

5

Cooperative *Homo economicus*

It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest.

Adam Smith, *Wealth of Nations* (1937[1776]), Book 1, Chapter 2

5.1 Introduction

In contrast to biology where cooperative behaviors have become a central research focus only in recent decades, a major goal of economic theory since its inception has been to show the plausibility of wide-scale cooperation among self-interested individuals. Since the middle of the twentieth century this endeavor involved the development of the so-called Walrasian model of economy-wide competitive exchange and the affirmation of Adam Smith's 'invisible hand' conjecture. The success of this endeavor culminated in the Fundamental Theorem of Welfare Economics (Arrow and Debreu 1954, Debreu 1959, Arrow and Hahn 1971), sustaining Smith's insight that self-interested behaviors might support socially valued economic outcomes. But, the theorem's essential assumption that all relevant aspects of all exchanges could be completely specified in contracts enforceable at zero cost to the exchanging parties is widely recognized as not applicable to any real world economy (Arrow 1971, Bowles and Gintis 1993, Gintis 2002, Bowles 2004).

A second major thrust of economic theory eschewed this implausible assumption and developed sophisticated repeated game models in which the outcomes of exchanges are determined by bargaining, collusion, and other forms of strategic interaction. These models refine and extend the insights of Shubik (1959), Trivers (1971), Taylor (1976), and Axelrod and Hamilton (1981) that retaliation against defectors by withdrawal of cooperation may enforce cooperation among self-regarding individuals. This literature culminates in the "folk theorems" of Fudenberg, Levine, Maskin, and others (Fudenberg and Maskin 1986, Fudenberg, Levine and Maskin 1994). A great virtue of these models, in contrast to the Walrasian paradigm in economics,

is that in recognizing the ubiquity of incomplete or unenforceable contracts, they describe the real world of interactions among most animals, including humans (Blau 1964, Gintis 1976, Stiglitz 1987, Tirole 1988, Laffont 2000, Bowles and Hammerstein 2003).

The folk theorems were not developed for the purpose of evolutionary explanation and have not been extensively used in this way. The most ambitious statement of this line of thinking, applied towards understanding the broad historical and anthropological sweep of human experience, is the work of Ken Binmore (1993, 1998), culminating in his *Natural Justice* (2005). This work offers an evolutionary approach to morality, in which moral rules form a cultural system that developed historically with the emergence of *Homo sapiens*. For Binmore, a society's moral rules are instructions for behavior in conformity with one of the myriad of Nash equilibria of a repeated n -player social interaction. Because the interactions are repeated, the self-regarding individuals who comprise the social order will conform to the moral rules of their society as a type of self-fulfilling prophecy (if all other individuals play their part in this Nash equilibrium, an individual has no incentive to deviate from playing his part as well).

Binmore's solution is one of a broad class of models developed to explain cooperation among self-regarding individuals as a result of repeated interactions. In this chapter, we will show that while the insight that repeated interactions provide opportunities for cooperative individuals to discipline defectors is correct, none of these models is successful. The reason is that even presupposing extraordinary cognitive capacities and levels of patience among the cooperating individuals, it is unlikely that a group of size greater than two would ever discover the cooperative equilibria that the models have identified, and almost certainly, if it were to hit on one, its members would abandon it in short order.

5.2 Folk Theorems and Evolutionary Dynamics

The folk theorem is based on a *stage game* played an indefinite number of times, with a constant, strictly positive, probability that in each period the game will continue for an additional period (§A2). The restrictions on the stage game tend to be minimal and rather technical (Fudenberg and Maskin 1986, Fudenberg et al. 1994). Player strategies in the repeated game are conditioned on the pattern of behavior, usually interpreted as cooperation and defection, in previous periods. The information concerning this pattern

of behavior is a signal that may be *perfect* (completely accurate and received by all individuals) or *imperfect* (inaccurate with positive probability and/or received only by a subset of individuals). An imperfect signal can be *public* (all players receive the same signal) or *private* (different players receive different signals, and some may receive no signal at all). As in the previous chapter, we may think of imperfect public signals as caused by *execution errors*, which are then seen by all other players, while private signals are caused by *limited scope*, in which players observe the behaviors of only a subset of their group members, or *perceptual error*, in which specific individuals incorrectly interpret cooperation as defection, or *vice-versa*. However execution errors can be private because they are seen only by a subset of individuals, and hence private signals need not involve perceptual error.

With either perfect or public imperfect signals, a folk theorem can be proved, asserting that any feasible allocation of payoffs to the players that dominates the minimax payoff of each player can be achieved, or approximated as closely as desired, as the equilibrium per-period payoff to the repeated game, for some discount factor strictly less than unity (Fudenberg and Maskin 1986, Fudenberg et al. 1994). A similar folk theorem can be proved for certain types of private signals. Significant contributions to this literature include Sekiguchi (1997), Piccione (2002), Ely and Välimäki (2002), Bhaskar and Obara (2002), and Mailath and Morris (2006). Private signaling creates especially grave problems. First, the sequential equilibrium requires strictly mixed strategies on the part of all players in all periods. Yet, individuals have no incentive to play these mixed strategies. Second, the equilibria require that private signals be sufficiently close to being public, so all individuals receive nearly the same signal concerning the behavior of any given group member. When this is not the case, the equilibrium will not exist. Thus, these models apply only to forms of cooperation where all members observe the actions of (nearly) all others with a high level of accuracy. The equilibrium concept employed is that of sequential equilibrium, which is a Nash equilibrium in which players choose best response and use Bayesian updating of beliefs at all information sets, whether on or off the path of play (Kreps and Wilson, 1982).

In Section 5.3 we show that repeated game models and their associated folk theorems are merely a first step in understanding social cooperation. Proving the existence of a sequential Nash equilibrium must be followed by an analysis of the dynamical out-of-equilibrium behavior of the system, with

the goal of showing that the equilibrium is asymptotically stable (i.e., has a basin of attraction) and the system is highly likely eventually to enter the basin of attraction of the equilibrium and remain there. Unless this can be shown, we say that the result is an “dynamically irrelevant Nash equilibrium”

Recent advances in the epistemological foundations of equilibrium concepts in game theory provide a possible way forward in dealing with out-of-equilibrium dynamics (Aumann and Brandenburger 1995). Common knowledge of rationality and even common priors do not ensure that player beliefs are sufficiently aligned to produce Nash equilibrium in all but the smallest and simplest games. Nor does game theory provide an explanation of how individual beliefs can be aligned in a manner allowing a group to coordinate on the kinds of complex behaviors required by the folk theorems.

However, sociologists (Durkheim 1933[1902], Parsons and Shils 1951) and anthropologists (Benedict 1934, Boyd and Richerson 1985, Brown 1991) have found that virtually every society has such processes, and that they are key to understanding strategic interaction. Borrowing a page from sociological theory, we posit that groups have *focal rules* specifying how a game ought to be played and that these rules are identified as social norms by group members. Learning a focal rule includes learning that the rule is common knowledge among those who know it, learning what behavior is suggested by the rule, and learning that a large fraction of group members know the rule and follow it.

Focal rules do not ensure equilibrium, because error, mutation, migration, and other dynamical forces may lead individuals to reject beliefs or behavior fostered by the rule, because the focal beliefs might conflict with an individual’s personal experience, or its suggested behavior may be rejected as not it the individual’s best interest; i.e., the action fostered by a focal rule must be a best response to the behaviors of the other group members, given the beliefs engendered by the focal rule and the individual’s Bayesian updating. Moreover, focal rules cannot be introduced as a *deus ex machina*, as if laid down by a centralized authority, without violating the objective to provide a “bottom up” theory of cooperation that does not presuppose preexisting institutional forms of cooperation. Focal rules are thus discretionary, because any institution that is posited to enforce behavior should itself be modeled within the dynamical system, unless plausible reasons are given for taking a macro-level institution as unproblematically given. Nor are focal rules fixed in stone. A group’s focal rules must themselves be subject to change, those groups producing better outcomes for their members displacing groups with

less effective focal rules, and changing social and demographic conditions leading to the evolutionary transformation of focal rules within groups.

The term “focal rule” generalizes Schelling (1960), who introduced the notion of a *focal point* as an informal process of attaching meanings to strategies and using the notion of salience to align the choices of individuals in a coordination game (Sugden 1995, Binmore and Samuelson 2006). Friends, for example, may share as a focal point a common idea of where to find each other should they become separated while shopping. The idea of focal *rules* is akin to Binmore’s notion of moral rules that choose among Nash equilibria, except that focal rules facilitate the attainment of a Nash equilibrium by appropriately aligning beliefs, rather than selecting among Nash equilibria. Employing the terminology of interactive epistemology (Aumann and Brandenburger 1995), a focal rule leads agents to alter their Bayesian priors, and generates a correlated equilibrium with the potential to coordinate cooperative activity and provide incentives for individuals to play their part in this activity.

Beginning with Section 5.4, we restrict consideration to the n -player public goods game (§4.4), which is the appropriate model for many social dilemmas in which contemporary humans exhibit a high level of cooperation, including team production, voting, common pool resource management, and collective action, as well as in common defense, information sharing, and hunting in Pleistocene ancestral communities.

Fudenberg et al. (1994) proved the folk theorem for stage games with imperfect public signals. Gintis (2007) has shown that their argument applies to the public goods game, deriving expressions linking the error rate ϵ , the group size n , and the discount factor δ , and showed that for any given discount factor δ , there is a maximum $n\epsilon$, order of magnitude unity, that supports cooperation (for a discussion of the discount factor, see the analysis surrounding equation 4.6). This means that cooperation may be sustained in groups experiencing less than one error per period, and otherwise not. Thus, either large groups or large error rates are incompatible with cooperation, despite the folk theorem, unless the discount factor is permitted to approach arbitrarily close to unity, which is ruled out by such demographic realities as the probability of mortality, as well as subjective time preference.

In Section 5.4 we use an agent-based simulation (§A1) to the public goods game with imperfect public signaling to show that without focal rules, a high level of cooperation can be attained only with very small group size ($n \leq 4$) or near zero error rates. When we introduce focal rules reflecting

the game-theoretic strategy of punishing defections but ignoring defections by others for whom defecting is a punishment of a third party, we can attain quite high levels of cooperation as long as we do not allow the focal rules themselves to evolve. However, when focal rules are subject to competitive pressure, they collapse, leading to the exceedingly low levels of cooperation characteristic of models without focal rules.

The reason for this unraveling is straightforward. When the error rate is low, the optimal focal rule is to tolerate zero defections. However, when all groups follow this focal rule, a group that tolerates a single defection has higher average payoff than the zero-tolerance groups. Thus, by adopting a rule that is very intolerant of defections, a group is providing a public good to the rest of the population at a cost to itself. Hence, other groups copy the less stringent focal rule until all allow a single defector. But, in this situation, a group that tolerates two defectors has higher average payoff than the groups that tolerate a single defector, so tolerating two defectors eventually becomes the universal focal rule. At some point a within-group selection process takes over: there are now so many defectors that individuals who ignore the focal rules altogether and merely tolerate zero defectors have higher payoffs than group members who conform to focal rules. Because defections are now present in all groups these zero-tolerance individuals defect at a high rate. This leads quickly to the abandoning of the focal rule and hence the unraveling of cooperation.

This exercise shows both the value of the focal rule approach, and the weakness of the repeated game solution in the context of the public goods game in an evolutionary setting. Our negative assessment of the folk theorem is due in part to the particular game we have studied. There is a serious problem with the public goods game as a model of cooperation: the only incentive mechanism is the threat of withdrawal of cooperation in response to an observed defection. If the public good in question is a vital service to the group, the idea of executing a coordinated failure to provide the service in response to an infraction is implausible. This form of punishment cannot be directed at the miscreant, but rather is shared by all. Thus, in large groups, or groups with imperfect signals, the efficiency costs of incentives can completely offset any gains from cooperation. For instance, (a) fishers cooperating to maintain a common pool resource cannot possibly respond to overfishing on the part of one member by all members' intentionally overfishing; (b) a band of hunters cannot respond to an observed shirking

incident by shirking in response; (c) in time of conflict, warriors are unlikely to punish cowards within their midst by refusing to fight.

Section 5.5 suggests that a general alternative in such cases is to use *directed punishment*, whereby a miscreant must pay a fine a cost to the individual imposing the fine. However, implementing a truly decentralized directed punishment mechanism is challenging. If punishment is costly, self-regarding individuals must have adequate incentives for carrying it out, and the signaling mechanism must include information on punishing activity. If punishing is rewarding to the punisher (e.g., failing to help a miscreant in need, as in the indirect reciprocity models of the previous chapter), then we must have a mechanism that limits punishing acts to miscreants alone. This problem is difficult when signals are public, but approaches being insurmountable when signals are private, so that a punishing act that is justified according to one observer may not be justified according to another. In short, directed punishment merely shifts the problem of cooperation from the stage game to the directed punishment game.

Altruistic punishment, which we discuss in the next chapter, restores the efficacy of the public goods game by dropping the requirement that the payoff to punishing must be sufficient to motivate punishing behavior on the part of self-regarding individuals. Not surprisingly, then, the public goods game induces cooperation only when accompanied by altruistic participants.

5.3 Dynamically Irrelevant Equilibria

John Nash (1950) developed the equilibrium concept that bears his name (§A2), the idea was elaborated upon by Kuhn (1953) and promoted in the influential volume by Luce and Raiffa (1957). John Harsanyi (1967) extended the notion of Nash equilibrium to games of incomplete information, and Reinhard Selten (1975) offered the first, and most important so-called equilibrium refinement, the *subgame perfect* equilibrium, which ruled out Nash equilibria involving incredible threats; i.e., actions that a player registers the intention of choosing, but are not best responses, so will not in fact be chosen should the occasion arise. There followed a decade of research into equilibrium refinements, including the well-known sequential equilibrium that we have already encountered. Jean Tirole (1990) documented a revolution in the economic field of Industrial Organization accomplished by searching for appropriately Nash equilibrium refinements. By the time Kreps (1990) and Fudenberg and Tirole (1991) published their influential

textbooks, it had become accepted wisdom that “solving” a game meant finding its subgame perfect or sequential equilibria.

But, these ‘solutions’ do not explain human behavior because there is no reason to believe that individuals would ever adopt the behaviors making up the equilibrium, or if they did, would not swiftly abandon them. While there are conditions under which individuals can “learn” to play a Nash equilibrium (Fudenberg and Levine 1997, Young 2006), these conditions do not obtain for repeated games, which are much more complex entities than their stage games. The strongest arguments in favor of the assertion that a Nash equilibrium will be played come from evolutionary game theory, where it is shown that every stable equilibrium of a dynamical system governed by a monotone dynamic, such as the replicator dynamic (Taylor and Jonker 1978), is a Nash equilibrium of the underlying game (Nachbar 1990, Samuelson and Zhang 1992). However, if there are multiple equilibria, as in the case in repeated game theory, this argument does not yield any prediction about behavior and does not imply that a high level of cooperation will occur even if full cooperation equilibria exist. Indeed, this argument does not even imply that an evolutionary system has a stable equilibrium, the alternative being a limit cycle or other non-equilibrium behavior.

It is surprising how little can be said about strategic interaction even assuming that individuals are predisposed and able to best respond given full knowledge of the game. We have known since Pearce (1984) and Bernheim (1984) that the assumption that it is common knowledge that individuals maximize their payoffs implies only that players will use only strategies that survive the iterated elimination of strictly dominated strategies, and there are many examples of games that cast doubt on the adequacy of even this assumption (Milgrom and Roberts 1990, Carlsson and van Damme 1993, Basu 1994, Vives 2005). Moreover, the equilibria in the repeated game models of cooperation are not achievable by the iterated elimination of strictly dominated strategies,

Recent research in interactive epistemology suggests that the conditions for achieving Nash equilibrium are quite stringent and rarely satisfied, except in the simplest of cases (Aumann and Brandenburger 1995). The problem with achieving a Nash equilibrium is that individuals may have heterogeneous and incompatible *beliefs* concerning how other players will behave, and indeed what other players believe concerning one’s own behavior. It is clear from this research that the epistemological requirements for Nash equilibrium in all but the simplest games cannot be deduced from the as-

sumption of rationality alone. This is because when there are multiple Nash equilibria, even the assumption that other players will choose a Nash strategy (an assumption that is itself difficult to justify) is insufficient to ensure a Nash equilibrium. Rather, there must be a social process leading to the alignment of conjectures and the constitution of common priors. This idea has a long history in the context of pure coordination games (Lewis 1969), but it in fact applies quite generally (Aumann 1987).

5.4 Focal Rules in the Public Goods Game

The concept of focal rules provides at least a partial solution. A cooperative equilibrium with focal rules as one in which not only is the equilibrium strategy mix evolutionarily stable, but focal rules are themselves an evolutionary adaptation, stable against invasion by competing focal rules. To see this, we return to the agent-based model of the public goods game developed in §4.4.

To implement the concept of focal rules, suppose each group G suggests a minimum cooperate level t_g , such that members cooperate as long as at least t_g members cooperated on the previous round. We assume that the group is in one of the states {Cooperate,Punish}. The game starts in state Cooperate, and remains there until fewer than t_g members signal Cooperate, whereupon the state Punish is entered. In state Punish, the focal rule is for the Defectors to Cooperate and all others to Defect. If the Defector succeed in cooperating, the system moves back into state Cooperate, where everyone cooperates, and otherwise remains in state Punish.

Suppose each individual is either a “Conformist” or an “Independent.” A Conformist follows the focal minimum cooperate level t_g of whatever group he is in, while the Independent follows his own strategy, given by his being an t -Cooperator for a given t . We begin by assuming that all groups have the same t_g , which is exogenously given and not subject to change during the simulation. The fraction of Conformists in the population and the fraction of each type of t -Cooperator among the Independents, however, evolve endogenously according to a payoff-based replicator dynamic. The results are exhibited in Figure 5.1 for $t_g = n$. Results are similar for values of t_g as small as 70% of group size. For smaller values of t_g , the fraction of Conformists declines to low levels and little cooperation can be sustained. In general, if a group’s focus does not provide the proper incentives, Independents will have higher payoffs than Conformists, and the focal rule

will be abandoned as the frequency of Independents increases and that of Conformists decreases.

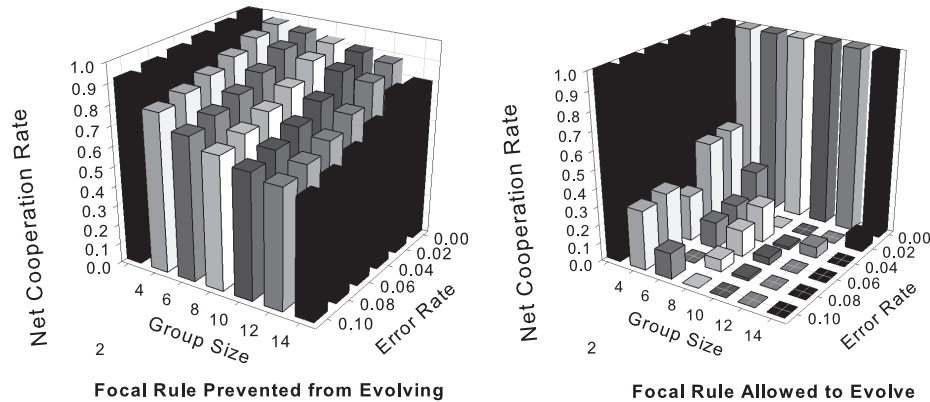


Figure 5.1. The Rate of Cooperation in the Public Goods Game for various Group Sizes and Error Rates, $b/c = 2$. The left pane assumes an exogenously fixed focal rule $t_g = n$. The right pane assumes that the focal rule is subject to evolutionary pressures.

It is clear from the left pane in Figure 5.1 that the addition of an appropriate focal rule entails high efficiency of cooperation, attaining 70% even for groups of size 14 with a 10% error rate. However, by maintaining a stringent defection threshold, a group is bearing a cost to provide a benefit to the rest of the population, because defectors punished as a result of the stringent threshold will then have low fitness, and hence will tend to be replaced by Cooperators, thus benefitting all members of the population. Thus, members of each group will do better if their group raises the number of defections it permits and if other groups maintain zero or low tolerance for defectors. If the defection rate that each group permits before triggering retaliation is permitted to evolve, it will therefore fall. The results of this dynamic are depicted in the right pane of Figure 5.1.

5.5 Directed Punishment

The analysis to this point suggests that even when we have the luxury of public signaling, the inability to direct punishment explicitly to the offending party renders the repeated game model of cooperation an infeasible or

inefficient instrument in many cases. The obvious alternative is to allow some form of punishment directed specifically at the miscreant.

Suppose defectors can be identified, and are punished an amount p by other members of the group. We must have $p \geq c$ for punishment to deter defection, which means each other group members must punish an offender an amount at least $c/(n-1)$. If there are execution errors occurring with probability ϵ , then in a fully cooperative equilibrium each member each period will punish others an expected amount $\epsilon(n-1)c/(n-1) = \epsilon c$, and will himself receive an amount of punishment $\epsilon p = \epsilon c$ if punishment is set to its minimum effective level $p = c$. Suppose the cost c_p of meting out punishment p is αp where $-1 < \alpha$ (i.e., the cost may be negative, providing an incentive to punish). Then, assuming full cooperation, the cost of punishing and being punished per individual is $\epsilon c(1 + \alpha)$, and the net payoff per period is

$$b(1 - \epsilon) - c(1 + \epsilon(1 + \alpha)). \quad (5.1)$$

Note that the cost of punishing is just $\epsilon(1 + \alpha)$ per period per group member, which is independent of group size. Thus, directed punishment appears to solve the problem of cooperation in large groups. Moreover, there is nothing in principle preventing $-1 < \alpha < 0$, so the directed punishment solution has the potential of being extremely efficient.

There is a catch, however: if $\alpha > 0$, players have no incentive to carry out the punishment, and if $\alpha < 0$, players have no incentive to limit their extractions to shirkers. Thus, an equilibrium of this type cannot be sustained by self-regarding individuals. To create incentives for individuals to punish defectors in the $\alpha > 0$ case, suppose members agree that any individual who is detected not punishing a defector is himself subject to punishment by the other players. Suppose with probability ϵ an individual who intends to punish fails to do so, or is perceived publicly by the other members to have failed. For simplicity, we choose ϵ to be the same as the error rate of cooperation. If all individuals cooperate and punish, the number of observed defections will be ϵn . Suppose all members must punish defectors equally. Then, the mean number of punishment of defector events per period will be ϵn^2 . But, of course, $\epsilon^2 n^2$ of these events will erroneously be viewed as non-punishing, so we must have $\epsilon^2 n^3$ punishing of non-punisher events (let us call this *second order punishment*). Similarly, we must have $\epsilon^3 n^4$ third-order punishment to enforce second-order punishment. Assuming we have punishment on all levels, the total amount of punishment per individual

per period will be

$$\epsilon n(1 + \epsilon n + \epsilon^2 n^2 + \dots) = \frac{\epsilon n}{1 - \epsilon n},$$

provided $\epsilon < 1/n$. If the reverse inequality holds, this mechanism cannot work because each order of punishment involves greater numbers than the previous. Assuming $\epsilon < 1/n$, the expected payoff to a Cooperator under conditions of complete cooperation (assuming one engages in one's own punishment) is given by the recursion equation

$$v = b(1 - \epsilon) - c - \epsilon n \frac{p + c_p}{1 - \epsilon n} + \delta v,$$

so

$$v(1 - \delta) = (b(1 - \epsilon) - c)(1 - \epsilon n) - \epsilon n(p + c_p). \quad (5.2)$$

which becomes negative when ϵ is sufficiently close to $1/n$. Thus, cooperation will not be sustainable for large n unless error rates are quite low.

5.6 Conclusion: The Missing Choreographer

The economic theory of cooperation based on repeated games proves the existence of equilibria with socially desirable properties, while leaving the question of how such equilibria are achieved as an afterthought, exhibiting a curious lack of attention to out-of equilibrium behavior. The folk theorem shares this defect with the even more celebrated Fundamental (“Invisible Hand”) theorem mentioned at the outset of the chapter. It purports to model decentralized interactions, but on close inspection requires a level of coordination that is not explained, but rather posited as a *deus ex machina*. We have shown that the focal rules on which the coordination must be based will not evolve spontaneously. Yet in these models, highly choreographed coordination on complex strategies capable of deterring defection are supposed to materialize quite without the need for a choreographer.

The failure of these models is hardly surprising, for the task we set for them, that of explaining the emergence and persistence of cooperation among large numbers of self regarding strangers without recourse to pre-existing cooperative institutions, is not only formidable, it most likely never occurred in the history of our species. Humans are indeed unique among living creatures in the degree and range of cooperation among large numbers of substantially

unrelated individuals. The global division of labor and exchange, the modern democratic welfare state, and contemporary warfare alike evidence our distinctiveness. These forms of cooperation emerged historically and are today sustained as a result of the interplay of self-regarding and social preferences operating under the influence of group-level institutions of governance and socialization that favor Cooperators, in part by helping to coordinate their actions so as to target transgressions for punishment and thus protecting them from exploitation by defectors.

The norms and institutions that have accomplished this evolved over millennia through trial and error. Consider how real world institutions addressed two of the shoals on which the economic models foundered. First, the private nature of information, as we have seen, makes it virtually impossible to coordinate the targeted punishment of miscreants. In many small-scale societies this problem is attenuated by such cooperative customs as eating in public so that violations of sharing rules can be easily detected. Cooperative shrimp fishermen in Japan deliberately land their catch at the same time of day for the same reason (Platteau and Seki 2001). But, where larger numbers are involved, converting private information about transgressions to public information that can provide the basis of punishment often involves civil or criminal trials, elaborate processes that rely on commonly agreed upon rules of evidence and ethical norms of appropriate behavior. Even these complex institutions frequently fail to transform the private protestations of innocence and guilt into common knowledge. Second, here and in the previous two chapters we have seen that cooperation often unravels when the withdrawal of cooperation by the civic-minded intending to punish a defector is mistaken by others as itself a violation of a cooperative norm, inviting a spiral of further defections. In virtually all surviving societies, this problem was addressed by the creation of a corps of specialists entrusted with carrying out the more severe of society's punishments. Their uniforms conveyed the civic purpose of the punishments they meted out, and their professional norms, it was hoped, ensured that the power to punish was not used for personal gain. Like court proceedings this institution works imperfectly.

Modeling the complex process by which we became a cooperative species is a major challenge of contemporary science. Economic theory, favoring parsimony over realism, has instead sought to explain cooperation without reference to other-regarding preferences and with a minimalist or fictive description of social institutions. This research trajectory, as we have seen, has produced significant insights. But it may have run its course.

The challenge thus shifts from that favored by biologists and economists over the last half century—showing why self-interested individuals would nonetheless cooperate—to explaining how the other-regarding preferences and group-level institutions that sustain cooperation could have emerged and proliferated in an empirically plausible evolutionary setting. The evolution of a predisposition to punish antisocial behavior even at substantial personal cost is emblematic of this process.