choice', in C. S. Russell (ed.) *Collective Decision Making: Applications from Public Choice Theory,* Baltimore, MD: Johns Hopkins University Press, pp. 137–60.

Pruitt, D. G. and Kimmel, M. J. (1977) 'Twenty Years of Experimental Gaming: Critique, Synthesis, and Suggestions for the Future', *Annual Review of Psychology* 28:363–92.

Putnam, R. D., with R. Leonardi and R. Nanetti (1993) *Making Democracy Work: Civic Traditions in Modern Italy*, Princeton, NJ: Princeton University Press.

Rabin, M. (1994) 'Incorporating Behavioural Assumptions into Game Theory', in J. Friedman, *Problems of Coordination in Economic Activity,* Norwell, MA: Kluwer Academic Press.

Rapoport, A. (1997) 'Order of Play in Strategically Equivalent Games in Extensive Form', *International Journal of Game Theory* 26 (1): 113–36.

Rocco, E. and Warglien, M. (1995) 'Computer Mediated Communication and the Emergence of 'Electronic Opportunism', Working paper RCC#13659, Universita degli Studi di Venezia.

Roth, A. E. (1995) 'Bargaining Experiments', in *Handbook of Experimental Economics,* J. Kagel and A. E. Roth (eds) Princeton, NJ: Princeton University Press.

Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M. and Zamir, S. (1991) 'Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study', *American Economic Review,* 81 (December): 1068–95.

Rutte, C. G. and Wilke, H. A. M. (1984) 'Social Dilemmas and Leadership', *European Journal of Social Psychology* 14(January/March) :105–21.

Sally, D. (1995) 'Conservation and Cooperation in Social Dilemmas. A Meta-Analysis of Experiments from 1958 to 1992', *Rationality and Society* 7 (January): 58–92.

Samuelson, C. D. and Messick, D. M. (1986) 'Alternative Structural Solutions to Resource Dilemmas', *Organisational Behavior and Human Decision Processes* 37 (February): 139–55.

— (1995) 'When Do People Want to Change the Rules for Allocating Shared Resources', in D. A. Schroeder (ed.) *Social Dilemmas: Perspectives on Individuals and Groups,* Westport, CT: Praeger, pp. 143–62.

Samuelson, C. D., Messick, D. M., Rutte, C. G. and Wilke, H. A. M. (1984) 'Individual and Structural Solutions to Resource Dilemmas in Two Cultures', *Journal of Personality and Social Psychology* 47 (July): 94–104.

Samuelson, L., Gale, J. and Binmore, K. (1995) 'Learning to be Imperfect: The Ultimatum Game', *Games and Economic Behavior* 8 (January): 56–90.

Samuelson, P. A. (1954) 'The Pure Theory of Public Expenditure', *Review of Economics and Statistics* 36 (November): 387–9.

Sandler, T. (1992) *Collective Action: Theory and Applications*, Ann Arbor: University of Michigan Press.

Sato, K. (1987) 'Distribution of the Cost of Maintaining Common Property Resources', *Journal of Experimental Social Psychology* 23 (January): 19–31.

Satz, D. and Ferejohn, J. (1994) 'Rational Choice and Social Theory', *Journal of Philosophy* 91 (February): 71–82.

Scharpf, F. W. (1997) *Games Real Actors Play: Actor Centred Institutionalism in Policy Research,* Boulder, CO: Westview Press.

Schelling, T. C. (1978) *Micromotives & Macrobehaviour*, New York: W. W. Norton.

Schlager, E. (1990) 'Model Specification and Policy Analysis: The Governance of Coastal Fisheries', PhD diss., Indiana University.

Schlager, E. and Ostrom, E. (1993) 'Property-Rights Regimes and Coastal Fisheries: An Empirical Analysis', in R. Simmons and T. Anderson (eds) *The Political Economy of Customs and Culture: Informal Solutions to the Commons Problem,* Lanham, MD: Rowman & Littlefield, pp. 13–41.

Schneider, F. and Pommerehne, W. W. (1981) 'Free Riding and Collective Action: An Experiment in Public Microeconomics', *Quarterly Journal of Economics* 96 (November): 689–704.

Scholz, J. T. (1998) 'Trust, Taxes, and Compliance', in V. Braithwaite and M. Levi (eds) *Trust and Governance,* New York: Russell Sage Foundation.

Schroeder, D. A. (ed.) (1995) *Social Dilemmas: Perspectives on Individuals and Groups,* Westport, CT: Praeger.

Schuessler, R. (1989) 'Exit Threats and Cooperation Under Anonymity', *Journal of Conflict Resolution* 33 (December): 728–49.

Sell, J. and Wilson, R. (1991) 'Levels of Information and Contributions to Public Goods', *Social Forces* 70 (September): 107–24.

— (1992) 'Liar, Liar, Pants on Fire: Cheap Talk and Signalling in Repeated Public Goods Settings', Working paper, Department of Political Science, Rice University.

Selten, R. (1975) 'Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games', *International Journal of Game Theory* 4 (1): 25–55.

— (1986) 'Institutional Utilitarianism', in F.-X. Kaufmann, G. Majone, and V. Ostrom (eds) *Guidance, Control, and Evaluation in the Public Sector,* New York: de Gruyter, pp. 251–63.

— (1990) 'Bounded Rationality', *Journal of Institutional and Theoretical Economics* 146 (December) :649–58.

— (1991) 'Evolution, Learning, and Economic Behavior', *Games and Economic Behavior.* 3 (February): 3–24.

Selten, R., Mitzkewitz, M. and Uhlich, G. R. (1997) 'Duopoly Strategies Programmed by Experienced Players', *Econometrica* 65 (May) :517–55.

Sen, A. K. (1977) 'Rational Fools: A Critique of the Behavioural Foundations of Economic Theory', *Philosophy & Public Affairs* 6 (Summer): 317–44.

Sethi, R. and Somanathàn, E. (1996) 'The Evolution of Social Norms in Common Property Resource Use', *American Economic Review* 86 (September): 766–88.

Shepsle, K. A. and Weingast, Barry R. (1984) 'Legislative Politics and Budget Outcomes', in G. Mills and J. Palmer (eds) *Federal Budget Policy in the 1980s,* Washington, DC: Urban Institute Press, pp. 343–67.

Simon, H. A. (1985) 'Human Nature in Politics: The Dialogue of Psychology with Political Science', *American Political Science Review* 79 (June): 293–304.

— (1997) *Models of Bounded Rationality: Empirically Grounded Economic Reason,* Cambridge, MA: MIT Press.

Smith, V. (1982) 'Microeconomic Systems as an Experimental Science', *American Economic Review* 72 (December): 923–55.

Snidal, D. (1985) 'Coordination versus Prisoner's Dilemma: Implications for International Cooperation and Regimes', *American Political Science Review* 79 (December): 923–42.

Tang, S. Y. (1992) *Institutions and Collective Action: Self-Governance in Irrigation,* San Francisco, CA: Institute for Contemporary Studies Press.

Taylor, M. (1987) *The Possibility of Cooperation,* New York: Cambridge University Press.

Thibaut, J. W. and Kelley, H. H. (1959) *The Social Psychology of Groups*, New York: Wiley.

Tocqueville, A. de [1835 and 1840] (1945) *Democracy in America,* 2 vols. New York: Alfred A. Knopf.

Trivers, R. L. (1971) 'The Evolution of Reciprocal Altruism', *Quarterly Review of Biology* 46 (March): 35–57.

van de Kragt, A., Orbell, J. M. and Dawes, R. M. (1983) 'The Minimal Contributing Set as a Solution to Public Goods Problems', *American Political Science Review* 77 (March): 112–22.

Walker, J., Gardner, R., Herr, A. and Ostrom, E. (1997) 'Voting on Allocation Rules in a Commons: Predictive Theories and Experimental Results', Presented at the 1997 annual meeting of the Western Political Science Association, Tucson, Arizona, March 13–15.

Walker, J., Gardner, R. and Ostrom, E. (1990) 'Rent Dissipation in a Limited-Access Common-Pool Resource: Experimental Evidence', *Journal of Environmental Economics and Management* 19 (November): 203–11.

Williams, J. T., Collins, B. and Lichbach, M. I. (1997) 'The Origins of Credible Commitment to the Market', presented at the 1995 annual meeting of the American Political Science Association, Chicago, Illinois.

Yamagishi, T. (1986) 'The Provision of a Sanctioning System as a Public Good', *Journal of Personality and Social Psychology* 51 (1): 110–6.

— (1988a) 'Exit from the Group as an Individualistic Solution to the Free Rider Problem in the United States and Japan', *Journal of Experimental Social Psychology* 24 (6): 530–42.

— (1988b) 'The Provision of a Sanctioning System in the United States and Japan', *Social Psychology Quarterly* 51 (3): 265–71.

— (1988c) 'Seriousness of Social Dilemmas and the Provision of a Sanctioning System', *Social Psychology Quarterly* 51 (1): 32–42.

— (1992) 'Group Size and the Provision of a Sanctioning System in a Social Dilemma', in W. B. G. Liebrand, D. M. Messick, and H. A. M. Wilke (eds) *Social Dilemmas: Theoretical Issues and Research Findings,* Oxford, England: Pergamon Press, pp. 267–87.

Yamagishi, T. and Cook, K. S. (1993) 'Generalized Exchange and Social Dilemmas', *Social Psychological Quarterly* 56 (4): 235–48.

Yamagishi, T. and Hayashi, N. (1996) 'Selective Play: Social Embeddedness of Social Dilemmas', in W. B. G. Liebrand and D. M. Messick (eds) *Frontiers in Social Dilemmas Research,* Berlin: Springer-Verlag.

Yamagishi, T. and Takahashi, N. (1994) 'Evolution of Norms without Metanorms', in U. Schulz, W. Albers and U. Mueller (eds) *Social Dilemmas and Cooperation,* Berlin: Springer-Verlag, pp. 311–26.

Yoder, R. (1994) *Locally Managed Irrigation Systems,* Colombo, Sri Lanka: International Irrigation Management Institute.

Chapter Six

Beyond Markets and States: Polycentric Governance of Complex Economic Systems[1]

Elinor Ostrom[2]

Contemporary research on the outcomes of diverse institutional arrangements for governing common-pool resources (CPRs) and public goods at multiple scales builds on classical economic theory while developing new theory to explain phenomena that do not fit in a dichotomous world of 'the market' and 'the state'. Scholars are slowly shifting from positing simple systems to using more complex frameworks, theories, and models to understand the diversity of puzzles and problems facing humans interacting in contemporary societies. The humans we study have complex motivational structures and establish diverse private-for-profit, governmental, and community institutional arrangements that operate at multiple scales to generate productive and innovative as well as destructive and perverse outcomes (North 1990, 2005).

In this chapter, I will describe the intellectual journey that I have taken the last half century from when I began graduate studies in the late 1950s. The early efforts to understand the poly-centric water industry in California were formative for me. In addition to working with Vincent Ostrom and Charles M. Tiebout as they formulated the concept of polycentric systems for governing metropolitan areas, I studied the efforts of a large group of private and public water producers facing the problem of an overdrafted groundwater basin on the coast and watching saltwater intrusion threaten the possibility of long term use. Then, in the 1970s, I participated with colleagues in the study of polycentric police industries serving US metropolitan areas to find that the dominant theory underlying massive reform

1. This chapter is a revised version of the lecture Elinor Ostrom delivered in Stockholm, Sweden, on December 8, 2009, when she received the Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel. This article is copyright © The Nobel Foundation 2009 and is published with the permission of the Nobel Foundation. Published initially in *The American Economic Review*, 100(3), 2010, pp. 641–672.

2. I wish to thank Vincent Ostrom and my many colleagues at the Workshop who have worked with me throughout the years to develop the research program that is briefly discussed herein. I appreciate the helpful suggestions given me by Arun Agrawal, Andreas Leibbrandt, Mike McGinnis, Jimmy Walker, Tom Wisdom, and by the Applied Theory Working Group and the Experimental Reading Group, and the excellent editing skills of Patty Lezotte. Essential support received over the years from the Ford Foundation, the MacArthur Foundation, and the National Science Foundation is gratefully acknowledged.