Chapter 7
Bringing Back Leviathan into Social Dilemmas

Mizuho Shinada and Toshio Yamagishi

Introduction

More than three and a half centuries ago, the great ancestor of the contemporary social dilemma researchers, Thomas Hobbes, published the most influential book on social dilemmas ever (Hobbes, 1651) and argued that social order cannot be maintained without authority that controls individuals’ unrestricted pursuit of self-interest. According to Hobbes, people who prefer peace (mutual cooperation) to a war of all against all (mutual defection) should agree to give birth to Leviathan, or a government or authority that enforces social order.

In “the tragedy of commons,” Garrett Hardin (1968) talks about the modern version of the Hobbesian problem of social order. In this influential article, Hardin uses a parable of English commons to illustrate how and why communities fail to maintain their resources. Consider the pasture that is open to all herdsmen in a village. Suppose there is no rule for regulating the use of the pasture or social institutions that enforce such rules. Each herdsman can freely add his sheep on the common pasture. The increase in the demand for wool in fact induced English villagers to add more sheep than the pasture could sustain. This “rational” behavior of each herdsman produced depletion of grasses on the commons. This familiar parable drew the attention of the general public in the 1970s when people became aware of resource problems on a global scale. Hardin’s (1968) recommendation for the prevention of the “tragedy” echoes that of Hobbes’; he recommended that coercion—“coercion, mutually agreed upon by the majority of the people affected” (Hardin, 1968, p. 1247)—not conscience, provides a solution to the tragedy. He even stated that “a call for voluntary compliance would be counterproductive” (Hardin, 1977, p. 129).

Hardin’s recommendation for coercion as the solution to the social dilemma problem created much controversy in various fields, inviting criticism because of its brutal Hobbesian appearance (Crowe, 1969; Fox, 1985; Lynn & Oldenquist, 1986; Stillman, 1975; Taylor, 1976, 1982). Subsequent studies of social dilemmas in psychology mostly ignored the recommendation—mutual coercion mutually agreed upon—by the two giants of social dilemma research, and mostly focused on how and among whom “voluntary compliance” emerges (for reviews of psychological studies of social dilemmas, see Dawes, 1980; Kollock, 1998; Messick & Brewer, 1983; Stroebe &
Frey, 1982), except for limited and sporadic attempts to pursue psychological implications of their recommendations (e.g., Yamagishi, 1988a, b). Meanwhile, the two giants’ recommendations have met much warmer greetings in economics. Another earlier giant in social dilemma research, Mancur Olson (1965), emphasized selective incentives as a solution to the public-good problem even before Hardin’s (1968) influential article, though not as early as Hobbes’. Especially during the last decade, experimental economists have revitalized social dilemma researchers’ interest in “coercion” as a solution to the social dilemma problem. The purpose of this chapter is to provide a review of the studies and findings on the use and effectiveness of sanctions conducted in recent years mostly by experimental economists, and add a few suggestions for future studies from social psychological perspectives.

Experimental research on social dilemmas in economics has grown rapidly for the past couple of decades. Economics was once regarded as a non-experimental science like astronomy (Samuelson & Nordhaus, 1985) when Hardin’s article and Olson’s book were published; it was taken for granted that economic activities were so complicated that manipulating theoretically relevant factors while controlling for relevant other factors was impossible. However, this view has changed drastically between 1948, when the first experiment in economics was conducted to test the neoclassical theory of perfect competition (Chamberlin, 1948), and 2002, when Vernon Smith, who participated in Chamberlin’s experiment as a graduate student, received a Nobel Prize in economics. In the interim, experimental economists have produced experimental evidence against the assumption of homo economicus, who cares only for his or her own self-interest. For example, the contribution rate for public good is persistently higher than the zero predicted by such a self-interested model of humans (Isaac et al., 1984; see Ledyard, 1995; Rabin, 1998 for reviews).

The discrepancy between the traditional theory and experimental results has driven several economists to develop a theory of “social preferences” that incorporates preferences for other-regarding behavior as well as self-interested behavior (Camerer, 2003; Fehr & Schmidt, 1999; Levine, 1998; Rabin, 1993). Similar to their counterpart in social psychology—researchers of social value orientation—they explain individual differences in social behavior in terms of differences in social preferences. Faced with experimental evidence indicating other-regarding social preferences including preferences for fairness and reciprocity, the main concern of rationally minded economists has increasingly focused on the development of theories to explain why humans have such other-regarding social preferences. For them (as well as for the authors of this chapter), explaining social preferences means finding adaptive advantages for acquiring such preferences. The question is, under what socio-institutional arrangement does the behavior guided by such preferences bring in self-interest (or interest to one’s kin who shares one’s gene)? One promising theoretical direction in this endeavor of explaining other-regarding preference or behavioral pattern has been provided by studies of indirect reciprocity (Brandt & Sigmund, 2005; Nowak & Sigmund, 1998; Takahashi & Mashima, 2003). In a generalized exchange system, being conditionally altruistic toward other conditional altruists makes the conditional altruist more successful than egoists in procuring resources. Another promising theoretical development is in the direction of “strong reciprocity” (Fehr & Fischbacher, 2003;
Gintis, 2000), which includes sanctioning behavior. The adaptive advantage of having a preference for strong reciprocity and sanctioning against free riders (often called the altruistic punishment or third-party punishment; Fehr & Fischbacher, 2004; Fehr & Gächter, 2002) has been proposed based on some evolutionary mechanisms such as cultural transmission and cultural group selection (Fehr & Fischbacher, 2003; Fehr et al., 2002). Our goal in this chapter is to give a sample of research on sanctioning behavior mostly conducted during the last decade by experimental economists. The rest of this chapter is organized into three sections. The first section provides a brief review of research findings about solving social dilemmas in psychology, especially focusing on “structural solutions” (Messick & Brewer, 1983). In the second section, we will look at some of the recent studies of sanctions, most of which are conducted by experimental economists. In the third section, we will propose possible contributions that psychologists can provide to the study of sanctions in social dilemmas.

The Structural Approach to Solving Social Dilemmas

According to Messick and Brewer (1983), solutions to the social dilemma problem are more or less classified as either individual solutions or structural solutions. Individual solutions are the ones that alter individual cognition and motivation to make them more willing to engage in cooperative behavior. Researchers who are interested in individual solutions study factors that affect individuals’ behavior in social dilemmas such as communication (Bixenstine & Douglas, 1967; Bixenstine et al., 1966; Rapoport et al., 1962), information about others’ choices (Dawes et al., 1977), trust in other group members (Messick et al., 1983), social values and responsibility (McClintock & Liebrand, 1988; Sweeney, 1973; Van Lange et al., 1997; Van Lange & Kuhlman, 1994), ingroup identity (Kramer & Brewer, 1984, 1986), self-monitoring (Boone et al., 1999; Danheiser & Granziano, 1982; De Cremer et al., 2001; Kurzban & Houser, 2001), and personal history and experience (Bettenhausen & Murnighan, 1985, 1991).

While advancing research on individual solutions, psychologists also pursued possibilities of structural solutions. Structural solutions alter the incentive structure of a social dilemma in such a way as to eliminate inconsistency between individuals’ incentives and collective consequences. To change the payoff structure directly is the most straightforward implementation of a structural solution. A large number of studies were conducted to investigate how changes in the payoff structure affect participants’ cooperative behavior. Results of these studies generally support the view that participants’ cooperation behavior is negatively related to the cost of cooperation and positively tied to the benefit of cooperation (Bonacich et al., 1976; Goehring & Kahan, 1976; Kelley & Grzelak, 1972; Komorita et al., 1980; Marwell & Ames, 1979, 1980; Stern, 1976; Yamagishi, 1988a). Changes in the payoff structure may also involve the form of the production function rather than the magnitude of the cost and benefit of cooperation. The form of the production function that translates contributions made by the group members to the level of the public good
produced include linear, disjunctive, conjunctive, step-level, and other forms (Dawes et al., 1986; Sato & Yamagishi, 1986; Van de Kragt et al., 1983; Yamagishi & Sato, 1986). Van de Kragt et al. (1983) reported that introducing the minimal contributing sets (MCS) enhances cooperation. Participants in their study were told that the public good would be supplied only if a specified number of contributors (or more) were reached. The provision rates were higher in the MCS condition than in a linear condition in which the level of public good was linearly related to the number of contributors. Dawes et al. (1986) found similar results. Yamagishi and Sato (1986) found that fear and greed play different roles with different forms of the production function. Other structural changes that were found to enhance cooperative behavior is small group size (Bonacich et al., 1976; Fox & Guyer, 1977; Marwell & Schmidt, 1972), territorialization of the commons (Cass & Edney, 1978; Messick & McClelland, 1983), and leadership (Messick et al., 1983).

Changes to the payoff structure can be introduced through administration of sanctions or selective incentives in the form of punishments of non-cooperators and rewards for cooperators. Caldwell (1976) was the first to examine the effect of mutual sanctioning in an n-person prisoner’s dilemma game. In this study, participants played several trials of a five-person prisoner’s dilemma game. They were randomly assigned to one of three conditions: the information-only condition, the talk-only condition, and the sanction condition. In the information-only condition, participants played a repeated n-person prisoner’s dilemma game with feedback information about other players’ decisions. Participants in the talk-only condition were allowed to talk with other members. Participants in the sanction condition were given opportunities to vote to impose penalties on each member, though the penalty did not cost any to the voters, in addition to talk with other members. The results clearly showed the effect of punishments; the opportunity for sanctions raised the level of cooperation from 46% to 65%, while communication per se did not significantly yield higher cooperation. Subsequent studies of punishment (Sato, 1987; Yamagishi, 1986a, 1988a, b, 1992) provided further evidence of the effectiveness of sanctions for enhancing cooperation in social dilemmas.

Demonstration of the power of punishment as a means to improve cooperation in social dilemmas failed to instigate much interest among social dilemma researchers; rather, it was met with serious criticisms. Most criticisms against the use of sanctions concerned either psychological or structural “side effects” of the use of sanctions. First, the negative psychological consequences of administering sanctions have been pointed out. The negative psychological implications are summarized in Taylor’s view that sanctions are like drugs; the more we use it, the more we need it (Taylor, 1976). Sanctions may work as a source of overjustification (Lepper et al., 1973) of intrinsic motivation for cooperation and thus destroy intrinsic motivation for cooperation (Yamagishi, 1990a). Enforcement of sanctions also destroys sense of community (Fox, 1985; Taylor, 1976, 1982) and sometimes makes people strive for uncontrolled means to free-ride (Bell et al., 1989; Mulder et al., 2003). The negative effects of sanctions to reduce intrinsic motivation for cooperation pointed out by these people have long escaped empirical investigation, but Mulder et al. (2006) recently demonstrated this effect in their study in which the experience of playing a
public-good game under the threat of sanctions reduces participants' level of cooperation once the threat is removed, even below the level that prevailed before experiencing sanctions.

The structural implications of sanctions concern several issues. The first issue is how to reach an agreement about the sanctioning rules. Hobbes' Leviathan is created by the consent of the people who prefer peace to war. Similarly, Hardin (1968) emphasized that mutual coercion as a solution to the social dilemma problem has to be mutually agreed upon. Given the diversity in interests and values among major players in the real-life social dilemma makes it extremely hard for them to agree on a particular rule for sanctions (Crowe, 1969). The second issue concerns the cost for enforcing sanctions. In the extreme example of a totalitarian regime, damages to psychological welfare and quality of life may be imposed by strict enforcement of sanctions (Lynn & Oldenquist, 1986). Even in a more liberal regime, the administrative cost of enforcing sanctions may exceed the benefits produced by the enhanced level of cooperation (Edney & Harper, 1978; Tullock, 1977). How to share and provide the cost for enforcing sanctions is the central issue in the structural implications of sanctions. This is often called the second-order social dilemma problem (Oliver, 1980; Yamagishi, 1986a, b). Once sufficient sanctions are provided and make all people cooperate for the provision of a public good, everyone benefits from the public good regardless of whether or not one participated in the sanctioning activities. In this regard, sanctions are a public good and face the same problem that the original public good faces. It is not individually rational—against one's own short-term self-interest—to engage in sanctioning activities, in the same way contributions to a public good are not rational. If this is, in fact, the case, sanctions create a new dilemma in the way of solving the original dilemma problem (Yamagishi, 1986a, b).

Partly because of these difficulties that sanctions as a solution to the social dilemma problem face, and partly because of psychologists' intrinsic interest in psychological issues rather than social structural issues, structural solutions in general, and sanctions as a means for structural changes in particular, have never received a wide array of attention among psychologists working on the problem of social dilemmas, except in applied studies of real-life problems such as traffic congestion (Van Vugt et al., 1996a, b) and surveillance and monitoring in companies (Tenbrunsel & Messick, 1999). During the last couple of decades, in which psychologists have been more or less silent on structural solutions, economists have picked up the topic and have developed an impressive array of research. In the following section, we will review the recent studies on sanctions conducted mostly by experimental economists.

Research on Sanctions and Sanctioning Behavior

As we mentioned above, most of the research on sanctions is conducted by experimental economists. We refer you to Camerer (2003), Kagel and Roth (1995), and Rabin (1998) for an overview of the research by experimental economists on social
dilemmas in general and, instead, focus our discussion on their studies of sanctions. Let us start with a brief discussion of a particular feature of the sanctioning studies conducted by experimental economists; most of their studies deal with sanctions as individual behavior to another individual rather than a behavior taken by an institution that imposes a penalty on free riders. This feature of sanctions that experimental economists study is appropriate when the goal of the researchers is in individuals' motivations or preferences that make them sanction other people, rather than the effectiveness of various enforcing schemes. In other words, their orientation in their sanctioning studies is very much psychological rather than structural. This orientation creates a fertile ground for exchanges of ideas and research activities between psychologists and economists, as we will discuss in the next section of this chapter.

Their rather psychological orientation goes back to the "surprising" finding that people actually expend their own resources (viz. money) to retaliate unfair treatments by others (Guth et al., 1982; Fehr & Fischbacher, 2003). This finding is no surprise to psychologists who have been studying such behavior with such descriptive theories as equity theory (Adams, 1963) and frustration-aggression hypothesis (Dollard et al., 1939). However, such findings were, in fact, surprising for economists who had been operating under the assumption that people maximize their own payoffs. Rejection of unfair offers in ultimatum game experiments with monetary costs to the participant that sometimes are worth a few months' earnings (Cameron, 1999; Fehr et al., 2002) surprised some economists and prompted them to pursue a new model of humans that is capable of explaining such behavior. The result of this endeavor came out as new theories of social preference that explain irrational, fairness-based behavior observed in ultimatum and related games (Bolton, 1998; Falk & Fischbacher, 2006; Sobel, 2005).

One way of explaining social preference that produces behavior reducing payoff to the individual such as cooperation, retaliation, and pursuit of fairness is to regard it as a means to enhance the long-term benefit to the individual or to those genetically related to him. For example, cooperation between genetically related individuals, or altruistic behavior toward kin members, can be a means to enhance inclusive fitness, as presented in the theory of kin selection (Hamilton, 1964). Cooperation in a long-lasting relationship can be explained in terms of reciprocal altruism (Axelrod & Hamilton, 1981; Trivers, 1971) that is equivalent to the tit-for-tat strategy in repeated PD games (Axelrod & Hamilton, 1981). The theories of indirect reciprocity (Milinski et al., 2001; Nowak & Sigmund, 1998) and costly signaling (Zahavi, 1977) suggest that cooperation can evolve even in large groups when cooperators build up reputations through cooperative, altruistic, and/or fairness-enhancing behavior. While admitting that the ultimate causes of social preferences are founded on fitness-enhancing mechanisms such as those mentioned above and others including cultural transmission and cultural group selection (Boyd et al., 2003), experimental economists who take the social preference approach aspire to demonstrate the existence of social preferences that make people engage in behavior that reduces their own immediate payoff. In order to demonstrate the existence of such social preferences, their experiments are designed to eliminate the possibility of explaining their target behavior in terms of long-term, indirect benefits to the actor herself within the
experimental setting. That is, their experiments are typically designed to allow them to draw a conclusion that the participant's social preferences can be the sole explanation of the target behavior; no long-term, indirect implications of the behavior to her self-interest that explain the behavior exist in the experimental settings.

Research on sanctioning behavior by experimental economists more or less share this design feature with their studies of social preferences in general. Fehr and Gächter (2002), for example, showed in their influential article experimental findings from a study using such a design feature that people engage in altruistic punishment that brings no benefit to the participants themselves while being costly to them. This study stimulated other researchers' interests in sanctioning behavior and prompted subsequent studies, some of which we show next.

Altruistic Punishment

Cooperation in public-good games has been widely studied in psychology and economics in both repeated and one-shot experiments. Yamagishi (1986a, 1988a, b, 1992) studied sanctions in repeated public-good games. And yet, Fehr and Gächter (2002) were the first to study a public-good game with sanctions in a repeated one-shot setting. They examined the effect of the opportunity of mutual sanctions among group members in a public-good game. Each round of their game consisted of two stages. In the first stage, participants decided on the level of their contribution for a four-person public good as described below. In each trial they received an endowment of 20 money units (MUs), and each contributed any units of it (between 0 and 20MUs) to a group project. They kept the money that they did not contribute. For every MU invested in the project, each of the four group members earned 0.4 MUs, whereas 1 MU that participants kept was exactly 1 MU. If all group members kept all MUs, each participant earned only 20 MUs, while if all of them invested their 20 MUs, each of them would earn 0.4 \times 80 = 32 MUs. After the decision, they were informed about the investments by each of the other participants. In the second stage, participants assigned in the punishment condition decided how many MU points (between 0 and 10) they would use to reduce earnings of each of the other group members. For each point a participant used, three points were subtracted from the target member's earnings. This game was repeated for six trials, while group membership changed from trial to trial such that no participant ever met another participant more than once. In this repeated one-shot design, there was no opportunity for reputation building and direct reciprocity since participants never met anyone who knew what they had done in the previous games; thus, punishers were not able to expect any indirect benefit (through enhanced reputation) from engaging in costly punishment. Their punishment might enhance the target member's cooperation in future trials, but when it happened, they would not be in the same group and thus would not enjoy the benefit of enhanced cooperation from the target member. This feature of the game makes it impossible to explain the participant's punishment behavior in terms of his long-term benefit. Punishment
behavior in this game is thus purely altruistic, since the benefit produced by the costly punishment goes only to others. Fehr and Gächter (2002) call such punishment in the repeated one-shot game "altruistic punishment." They found that 84% of the participants in the punishment condition punished at least once, and 34% punished more than five times during the six trials. The effect of punishment was very clear. The average contribution in the punishment condition was higher than that in the no-punishment condition in any of the periods. In the final period, 59% of the participants in the no-punishment condition contributed nothing, while 39% in the punishment condition contributed their entire endowment. Fehr and Gächter (2000) also included another condition in which participants kept interacting with the same set of members (partner condition) and compared the effect of punishment in the repeated one-shot game (stranger condition). They found that the contribution rate was much higher in the punishment/partner condition (close to 100% toward the end of repeated trials) than in the punishment/stranger condition (of which the overall cooperation rate was 58%), whereas contributions in the no-punishment/stranger condition approached zero toward the end of repeated trials. They found not much difference in the use of sanctions between the partner and the stranger conditions. In either condition, those whose contribution levels were above average were hardly punished at all, whereas among the below-average contributors, the degree to which their contributions fell short of the average was strongly related to the amount of punishment they received. Furthermore, the size of punishment given to free riders at each level of free riding (i.e., the size of deviation from the average) was not much different between the two conditions either. These results suggest that punishment behavior is mostly based on a social preference rather than considerations of long-term self-interest.

Carpenter and Matthews (2002) reported that participants pay costs for punishing free riders in another group as well as in their own group. Their participants were assigned to a 4-person group whose members stayed through all 10 periods (i.e., equivalent to the partner condition in Fehr & Gächter, 2000). The experimental session consisted of two separate four-person groups playing the same public-good game simultaneously and independently of each other. Participants benefited only from their group's contribution to a public good. However, they were provided with opportunities to punish any one of the other seven players regardless of which group they belonged to. Since the group membership stayed the same during the 10 trials, punishing members of their own group may serve the punisher's own interest through enhancing other members' cooperation level. However, punishing members of the other group was useless as a means to improve their self-interest; it only helped the cooperation level in the other group. They found that opportunities to punish only members of their own group raised contributions as in the partner treatment in Fehr and Gächter (2000). They also found that when participants were provided with opportunities to punish outgroup members as well as ingroup members, cooperation rates increased even more, higher than in the condition in which only ingroup punishment was allowed. About 50% of participants punished outgroup members at least once, whereas they punished ingroup members much more heavily than outgroup members.
These experimental results suggest that people punish non-cooperators even when there is no possibility that their punishment behavior improves their self-interest. These findings made the researchers conclude that at least some of the punishment behavior observed in their studies is based on the participants’ social preference that internally instigates them to engage in punishment behavior. The social preference for punishing non-cooperators, however, may not be strong enough to overcome fear of retaliation. Most experimental studies of sanction behavior share a design feature of anonymity of punishers. That is, participants are informed of who contributed how much, but are not typically informed who punished whom. This feature provides the punisher immunity from retaliation from the punished. When this immunity from retaliation was not provided in an experiment in which the punisher’s identity was revealed, Nikiforakis (2004) found that opportunities of punishment did not have a positive effect on cooperation.

While the above studies suggest that punishment behavior is not a strategy to ensure long-term self-interest, it has been shown not to be free of cost-benefit considerations. Anderson and Putterman (2006), using a repeated one-shot game, showed punishment behavior to increase as the cost for the punishment was reduced. Their participants repeated a voluntary contribution game in a three-person group with a different set of players for each trial. They manipulated the costs of punishment as a within-subject factor over five trials; 0 to 120 experimental cents were required to subtract one experimental dollar from the earnings of the target of punishment. Their results indicate that the size of punishment administered by the participants was negatively related to its cost for the punisher, and positively related to the target member’s degree of free riding. The negative relationship between the size of punishment and the cost of punishment was robustly observed at each level of the target’s free riding. Carpenter (forthcoming-a) conducted an experiment similar to Anderson and Putterman’s (2006) and reported that a 1% increase in the cost of punishment reduces the quantity of punishment delivered by 1.22%. There results, taken together with the result that people punish free riders even when they earn nothing from such behavior, suggest that people derive psychological satisfaction from punishing free riders, but they pay to “purchase” the satisfaction.

**Motivations Behind Sanctions**

Based on the evidence provided by several studies that people pay substitute costs to deliver punishment while expecting no return for it, some researchers think that they are ready to pose a new question: “It is no longer the question whether there is informal sanctioning. The problem, which is not yet understood, however, is why people sanction and, in particular, why they sanction others’ cooperative or defective behavior” (Falk et al., 2005, p. 2017). As shown below, the answer to this question requires information about who punishes whom. In earlier studies that used punishment schemes that were prepared by the researcher (McCusker &
Carnevale, 1995; Yamagishi, 1986a, b, 1992), the target of punishment—the least cooperative member—was fixed and thus was not the topic of the study. The earlier researchers overlooked this issue simply because they assumed that punishments would be directed to free riders. This assumption is mostly true. Most punishments are imposed by cooperative members on less cooperative members. In Fehr and Gächter (2002), 74.2% of all punishments were delivered by above-average contributors and directed to below-average contributors. Also, as mentioned earlier, Fehr and Gächter (2000) and Carpenter and Matthews (2002) found that most punishments were directed toward below-average contributors. How shall we explain this pattern? Price et al. (2002) proposed an evolutionary explanation, fitness differential theory, of why cooperators punish defectors. According to this explanation, cooperative behavior cannot evolve in a population of “egoists” unless cooperators eliminate the benefits free riders earn. They further argued that punishment eliminates free riders’ fitness advantage by reducing free riders’ earnings. Since the “goal” of punishment in evolutionary terms is in the reduction or reversal of free riders’ fitness advantage, the proximate or psychological mechanism to execute this “goal” should be found in a punitive sentiment or a desire to harm free riders. This evolutionary “goal” is achieved by harming free riders more than the cost they bear to impose such harm to free riders. In other words, people should have a competitive social value orientation—maximization of relative gain—toward free riders. This kind of social value or social preference, they argue, is likely to be “hard-wired” into our motivational circuitry, and that is why game players punish free riders even when it is costly to do so.

According to the fitness differential theory briefly summarized above, cooperators should punish free riders because a penalty reduces free riders’ payoffs more than it reduces the punisher’s payoffs. However, Falk et al. (2005) found that cooperators punish defectors even when the punishment does not reduce the payoff difference between the punisher and the punished. Participants in their study played a one-shot, three-person prisoner’s dilemma game. They first decided whether or not to cooperate and then were provided with an opportunity to subtract money from each of the other group members. However, delivering punishment in one condition of this study was very costly; the cost for delivering punishment is as expensive as the harm delivered to the target. That is, they were required to pay one token to subtract one token from the target. They compared this condition with another condition in which punishment was less costly—1 token afforded 2.5 to 3.3 tokens of punishment. They were informed of all members’ choices in the public-good game and decided how much they should pay to subtract money from each of the other three. In the less costly punishment (1 token affording 2.5 to 3.3 tokens of penalty) condition, 69% of cooperators paid costs to reduce defectors’ earnings, and only 7% of them paid costs to reduce cooperators’ earnings. In the high-cost condition in which one token afforded one token of penalty, the pattern was almost exactly the same; 60% of cooperators punished defectors and none of them punished cooperators. Furthermore, a similar but less pronounced pattern of punishing defectors regardless of the fine-to-fee ratio was observed among
defectors. When punishment was effective (1:2.5–3.3), 40% of defectors punished other defectors, and 34% of them punished cooperators. When punishment was costly (1:1), only 2% of defectors punished other defectors, and none of them punished cooperators. These findings are inconsistent with the fitness differential theory proposed by Price and his associates, since, first of all, defectors punish other defectors, and, second, cooperators (and defectors) deliver punishment in the same way when such behavior does not eliminate or reduce free riders’ advantages as when it eliminates or reduces the difference. That is, cooperators’ punitive behavior was not driven by competitive social preference as assumed in the fitness advantage theory. Falk et al. (2005) suggested that retaliation (i.e., the desire to harm those who committed unfair acts, rather than to reduce the fitness advantage of free riders) seems to be the most likely motive for punishment by cooperators, not spiteful or competitive motivation (i.e., the desire to maximize the payoff difference between self and other).

Falk and his associates also compared the results from the “strategy method” with those from the “specific response method” and found some significant differences between the two. This advises us to use some caution in interpreting the finding of punishment by defectors. With the use of strategy method, participants are provided with a number of scenarios in which other participants’ choices are systematically varied and are asked to make their own decision. They are instructed that they will be rewarded for their own and other participants’ choices in a randomly chosen scenario. With the use of the specific response method, participants actually face a decision environment and make decisions in a particular environment. The strategy method is similar to the “role-playing experiment” advocated in social psychology in the 1960s and 1970s as a means to avoid using deception (Kelman, 1967). The only difference from the role-playing experiment in social psychology is that participants in the strategy method are actually paid according to their choices in one scenario. While social psychologists have never resolved the problem of internal validity associated with role-playing experiments, economists are more convinced of the validity of the strategy method (Hertwig & Ortmann, 2001). The strategy method has been used frequently in experimental economics because researchers can elicit a large number of responses without actually letting their participants face real decision tasks repeatedly. It also has the advantage of studying the effect of stimuli combinations that rarely occur in natural settings; it is a convenient means to study such effects without artificially creating the stimuli combination by deceiving participants. While some studies report no difference in the results produced by the two methods (e.g., Cason & Mui, 1998; Brandts & Charness, 2000) on some behavior, Falk et al. (2005) reported that 34% of the defectors punished cooperators with the strategy method while only 17% of them did so with the specific response method. Interpretation of such a difference is difficult at this stage. On the one hand, defectors may simply be thinking that they would behave in a socially desirable manner when they respond to the strategy method, whereas concerns for self-interest become more salient to them in the specific response method.
Group Size

Group size has been known to be negatively related to the level of cooperation in social dilemmas (e.g., Kerr & Bruun, 1983; Marwell & Schmitt, 1972). The negative effect of group size on cooperation has mostly been attributed to correspondent changes in the incentive structure; for example, benefit from one’s own contribution is reduced in larger groups. However, the negative effect of group on cooperation is also observed even when the incentive parameters are kept constant over different group sizes (Bonacich et al., 1976; Sato, 1988; Yamagishi, 1990b), while the negative effect disappears in others studies with the constant parameters (Isaac & Walker, 1988; Isaac et al., 1994). The “residual” negative size effect has been explained in terms of a decline in the influence a change in one’s behavior has on other participants (Bonacich et al., 1976; Dawes, 1980; Kerr, 1989) or reduced expectation for other members’ cooperation (Yamagishi, 1990b). Compared to the abundance of studies of group size on cooperation, experimental studies of the effect of group size on sanctions is limited. Yamagishi (1992) partially confirmed his prediction, in a comparison of four-person and eight-person groups, that people would be more willing to contribute to the provision of a punishment system in larger groups than in smaller groups. His results showed that high-trusters increased their contribution to the punishment system in larger groups, whereas the effect of group size on punishment was in the opposite direction among low-trusters. Bowles et al. (2001) compared 5-person groups with 10-person groups in a repeated one-shot public-good game with punishment. They also compared MPCR (marginal per capita return; the ratio of benefit from public good allocated to each player to the cost for its provision) in two conditions; it was set at 0.3 in one condition, and 0.5 in the other. After participants were informed of the other members’ contribution level, they were provided with an opportunity to reduce other participants’ earnings. Each point they paid reduced the punished member’s payoff in the same trial by 10%. The results of their study revealed a positive effect of group size on both cooperation and punishment; participants paid more money for delivering punishment and cooperated more in 10-person groups than in 5-person groups. They also showed that MPCR was positively related to both cooperation and punishment. Carpenter (forthcoming-b) also manipulated both group size and MPCR and partially replicated Bowles et al.’s (2001) findings. As in Bowles et al. (2001), Carpenter found a positive relationship between MPCR and both cooperation and punishment. Furthermore, he found more cooperation in larger groups than in smaller groups. However, the effect of group size was in the opposite direction to the one found by Bowles and his associates. Carpenter found that participants delivered less punishment in larger groups than in smaller groups.

Why does group size have a positive effect on cooperation and punishment (though not observed in Carpenter, forthcoming-a) with sanction opportunities present, while it has a negative effect on cooperation in the absence of sanction opportunities? Yamagishi (1992) predicted the positive effect of group size on
punishment based on his structural goal/expectation approach. According to the goal/expectation approach proposed by Pruitt and Kimmel (1977), people are motivated to cooperate when they realize the need for mutual cooperation. However, motivation to cooperate is not sufficient for them to actually cooperate. The motivated people need to feel that their goodwill will not be exploited by others. Based on this goal/explanation approach, Yamagishi (1986a, b) proposed that people realize the need for establishing a sanctioning system when the need for mutual cooperation is strongly felt and, at the same time, when they realize voluntarily based mutual cooperation is hard to achieve. Yamagishi (1988a) conducted an experiment to test this idea and found that the seriousness of a social dilemma (greater benefit from cooperation and greater cost of cooperation) induced his participants to contribute more to the establishment of a sanctioning system. Group size is a factor that makes voluntarily based cooperation harder to achieve, and thus, Yamagishi (1992) argued, people are more strongly motivated to contribute to establish a sanctioning system.

In contrast to Yamagishi's (1992) psychological explanation, Casari (2005) proposed a structural explanation of the positive effect of group size on punishment. Casari (2005) argued that how the cost of punishment each player bears is translated to the penalty delivered to a particular player affects the level of punishment provided in general, and how group size is related to the level of punishment, in particular. According to Casari (2005), punishment technologies commonly used in experiments are classified into two categories depending on the fine-to-fee ratio: the "neutral" and the "non-neutral" punishment technologies. A neutral technology lets a player punish another with a constant fine-to-fee ratio in all circumstances. Studies that used this technology include Fehr and Gächter (2002), Sefton et al. (2002), Page et al. (2005), Carpenter (forthcoming-a), Andreoni et al. (2003), and Anderson and Putterman (2006). In contrast, a non-neutral technology allows the fine-to-fee ratio to vary with members' behavior in the first stage. The most extreme case of this technology is when only defectors can be punished, like in a legal system. Fehr and Gächter (2000), Bowles et al. (2001), Carpenter (forthcoming-b), Masclet et al. (2003), and Nikisforakis (2004) adopted this type of technology. For example, in Bowles et al. (2001), the earnings of the punished member in a public-good game were reduced by 10% for each punishment point received. This makes the size of the fine vary according to the earnings of the punished in the public-good game. With the use of this technology, it costs less to deliver the same fine to a cooperator than to a defector who earned more than a cooperator in the public-good game. This should produce the finding that participants spend more of their money to punish defectors than to punish cooperators, even when their goal is to deliver the same level of punishment to defectors and cooperators. If this is, in fact, the case, the generally observed pattern that participants spend more of their own money to punish defectors than cooperators can be artificially inflated than the actual pattern produced by social preference for retaliation, fairness, or spite. Casari (2005) further argued that the positive effect of group size on punishment may be mediated by an increase in the fine-to-fee ratio when a non-neutral technology is
used. This is because the total benefit each player receives from a public good increases in larger groups, as in the case of a public good characterized by non-excludability from consumption. The increase in the benefit to each member from the public good also increases the fine-to-fee ratio. The positive effect of group size on cooperation and punishment observed in experiments using a non-neutral technology can be a result of the increase in the fine-to-fee ratio rather than a product of some social preferences. Casari (2005) advised using a neutral punishment technology in the study of social preferences as a motivator of punishment behavior in social dilemmas. Providing a definite answer to the question of whether group size increases sanctioning behavior and cooperation needs to wait for further studies; it is an important topic for future research since it has serious implications for possible solutions in the large-scale social dilemmas we face in real life. Experimental research on social dilemmas has been criticized for a lack of generalizability of its findings, such that what works well in small groups may not work well in larger groups. The positive effect of group size seems to imply that what works fine in small groups will work better in larger groups.

Punishments and Rewards

Comparisons of rewards and punishments as reinforcers are one of the most traditional topics in psychology (Skinner, 1938). In the behaviorist tradition, positive reinforcers (i.e., rewards) are usually more effective than negative ones (i.e., punishments) except in certain conditions (Bandula, 1969; Millenson, 1967). Platt (1973), who analyzed the commons dilemma from the behaviorist perspective, proposed a solution in terms of administration of immediate negative consequences, such as punishment of inappropriate behavior. While the distinction between positive and negative frames drew the attention of social dilemma researchers (social dilemmas versus social fences—see Messick & Brewer, 1983; give-some versus take-some type of social dilemmas—see Hamburger, 1974; Hamburger et al., 1975), and a general conclusion that a positive frame is more conducive to cooperation than a negative frame even when the payoff structures are held constant (Komorita & Barth, 1985), studies that compared positive sanctions (i.e., rewards) and negative sanctions (punishments) are relatively few.

Rapoport and Au’s (2001) experiment is an example of such a study; they compared the reward condition with the penalty condition using a repeated one-shot game with no feedback about outcomes in each trial. In each trial, participants played a common pool resource dilemma (Gardner et al., 1990), taking a private share of the resource from a common resource pool. The private share of the common resource pool became theirs insofar as the total amount requested by all the members of a five-person group was kept below a threshold; once the total amount was exceeded, they lost their share. In the reward condition, the member who
requested the least amount was given a bonus by the experimenter. In the penalty condition, the experimenter imposed a penalty on the member who requested the largest amount. The size of the bonus and the penalty were the same. Both rewards (bonus) and punishments (penalty) had a positive effect on cooperation, but the effect of punishments was stronger than that of rewards.

However, McCusker and Carnevale (1995) reported a finding from their experiment that is opposite to the finding by Rapoport and Au (2001). Their participants played a repeated dilemma game in a seven-person group consisting of the same set of participants across trials. Information about other players' decisions—actually pre-programmed by the experimenter—was given to the participant. In each trial, participants decided how much to contribute in a public-good game, and then they were given an opportunity to contribute to a “reward fund” (reward condition) or a “penalty fund” (penalty condition). In the reward condition, the sum of money contributed to the reward fund was given to the most cooperative member. In the penalty condition, the sum of money contributed to the penalty fund was deducted from the least cooperative member's earnings. Participants contributed more to the reward fund than to the penalty fund, and their cooperation level was higher in the reward condition than in the penalty condition.

Sefton et al. (2002) compared the effects of rewards, punishments, and the combination of both using a repeated public-good game in a four-person group in which they played with the same set of members while their ID was varied in each trial. After receiving feedback about the members’ contributions in the public-good game, their participants were given an opportunity to assign between zero and six tokens to each of the other three members for either punishing or rewarding him. Each token cost 10 cents to the rewarder/punisher and delivered the same 10 cents of reward or fine to the target member. Participants in the reward and punishment condition were given a choice of using the tokens for either rewarding or punishing each of the other members. At the end of each trial, participants were informed of their earnings in that trial, including rewards and/or fines they received. The cooperation level in any one of these three conditions was higher than that in the baseline condition (i.e., no-punishment, no-reward condition). The cooperation level in the public-good game was the highest when participants were able to provide both punishments and rewards (i.e., reward and punishment condition), while no difference was found between the punishments-alone condition and the rewards-alone condition. While the cooperation level was not statistically different between the punishment condition and the reward condition, the use of rewards in the reward condition was less frequent than the use of punishments in the sanction condition. At the same time, the use of rewards was more frequent than the use of punishments in the reward and punishment condition.

Walker and Halloran (2004) found no effect of either punishments or rewards despite the fact that they used a similar design to the one used by Sefton et al. (2002) with the following changes. First, Walker and Halloran (2004) used a repeated one-shot game instead of the repeated game with no history used by Sefton and his associates. Second, no feedback of the other members' rewarding
and/or punishing behavior was provided. Finally, the value of each token for punishment or rewarding was much smaller than in the Sefton et al. study; one token for punishments or rewards was worth one-fifth of one token used in the public-good game, while the ratio was one to one in Sefton et al. (2002). The lack of effect of sanctions (either punishments or rewards) and cooperation in this study may be attributed to any one of the three changes in Walker and Halloran’s (2004) study. First, the effects on punishments and rewards may require that a game be repeated among the same set of people. Second, the lack of the effect may be due to the unavailability of information about other members’ sanctioning behavior rather than to the repeated nature of the game. Finally, the size of punishments and rewards may have been too small to have an impact on participants. To identify the roles that these factors play in sanctioning behavior is an important topic for future study.

An interesting twist to the reward versus punishment controversy concerns the second-order sanctions (Cinyabuguma et al., 2006; Kiyonari et al., 2005, 2007). Second-order sanctions are sanctions on sanctions (punishment of non-punishers, rewarding of rewarders, etc.). One of the criticisms of sanctions as a solution to the social dilemma problem is that sanctions themselves are a public good and thus create a “second-order social dilemma” (Oliver, 1980; Yamagishi, 1986a, b). Axelrod (1986) provided an answer to this criticism in the form of “meta-norms.” According to him, meta-norms involve sanctions of non-enforcers of norms (e.g., those who do not punish defectors). If those who enforce norms (those who punish defectors) also enforce meta-norms (punish those who do not punish defectors), and those who do not enforce norms do not enforce meta-norms, the second-order social dilemma problem disappears within a certain range of parameters. Yamagishi and Takahashi (1994) reanalyzed this issue with a computer simulation and concluded that meta-norms are not really needed insofar as the “linkage” or behavioral consistency between two levels of dilemmas exists between the cooperation and punishment, instead of punishment and meta-punishment as argued by Axelrod (1986). In order to test Yamagishi and Takahashi’s (1994) claim, Kiyonari et al. (2005, 2007) examined three kinds of behavior—cooperation, punishment, and meta-punishment—in a four-person group. Results of their study show a fair amount of consistency between cooperation and punishment as observed in other studies in which those who punish defectors are mostly cooperators. However, they did not find consistency between punishment and meta-punishment; actually, meta-punishment occurred only very rarely. Kiyonari et al. (2005, 2007) replicated this finding and, in addition, found that the meta-sanctions in the positive direction—i.e., rewards to those who reward cooperators—occur frequently and are consistent with the first-order rewarding (rewards to cooperators). The positive chain of reward giving is likely to survive and prosper as suggested by the indirect reciprocity literature (Brandt & Sigmund, 2005; Milinski et al., 2001; Nowak & Sigmund, 1998; Takahashi & Mashima, 2003), but the negative chain of punishment giving seems to stop at the second step.
Social Sanctions

The studies we presented above all used monetary sanctions in the form of a fine or rewards. This fixation with money is a matter of convenience in conducting research; it is not based on a conviction that only monetary sanctions are important. Social approval and disapproval are important means to control our behavior in social life. Earlier exchange theorists in sociology (HOMANS, 1961; EMERSON, 1972) emphasized the importance of non-monetary rewards and sanctions. BLAU (1964) argued that informal sanctions such as peer pressure, gossip, and social ostracism are effective deterrents against defection in social exchange. Some economists who recognize the effects of social sanctions have constructed theoretical models to include peer pressure (KANDEL & LAZEAR, 1992; BARRON & GJERDE, 1997) and avoidance of social disapproval (akerlof, 1980; HOLLANDER, 1990). Bowles and Gintis (2003) proposed an evolutionary model of social emotions according to which some social emotions such as shame coevolve with other emotions motivating punishment of antisocial actions.

MASCLET et al. (2003) argued that a sanctioning behavior using monetary fines in Fehr and Gächter (2000) was a vehicle to express disapproval of others’ free-riding behavior. They argue that facing expression of disapproval from other members in itself, even without being monetarily penalized, would increase players’ level of contribution. They conducted an experiment in which they compared monetary sanctions with non-monetary sanctions. Their participants played a repeated public-good game in a four-person group. They decided on the level of contribution for a public good between 0 to 20 experimental currency units (ECUs). In addition, participants in the non-monetary punishment condition had an opportunity to inform the target member of their level of disapproval. They did not pay for their ECUs for this action. For participants in the monetary punishment condition, the ECUs they spent on punishment were used to reduce the target member’s earnings. In this experiment, the non-monetary sanctions—expressions of disapproval—raised the average contribution from 6.6 ECUs in the no-punishment treatment to 8.97 ECUs, while monetary punishment had about twice as strong an effect and raised it from 6.0 ECUs to 11.1 ECUs. The effect of the non-monetary punishment did not emerge in a repeated one-shot game.

NOUSSAIR and Tucker (2005) compared three conditions: monetary punishment, non-monetary punishment, and bother punishment—both monetary and non-monetary punishments. The average contribution to the public-good game was higher in the monetary punishment condition than in the non-monetary punishment condition. Furthermore, the average contribution level was not higher in the bother punishment condition than in the monetary punishment condition, indicating that adding non-monetary punishment to monetary punishment did not make people cooperate more. On the other hand, participants’ earnings after deducting costs for monetary punishment and fine were larger in the bother punishment condition than in the monetary punishment condition. That is, about the same effect was achieved in the
bother punishment condition despite the fact that participants spent less money for punishment when they could express their disapproval. Non-monetary punishment replaced some of the effects of monetary punishment.

The relatively weak power of non-monetary punishment observed in these studies may be due to the fact that disapproval was communicated without face-to-face interactions. The cooperation-enhancing effect of face-to-face communication has been extensively studied since the early days of social dilemma research (e.g., Bixenstine et al., 1966; Brechner, 1977; Dawes et al., 1977; Edney & Harper, 1978; Jerdee & Rosen, 1974; Rapoport et al., 1962), and the well-established effect is, at least partly, due to social approval and disapproval exchanged in face-to-face communication (Ostrom et al., 1992).

The effect of a non-monetary sanction in a face-to-face situation has also been examined outside the laboratory. Barr (2001) reports results of a public-good game with punishment conducted in rural communities in Zimbabwe. Interestingly, participants in this study who had witnessed the disapproval of those who contributed only a small amount increased their contribution by even more than those directly criticized. This result suggests the indirect effect of social disapproval as a warning to the third party, in addition to its direct effect as a punishment to a transgressor. Single-mindedly focusing on the direct effect of non-monetary punishment and concluding that it has only weak power on enhancing cooperation is thus premature; more research is needed on the indirect effect on non-monetary sanctions.

One issue surrounding non-monetary sanctions concerns its costless nature; does the power of non-monetary sanctions depend on its cost-free nature? A study by Carpenter et al. (2004) showing that their participants were willing to pay money to signal their disapproval of free riding suggests that this is not the case. In their study, Thai and Vietnamese participants in urban slums in their respective countries who played a repeated voluntary contribution game with social sanctions paid monetary costs to display a picture of an unhappy face (34% of Thai participants and 25% of Vietnamese participants disapproved at least once).

**Net Benefits of Sanctions**

Early critics of sanctions as a means to solve social dilemma problems raised the issue that the costs of sanctions may exceed their benefits (Crowe, 1969; Fox, 1985; Lynn & Oldenquist, 1986; Stillman, 1975; Taylor, 1976, 1982). The costs of sanctions include long-term as well as short-term costs. Short-term costs are monetary costs for monitoring free riders and delivering sanctions upon them. Long-term costs include reduction in intrinsic motivation for cooperation (Taylor, 1976; Yamagishi, 1990a) and the sense of community (Fox, 1985; Taylor, 1976, 1982). We will later discuss recent studies on the negative psychological effect of sanctions; we focus on the short-term, monetary costs in this section. As presented
earlier in this chapter, most studies of sanctions report the cooperation-enhancing effect of punishment (e.g., Bowles et al., 2001; Carpenter, forthcoming-b; Carpenter & Matthews, 2002; Masclet et al., 2004; Fehr & Gächter, 2002; Yamagishi, 1986, 1988a, b, 1992). However, the short-term monetary costs including both the cost for the punisher as well as the penalty paid by the punished often exceeded the extra benefits derived from the enhanced level of cooperation (Bochet et al., 2006; Carpenter & Matthews, 2002; Ostrom et al., 2002). It seems to be obvious that the net benefit of sanctions depends on several structural factors such as the MPCR of contribution in the public-good game and the fine-to-fee ratio in the sanctioning game. When the MPCR is high, only a small improvement in the cooperation level produces a large benefit, and thus the net benefit of punishment is more likely to be in the black. Also, an improvement in cooperation and the benefit from a public good can be achieved by a small fee when the fine-to-fee ratio is large. In lieu of this interpretation, Yamagishi (1986) found that the net benefit of punishment was positive with a high fine-to-fee ratio, whereas it was negative with a low fine-to-fee ratio. Given the fact that the net effect of sanctions is positive or negative depends on the nature of some parameters characterizing the game, it is impossible to draw a general conclusion on the overall benefit of sanctions. Instead of drawing a general conclusion, we need to pay attention to the specifics of the social dilemma problem to see if sanctions are useful and desirable as its solution.

Future Directions

As presented above, most of the recent research on sanctions in social dilemmas has been conducted by experimental economists. This is not surprising given the fact that psychologists are not as interested in social structural factors as social scientists. An interesting irony, however, is that the interest of experimental economists who work on sanctions is mostly focused on the psychology of sanctioning behavior, as briefly discussed earlier. Their primary research goal is to demonstrate the operation of social preferences that make people sanction free riders. It seems to the authors of this chapter that there is a lot for psychologists to contribute to this endeavor. We will present below some recent developments by social psychologists which we think enrich studies of sanctions in social dilemmas. Strong methodological positions taken by experimental economists (cf., Hetwig & Ortmann, 2001)—such as the use of monetary incentives and rules against the use of deception—may keep the fruit of some psychological research from having an impact on economists, but we believe that it will not take too long for an eventual collaboration between the two to emerge since sanctions are the place where psychology meets structure. We will present a few research topics that may turn out to be fertile grounds for psychologists’ contributions to the sanctioning research in social dilemmas.
The Dark Side of Sanctions

As mentioned earlier, the use of sanctions advocated by the precursors of social dilemma researchers was heavily criticized because of its potential negative consequences. Social psychologists had been concerned with the negative consequences of extrinsic incentives that occur in the form of depletion of intrinsic motivation (Deci & Ryan, 1985; Lepper et al., 1973). It was thus natural for those who had been trained in social psychology to think about the dark side of sanctions in the context of social dilemmas. The anticipated negative consequences included loss of intrinsic motivation to cooperate and of sense of community (Fox, 1985; Taylor, 1976, 1982; Yamagishi, 1990a) in addition to excessive cost for administering sanctions (Edney & Harper, 1978; Tullock, 1977). Economists' focus on the motivational foundations of sanctioning behavior, on the other hand, made them less concerned with such negative consequences of sanctions. It is of some interest to note that it is psychologists working in business schools who conducted experimental research on negative psychological consequences of sanctions. Tenbrunsel and Messick (1999) conducted the first experiment focused on the negative psychological consequences of sanctions. In their studies, Tenbrunsel and Messick (1999) compared the strong sanction condition and the weak sanction condition with the no-sanction condition. They found that the cooperation level in the weak sanction condition was lower than that in the no-sanction condition, whereas the cooperation level in the strong sanction condition was higher than that in the no-sanction condition. They explained these findings in terms of the framing effect. Threats of sanctions make participants think of the game situation as one in which people are driven by extrinsic incentives. That is, participants see the game situation in a “business frame” when the threat of sanctions exists. In the absence of such salient extrinsic incentives, participants tend to see the game situation as one in which people seek to cooperate. They see the game in a moral frame (cf. the might-over-morality literature on this issue; Kelley & Stahelski, 1970; Liebrand et al., 1986). In the weak sanction condition, participants see the game in a business frame, and yet the sanctions are not strong enough to make them cooperate as rational actors. When the sanction is strong enough, even those who have adopted a business frame do cooperate since cooperation is a rational strategy under strong sanctions.

Another, more recent study by Mulder et al. (2006) demonstrates that a sanctioning system in social dilemmas undermines trust—the belief that other members are motivated to cooperate. They developed a “removing the sanction” paradigm in which participants play a social dilemma twice. In the first phase, they play a public-good game in which a sanctioning system is either present (sanction condition) or not (no-sanction condition). In the second phase, they play the same game without a sanctioning system in both conditions. The goal of their experiments was to demonstrate that the cooperation level in the second phase is lower in the sanction condition than in the no-sanction condition. This would indicate that the experience of playing a public-good game under a sanctioning system makes them less cooperative once the sanctioning system is taken away. In one of
their experiments, participants played a 4-person public-good game in which each decided how many of 100 chips (each worth EUR 0.05) to contribute for the provision of a public good. In addition, the participant who contributed the smallest number of coins faced a fine of EUR 5.00. Their trust—expected contribution by other members—was measured at each phase. The average level of trust in other members in the second phase was found to be lower in the sanction condition than in the no-sanction condition, as expected. Further, participants’ level of cooperation in the second phase was found to be lower in the sanction condition, though the difference was not statistically significant. Their second experiment found a significant difference in cooperation between sanction and no-sanction conditions among high-trusters [whose levels of trust were measured through Yamagishi’s (1986a) trust scale administered in a pre-experimental questionnaire], though not among low-trusters. Sanctions in this study undermined participants’ willingness to trust others and cooperate, when and only when the level of their trust in the absence of sanctions is already high.

**Social Value Orientations**

Social value orientations in psychology are equivalents of social preferences in economics. One of the social preferences economists consider relevant to sanction behavior is *inequity aversion* (Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000). Its equivalent—equality orientation—has been discussed in the social value orientation (SVO) literature to explain reciprocal or conditional cooperation behavior (Van Lange, 1999; Van Lange et al., 1997). Although the social value orientation literature has focused on behavior in social dilemmas (i.e., cooperation and defection), implications of SVO for sanctioning behavior will be a challenging topic for future research. Since sanctions (either rewards for cooperators or punishments of defectors) reduce the difference in payoffs to cooperators and defectors, equality-seekers are the ones who should deliver sanctions. No one has directly examined if this prediction is supported. Pinpointing a particular social value orientation that is directly tied to sanctioning behavior may provide an interesting and challenging avenue for SVO researchers.

In extending the SVO research to include sanctions, one challenge is how to explain sanctions against unintended defection. The concept of SVO has been used in two contexts—consequence-based and rule-based contexts. In the consequence-based model of SVO, inequity-aversions are motivated to reduce the payoff difference between cooperators and defectors by rewarding the former and punishing the latter. Intentions of cooperation and defection should not matter in this model. As experiments using truncated ultimatum games have shown, punishment of unintended unfair behavior is rather rare (Falk et al., 2003; Ohmura & Yamagishi, 2005). According to the rule-based model (Messick, 1999; Weber et al., 2004), SVO involves more than differential assignments of weights to various components of utilities. It is also related to the way a game player defines the game situation
("What kind of situation is this?") and the appropriate behavior in that situation ("What would a person like me do in a situation like this?") (Messick, 1999; Weber et al., 2004). In this approach, players make decisions not by comparing utilities of outcomes, but by following rules or heuristics that are appropriate for the situation. SVO is important as an indicator of how a game player defines the game situation and what kind of decision rules or heuristics she applies to it. According to the rule-based view of SVO, the intentions of cooperators and defectors should matter. This is because some of the rules or heuristics are directly tied to the perceived intentions of others. Psychological contributions are of critical importance on this issue.

**Ingroup, Outgroup, and Sanctions**

Are sanctions directed more toward members of one's own group, or toward members of another group? Does it matter if the sanctions are in the form of rewards or sanctions in determining which group they are directed to? Are rewards directed toward members of one's own group, and punishments more toward members of another group? These are questions of both theoretical and practical importance. There are several competing theoretical grounds to answer these questions. Social identity theory would predict that rewards are directed toward ingroup members (i.e., members of one's own group) and punishments toward outgroup members. This is because providing extra rewards to ingroup members and punishments to outgroup members increases the advantage of ingroup over outgroup, enhancing the positive distinction between the two groups. This prediction competes with another prediction based on the view of sanctions as a means of improving cooperation within one's own group. According to this perspective, paying costs to reward cooperators and punish defectors outside one's own group is a waste, and thus both rewards and punishments should be directed toward ingroup members. And yet, another prediction is possible by adding a "group-selection" aspect to the second perspective. This perspective provides a similar prediction as that from the social identity perspective; rewards toward ingroup members and punishments toward outgroup members provide a fitness advantage to one's own group vis-à-vis other groups. This perspective, however, provides a prediction concerning who in the other group is different from the social identity perspective; punishing cooperators, not defectors, in other groups provides a fitness advantage to one's own group. As discussed in the previous section, Carpenter and Mathews' (2002) participants punished ingroup members more than outgroup members in a repeated game, supporting the second approach. Shinada et al. (2004) conducted an experiment in which participants were provided with opportunities to punish members of their own group and those of another group. When they had such opportunities, they had been informed that there would be no future trials, eliminating consid-
erations for future self-interest. Two three-person groups played a public-good game within each group. When the subject was chosen as a “monitor,” he was given an opportunity to spend money to reduce the earnings of each of the other two members of his own group and of the three members of the other group. The results of this experiment were rather mixed. Cooperators punished mostly ingroup defectors, whereas defectors punished mostly outgroup defectors. Furthermore, in a partial replication of this in which the fine-to-fee ratio was reduced from 3:1 in the above study to 1:1, defectors stopped punishing outgroup defectors (Shinada et al., 2005). None of the above three approaches can explain these findings. These studies provide more puzzles than answers and invite further theoretical and empirical research efforts.

Goren et al.’s (2005) study addresses another issue of the effect of intergroup conflicts on sanctions. Intergroup conflicts have been known to make people cooperate more within each group (Bornstein & Ben-Yossef, 1994). Goren et al. (2005) hypothesized in this study that intergroup conflicts also promote sanctioning behavior within each group as a means to enhance cooperation there. Although their findings were not strong enough to draw a conclusion on this hypothesis either way, this study points to an important issue to be pursued. Furthermore, recent studies in experimental economics report the “parochial” nature of punishment in real social groups where the members have strong mutual rapport (Bernhard et al., 2006; Goette et al., 2006). Parochial nature of punishment means that third parties are more lenient if the norm violator belongs to their group, whereas they punish outgroup members who harm an ingroup victim. These results are intriguing, but we need to conduct further carefully designed studies before we draw firm conclusions based on the data obtained from real social groups.

Direct and Indirect Effects of Sanctions

We started this chapter with Hobbes. While studies of sanctions seem to represent our renewed interest in Hobbes, the most important insight in his discussion of Leviathan is missing from the contemporary studies of sanctions in social dilemmas. What is missing in the contemporary research on sanctions is Hobbes’ insight that Leviathan’s role is not in the direct control of its subjects but in the protection of peace. Hobbes starts his argument with the assumption that people come to prefer peace to war of all against all. Using terminology of the contemporary social dilemma research, Hobbes assumed that people are conditional cooperators instead of rational defectors who care only about their own short-term self-interests. People are willing to cooperate if they are assured that their cooperation will not be exploited by others. Leviathan’s role is in providing this assurance rather than forcing everyone to cooperate against their will.
Yamagishi's (1986a, b, 1988, 1992) earlier work on the provision of a sanctioning system was based on this idea of Hobbes. In these studies, Yamagishi emphasized the indirect effect of sanctions as a protector of peace rather than the direct effect as a coercer of peace. The costs, psychological as well as monetary, of providing sanctions, he argued, are much lower if the aim of sanctions is to convince conditional cooperators that it is safe to cooperate than if it is to force everyone to cooperate against his or her will. In his earlier studies, however, Yamagishi did not directly investigate the importance of the indirect effect of sanctions vis-à-vis the direct effect of sanctions. Almost two decades after Yamagishi pointed out the importance of the indirect effect of sanctions (and close to four centuries after Hobbes pointed out the same idea), Eek et al. (2002), Loukopoulos et al. (2006), and Shinada & Yamagishi (in press) conducted experiments to demonstrate the importance of the indirect effect of sanctions. Eek et al. (2002) called the indirect effect a “spill-over effect” and studied the relative size of the two types of effects by varying the target of punishment. In one condition, the participant alone was subject to punishment; since other members were not subject to punishment, only the direct effect was expected in this condition. In another condition, all members excluding the participant were subject to punishment; only the indirect effect was expected in this condition. They found a substantial indirect effect among prosocials and no indirect effect among proselfs. In a replication of this experiment, Loukopoulos and his associates (2006) found a significant indirect effect in both prosocials and proselfs, although the size of the indirect effect was only a fraction of the direct effect.

Shinada & Yamagishi (in press) noticed a potential problem in the above experiments that might have artificially reduced the size of indirect effect; the indirect effect was not completely excluded from the direct effect in those experiments. Participants in these studies knew that the other members knew they were subject to punishment; thus, they would be less afraid of their defection. The direct effect manipulation thus could have produced the indirect effect as well. Shinada & Yamagishi (in press) used a different technique and demonstrated that the indirect effect substantially explains the effect of sanctions. Their participants played a three-person, one-shot public-good game, under two punishment conditions. In the global punishment condition, all three members faced punishment. When punished, the amount of money the participant kept was reduced by half. In the direct effect condition, only the participant was told that only he or she had a chance of being punished and that the other two members would not be informed of anything about punishment. Therefore, the participant in the direct effect condition couldn’t expect the other members to cooperate under the sanction system since they knew nothing about sanctions. The average cooperation level of participants in the direct punishment condition who faced the possibility of punishment alone (0.43) was greater than that observed in the no-punishment condition (0.30), indicating that the direct effect of punishment can boost cooperation. Players in the global punishment condition who faced punishment together with their group members contributed significantly more money (0.54) compared to those in the direct effect condition. Thus, the size of the indirect effect is almost as big as that of the direct effect.
This result—increased cooperation over and above what was possible by the direct effect alone—provides evidence for the strong indirect effect of punishment. Furthermore, regression analysis indicated that the effect of punishment was mediated by the expectation in the net effect condition; a sanction system has indirectly increased contribution by changing beliefs about others. While not ruling out the importance of the direct effect of punishment, these results demonstrate that "ruling by the sword" alone is insufficient to convince people to behave in a mutually beneficial manner. As Hardin (1968) and Ostrom (1990) argued, the key to a successful sanctioning system is consent by the people involved; voluntary acceptance enhances the efficacy of punishment with the indirect effect. While the direct effect depends more on the actual controlling power of a social institution, the indirect effect depends more on the conviction that other members believe in its power. A sanctioning system supported by a shared belief system would be more effective than sanction by "sword" alone.

Let us conclude this chapter with the following final remark. Sanctions are where psychology meets structure. Studies of sanctioning systems and sanctioning behavior, we believe, will be a fertile ground for a truly interdisciplinary research endeavor to study human cooperation.

References


